A Comparison of GLM, GAM, and GWR Modeling of Fish Distribution and Abundance in Lake Ontario

by

Robert Edward Alexander

A Thesis Presented to the
Faculty of the USC Graduate School
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
Master of Science
(Geographic Information Science and Technology)

May 2016

To Meg Harris who introduced me to the world of GIS. A teacher, mentor, and friend.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I'd like to thank Dr. Brian Weidel and the crews of the USGS and NYDEC trawling vessels for their hard work and dedication in developing an extensive Lake Ontario fisheries database. Chris Castiglione of the US Fish and Wildlife agency for the Lake Ontario habitat data he and Dr. James E. McKenna of the USGS provided. I'd also like to thank Dr. Karen Kemp for her help as my thesis advisor as well as Dr. Laura C. Loyola and Dr. Travis Longcore for their participation on my committee.

List of Abbreviations

AIC	Akaike Information Criterion
AICc	Akaike Information Criterion with correction
ANN	Artificial Neural Network
CPUE	Catch per Unit of Effort
EPA	Environmental Protection Agency
Esri	Environmental Systems Research Institute
GAM	Generalized Additive Model
GIS	Geographic Information System
GLM	Generalized Linear Model
GWR	Geographically Weighted Regression
k	Cohen's Kappa
MGET	Marine Geospatial Ecology Tools
NOAA	National Oceanic and Atmospheric Administration
NOS	US National Ocean Service
OLS	Ordinary Least Squares
R^2	Coefficient of Determination
SDM	Species Distribution Model
TSS	True Skill Sensitivity
USC	University of Southern California
USGS	United States Geological Survey
VIF	Variance Inflation Factor
ZIP	Zero Inflated Poisson

Abstract

With advancements in GIS technology and computer capabilities there has been an increased interest in species distribution modeling (SDM). Previous works have focused on creating SDMs to determine presence while many ignore how the environment interacts with the species abundance levels. This study attempted to determine the most suitable method for predicting spatial distribution as well as the abundance of several different fish species in Lake Ontario. Ten fish species that were observed in Lake Ontario benthic trawling surveys at least 5% of the time between 1978 - 2014 were used to develop models. Subsets of the original dataset were also used to account for periods of time that saw major changes in Lake Ontario. This included a dataset before the invasion of dreissenid mussels, a dataset after the invasion of dreissenid mussels, and a dataset for the years limited to when Round Goby (*Neogobius melanostomus*) occurred within the trawling surveys. Generalized Linear Models (GLM), Generalized Additive Models (GAM), and Geographically Weighted Regression (GWR) were compared to each other to determine the best method. Habitat variables used to determine abundance relationships consisted of depth, fetch, fishing depth temperature, distance to major rivers and wetlands, as well as the presence of other fish species in the trawl. Adjusted R^2 and Cohen's Kappa were the primary indicators for determining the best method. None of the methods were able to produce good models with the habitat and biological data used. GWR did show an improvement in overall modeling performance, based on this study's criteria, over GLM and GAM. This was done by producing adjusted R^2 and Cohen's Kappa values similar to the GAM models while using a less complex regression model by using fewer predictive variables.

Chapter 1 Introduction

The ability to model species distribution has benefited ecologists, wildlife managers, and conservationists by helping them determine where a species can be found, what areas are most important for preservation, or what environmental and biological factors significantly influence the targeted species. Species distribution models (SDM) not only assist in better understanding for desired or endangered species but could also be used to determine how invasive species spread and to track their progress. SDMs can also assist in determining what possible actions could be done to influence the distribution of a species or if altercation of an area could impact a species distribution (Guisan et al. 2013). While SDMs should not be used as the sole factor in conservation decisions, they can be a powerful tool in identifying the best locations to focus on, what actions can be taken to remedy the issue, and to investigate how changes in the environment could affect distribution.

Geographic information systems (GIS) and remote sensing technology have allowed researchers a greater ability to assign habitat data to species observational data. As GIS and remote sensing technology advances, better resolutions of data are becoming more readily available. In the past, broad scale representations of environmental variables such as land cover, elevation, and other variables were available but these environmental variables were often too different in scale compared to observational data to properly represent the environment of the observed location. This often led to the aggregation of observational data into larger areas that could match the environmental variable resolution (Graham et al. 2001, Nishida & Chen 2004). With finer scale remotely collected data as well as more effort to collect habitat data from the field becoming more available, the ability to aggregate observational data in smaller areas or to keep as individual observational records could become more common.

Observational data for distribution modeling typically comes in two different forms, presence and absence data or presence only data. Presence and absence data is collected when an observer studies an area and documents if the target species is present or not. While data can be collected on the habitat characteristics of absence locations it does have the potential to include numerous zeros into the dataset. While access to presence and absence data is preferred, it is not often available.

Presence only data is collected when an observer documents only where the target species is encountered. Thus, locations where the species does not occur (absences) are not recorded. The analysis of presence only data sometimes requires the fabrication of pseudo-absences, which are randomly generated points within the study area that are considered absences but were not actually observed. Since absences were not recorded by an observer, it is possible that a pseudo-absence can occur at a location that is truly a presence (Graham et al. 2004). Observational data that does not properly sample the entire study area could lead to biased results because some habitat types might have not been sampled at all even though they may be included in the study area over which the model was developed. This research depended upon presence and absence data that contained species count numbers, allowing not only for the habitat of absences to be known but different abundances as well.

The choice of the SDM method to use is highly dependent on the availability of observational data and the dispersal of that observational data. The choice of SDM methods is also subject to the needs of the analyst. SDMs that simply predict where a species could or could not be have commonly been used for plants and animals. Methods such as Bioclim, Domain, GARP, and Maxent have been used for numerous studies but only allow for results ranging in values of 0 to 1, with 0 being an absence and 1 being the highest probability that the target

species is present (Hernandez et al. 2006). Presence and absence distribution can be useful with rare species where abundances are overall low. When knowing where the lowest and highest abundances are likely to be found is important, these methods would not offer as much insight as Generalized Linear Model (GLM), Generalized Additive Model (GAM), Geographically Weighted Regression (GWR), and Artificial Neural Networks (ANN). The use of regressions and ANN allow for not only presence distribution to be determined but also abundances to be calculated from observational data. Regressions and ANNs are also beneficial when relationships between the target species and environmental variables are not completely understood.

1.1 Motivation

Having the ability to predict the distribution and abundance of an aquatic species would allow lake managers, conservationists, and commercial fisheries to better utilize their time and resources for accomplishing their mission statements. Knowing the best modeling method to use in a particular situation should lead to the development of better predictions and more informed species management decisions.

Since presence and absence only prediction results can only predict distribution, they are unsuitable for use in predicting abundance. This limited function of presence and absence only predictions, which is utilized by methods like Bioclim, Domain, GARP, and Maxent would not be useful for this study. ANN has been considered a powerful tool due to its ability to determine non-linear relationships to predictors which could allow for better identification of sparse distribution (Pearson et al. 2002). Although ANN has the capability to predict abundance values and has had success outperforming other methods it was been observed to overfit (Heikkinen et al. 2006). This tendency to overfit could lead to some variables to being seen as more significant than they truly are or significant when they are not. If researchers focus on variables made to fit

the model, they could possibly ignore other available data that could be better suited for use in the model or prevent them from considering other environmental variables that can be collected from the field to improve their models. To overcome this setback, a method is needed that can allow analysts to predict abundance values as well as control the possibility of overfitting the model.

Environmental conditions can assist in species distribution modeling because of the requirements and preferences of each species. For example, water temperature can be used to determine where cold or warm water species can be found. Depth could assist in helping to predict where light sensitive species are located. Biological data could be used to indicate food availability or the existence of competition.

While all the mentioned methods have their merits, this study focused on GLM, GAM, and GWR. These methods were chosen because they can develop predictions that not only note presence but provide an abundance estimate. The possibility of overfit can be somewhat limited by limiting the number of variables used to explain the response. An additional reason these modeling method were selected was that it is not entirely clear what environmental influences are most significant for predicting abundance of fish species. Rather than have an ANN that could overlearn using a non-significant predictor until it fits the data, the selected regression methods can potentially help spot poor performing predictors. GLM was specifically chosen because it is a linear regression and would be a good base to compare with a GAM which utilizes smoothers that can allow for nonlinear relationships. GWR was specifically chosen to compare the difference between global regression equations (from GLM and GAM) and a local regression equation. In a small study area, a global regression equation could be suitable due to less variation in environmental variable values. However, in a study area that encompasses a large

area with sampling events dispersed throughout, model performance could do better with a local regression equation.

Lake Ontario was chosen as the study area for a number of reasons; (1) Lake Ontario is smaller than a marine ecosystem and is more contained than a marine ecosystem but larger than most lakes so that a local regression equation could be better tested, (2) Lake Ontario, due to its commercial and recreational importance, has a wide variety of environmental and biological data available that can be used for analysis, (3) Lake Ontario, again due to its commercial and recreational importance, could greatly benefit from fish abundance prediction models.

Better SDMs for game fish species could be used by game and wildlife officials to create more appropriate guidelines for catch limitations and even special regulations for specific regions for the lake. By creating stable fish populations the recreational fishing industries involved in a number of the cities and towns that border Lake Ontario could have long term fisheries security. Industrial fisheries could use the SDMs to determine the most efficient areas to harvest while limiting the number of harvesting events, avoiding sensitive areas for rare species, and reducing overall bycatch. Lake mangers and conservationists could better determine which areas of the lake are most significant to a rare or endangered species and best utilize time and funding by focusing their efforts on those areas. Research biologists could utilize SDMs for better time and fund management by focusing on areas with the highest abundances for collecting samples as well as lower abundance areas to study reasons for lower numbers.

The proper use of SDMs to support fisheries management decision making for Lake Ontario could help sustain stable fish population by promoting better understanding of the key environmental variables. These models can also be used to gain understanding of interaction relationships between different species.

1.2 Research Questions

This study aims to identify which regression method performs best with predicting the distribution and abundance of Lake Ontario fish species. Secondarily, this study attempted to develop models that could better predict both distribution and abundance. The predictor variables used in this study include environmental data collected at the time and location of the trawling event as well as environmental data developed from remote sensing. Environmental data obtained from remote sensing largely focused on the trawling event's distance to an object (i.e. river mouth or wetland) since the values of distance vary smoothly, do not change over short time ranges, and can be directly computed. While spatially and temporally dynamic variables such as primary production, dissolved oxygen, or other water quality characteristics could also be determined by remote sensing or extrapolation from point sources, it would be difficult to capture this data at the specific time and place of the trawling events. Therefore, the dynamic environmental variables used for this study were collected in situ at the same moment as the response variable so that they match both the temporal and spatial scale of the species density data. The presence and abundance of other species will also be used as a predictor variable for developing models.

By creating a variety of SDM models from a large collection of both fish density and environmental data collected in Lake Ontario, this study aims to answer the following research questions:

1) Which modeling method yields the best overall results; Generalized Linear Model, Generalized Additive Model, or Geographically Weighted Regression?
2) Does a local regression (GWR) perform better than a global regression (GLM and GAM)?

3) Given the various measures of model success discussed in this thesis, can good distribution or abundance models be produced for any fish species using these methods?

1.3 Summary

In order to develop SDMs that can predict abundance as well as presence, the best modeling method must first be identified. Generalized Linear Model, Generalized Additive Model, and Geographically Weighted Regression methods were investigated to determine which performed the best using both environmental and biological predictor variables. This study investigated which method worked the best with fisheries data and if a local regression outperforms a global regression. This study also investigated if models with high adjusted R^2, ≥ 0.70, and Cohen's Kappa values that suggest moderate or better agreement between observed and predicted values could be produced with the predictor variables used in this study.

The next chapter details preliminary research that was done to understand problems that exist with fisheries datasets and to explore how successful previous research was at using GLM, GAM and GWR methods for prediction of fisheries distribution and abundance. Chapter 3 outlines the source and type of data used for the analysis. Chapter 4 outlines the procedure that was used to develop and the criteria used to compare the results for the GLM, GAM, and GWR methods. Chapter 5 provides the results from all three modeling methods. Chapter 6 summarizes conclusions arising from the results and explains which method was determined to perform best at developing Lake Ontario fish distribution and abundance models.

Chapter 2 Background

To help identify issues in performing the intended GLM, GAM, and GWR a review was done on the general issues that exist in dealing with aquatic or fisheries data. Research into past studies using fisheries data was also performed to determine the successes and failures of past attempts. This insight into past research made it possible for solutions to be made in this study to correct issues that have arisen for others. The history of Lake Ontario was also studied to make an account of any major events that occurred within the lake that may impact the fish populations.

2.1 Fish Observation Methods and Issues

Unlike terrestrial or avian species that can be tracked with GPS, satellite imagery, and remotely or personally observed in the field, an aquatic species is very difficult to track and monitor. GPS and radio transmissions are often limited to larger fish and aquatic animals due to device size and concerns on how it may negatively affect the subject (Bryne et al. 2009). GPS also requires the subject to remain close to the surface, ~1.5 meters, for the device to broadcast (Sims et al. 2009). While advancements with satellite imagery and its ability to penetrate surface water are being made, there is still a severe limit of how deep the view can go and it is highly dependent on the clarity of the water system. Landsat imagery used for determining water depth through remote sensing showed that results rarely exceeded 25 meters due to visual limitations of the equipment (Stumpf et al. 2003). While this may allow researchers to track large aquatic animals like whales and some sharks that are near the surface or to monitor coral reefs, it would be of little to no use in assisting to identify benthic species.

The biggest difficulty in studying an aquatic species is that it cannot be studied in its native environment successfully. While some submersibles, both manned and unmanned, can view the target subject in their habitat, the observer's intrusion into the environment alone can

likely influence the target's behavior, as would a diver (Usseglio 2015). The inability to observe the target species without altering its behavior creates a profound lack in the understanding of what influences the target.

Traditionally the most common method of determining a fish species presence and abundance is through sampling with either a passive or active method (Schlieper 1972). Passive methods would include stationary traps or nets that fish swim into and become entrapped. Active methods include field personnel using moving nets/seines, hooks, or electricity to collect fish. More modern techniques now involve tags detected by buoys and handheld devices as well as cameras (Sippel 2015, Fukuba 2014).

Both traditional and new techniques have their pros and cons. Traditional methods, like nets and electric shock, generally offer a snap shot of the conditions at a single time, but are well tested and are capable of capturing large numbers of individuals. With new methods, like tags and cameras, obtaining movement data is now possible. However, with tags the individual fish needs to be captured for implantation and observation cameras rely heavily on light conditions, depth, and water clarity. While these new methods offer new insights into fish behavior, they are unlikely to be able to assist in large scale abundance studies for multiple wild population species.

2.2 Previous Fish Modeling Research

Species distribution modeling has been a growing tool for the last few decades as biological understanding and technological advancements have been made. However, as mentioned before, most of that research has been focused on plants and animals in terrestrial settings. For this study, emphasis was placed on reviewing research using GLM, GAM, and GWR in marine and lacustrine ecosystems.

2.2.1. Great Lakes fish species distribution and abundance research

McKenna and Castiglione (2014) produced distribution and abundance models for Silver Chub (*Macrhybopsis storeriana*) in the western basin of Lake Erie using an ANN. The observational data used by McKenna and Castiglione was collected by a combination of electrofishing and trawling events undertaken by US and Canadian agencies. While the study did include a number of environmental variables, McKenna and Castiglione did not include any biological predictors. This lack of biological data could be due to the absence of any useable datasets, which can be the case when working with targeted surveys and not community wide surveys.

McKenna and Castiglione were also interested in extrapolating the results of the model into unknown areas of the western basin of Lake Erie to see what the theoretical range and abundances of Silver Chub would be. This extrapolation was based solely on environmental habitat needs and could explain why biological data was not used for the analysis. Since biological predictor data is rarely available for large areas it would be difficult to extrapolate the model predictions into unknown locations. The environmental predictors that were most significant in their datasets were the cosine of direction to nearest delta wetland, cosine of direction to nearest open wetland, sinuosity of the nearest shoreline, submerged aquatic vegetation covering ≥50% of the bottom substrate, water depth, distance to nearest open wetland, distance to nearest protected wetland, distance to nearest large river mouth, nearest shore line geomorphic type, coastal reinforcement condition of nearest shoreline, mean surface water temperature, and coefficient of variation of surface water temperature.

The ANN used for this study produced R^2 values greater than 0.8. Cohen's Kappa was also used to determine how much agreement not associated with chance existed between predicted and observed values. Kappa values for their moderate and high abundances were

considered to be fair or substantial using a ranking system by Landis and Koch (1977). While these results from the ANN were good, the paper fails to indicate if or how observations were aggregated before modeling.

2.2.2. Generalized Linear Model only fisheries research

Nishida and Chen (2004) utilized the GLM regression method in an attempt to see if the inclusion of a spatial autocorrelation parameter in the GLM could improve performance for determining longline catch per unit of effort (CPUE) of Yellowfin Tuna (*Thunnus albacares*) in the Indian Ocean. Nishida and Chen included this spatial autocorrelation component as covariograms using an exponential model, spherical model, Gaussian model, and a linear model. To avoid issues with zero value data, Nishida and Chen added to all the data a constant that was equal to 10% of the global CPUE mean. Data was also aggregated in units of 5° latitude by 5° longitude. Other variables beside the spatial component used in the models were year, month, sea surface temperature, and thermocline depth.

A total of 10 models were run, two with no spatial component and eight which incorporated the spatial component. The two GLM with no spatial component models, one incorporating a habitat based model and the other not, resulted in coefficient of determination (R^2) values of 0.585 and 0.602 respectively, both with the highest Akaike Information Criterion (AIC) values out of all 10 models. The eight GLM with a spatial component, four incorporating a habitat based model and the other four not, produced R^2 values ranging from 0.711 to 0.768. Of the four methods to calculate the spatial component, the Gaussian model achieved the highest R^2 value and the lowest (i.e. the best) AIC value in their groupings. The Gaussian model was followed by the spherical model, the exponential model, and the linear model in decreasing order of success.

Nishida and Chen's study concluded that the GLM with a spatial component performed better in analyzing Yellowfin Tuna CPUE than a GLM without a spatial component. However, they also discuss how this method has not been thoroughly applied to other species, datasets, or location. They also discuss how this method is intended for strongly spatial autocorrelation datasets.

It should be noted there is some criticism by including a spatial variable into regression to correct the effect of spatial autocorrelation. Carsten Dormann (2007) argues that the inclusion of this spatial variable can underestimate the environmental variables effect on the model and may introduce additional bias to the model when compared to models that do not use a spatial variable.

2.2.3. Generalized Additive Model only fisheries research

Graham et al. (2001) use a GAM as the primary modeling method when developing a GIS for a cephalopod fishery. Other methods of modeling are suggested in their work such as ARIMA, GLM, and tree-based models but there are no results or details on procedures, possibly suggesting a lower performance than the highlighted GAM model. Analysis was done for a study area encompassing a large section of European waters in the Northeastern Atlantic with observations aggregated to a resolution of l° longitude by 0.5° latitude cells. Environmental variables used for analysis were sea surface temperature, sea bottom temperature, sea surface salinity, sea bottom salinity, and bathymetry.

The results for the 1990 Cuttlefish predictions were deemed "poor" at fine spatial resolution even though the authors observed a broad agreement between predicted and actual abundance at a broader scale. The authors did not report any statistical indicators of model performance which prevents the reader from understanding what the authors consider "poor" or

drawing their own conclusions on how the model performed. The poor results could be connected to using aggregated zones having different sized areas dependent on their latitudinal position. The use of predator or food source variables could have increased the performance of the GAM.

2.2.4. Geographically Weighted Regression only fisheries research

GWR is a relatively new method used to model species distribution; the earliest use of this method in fisheries that could be found was in 2009 in a methods comparison between GLM, GAM and GWR (Windle et al. 2009). A more focused study on the improvement of an OLS model with the use of a GWR model is the thesis of Jamie Kilgo (2012). Finding research focusing solely on GWR even outside the field of fisheries is difficult, an ordinary least squares (OLS) model is often included in the study to first determine if the data should be run with a GWR as well as a means to measure how much improvement the model gained.

In Kilgo's study fish biomass of five different reef species were modeled with seven available variables. The landscape variables used were nearest patch or linear reef, distance to boundary, coral reef proportion, seagrass proportion, macroalgae proportion, non-colonized pavement proportion, and sand proportion. Data exploration showed that seagrass and sand proportion were not suitable variables for any of the species and were not used in any of the models. The results of the OLS had R^2 values ranging from 0.0191 to 0.2841 and the GWR with R^2 values ranging from 0.0507 to 0.3467. Each species had an R^2 value increase and a lower AIC value from the use of a GWR.

2.2.5. Comparison of methods for fisheries research

Murphy et al. (2015) compared GLM and GAM along with a number of other modeling methods to determine the best method for predicting the distribution of a highly tolerant invasive species, Eastern Mosquitofish (*Gambusia holbrooki*) throughout the Iberian Peninsula. Environmental variables consisted of elevation, slope, topographic index, flow accumulation, urban land use, agriculture land use, annual precipitation, annual mean temperature, thermal range, population density, number of local dams, and total number of dams upstream. Cohen's Kappa (k), area under receiver operating characteristic curve (AUC), and true skill sensitivity (TSS) were used to determine overall model performance.

In this study, the GAM method (k = 0.43, AUC = 0.82, TSS = 0.48) slightly outperformed the GLM method (k = 0.40, AUC = 0.81, TSS = 0.45). Compared to the other methods used, the GLM and GAM methods were only superior to the surface range envelopes method (k = 0.10, AUC = NA, TSS = 0.16). The GLM and GAM methods performed closest to the multiple adaptive regression splines (k = 0.47, AUC = 0.82, TSS = 0.49), mixture discriminate analysis (k = 0.43, AUC = 0.80, TSS = 0.44), classification tree analysis (k = 0.49, AUC = 0.81, TSS = 0.61), and artificial neural networks (k = 0.49, AUC = 0.78, TSS = 0.41) methods. Of the methods used boosted regression trees (k = 0.64, AUC = 0.92, TSS = 0.41), Maxent (k = 0.77, AUC = 0.96, TSS = 0.77), and random forests (k = 0.88, AUC = 0.98, TSS = 0.85) methods performed the best. While results for this study showed GLM and GAM among the lowest performing modeling methods, this may have been due to the use of presence only data and was focused only on determining presence and absence rather than abundance levels.

Windle et al. (2009) explored the possibility of a local regression outperforming a global technique. The species of interest was Atlantic Cod (*Gadus morhua*) in the Northwest Atlantic. Observation data was classified as present or absent, with absences consisting of catch weights

five kilograms or less. This was done to prevent small catches from influencing the model too much. Point events were not aggregated into a larger area but kept as individual events. Environmental variables used for the models were temperature, mean bottom depth, salinity, distance to shore, crab biomass, and shrimp biomass were available for analysis. The variables found best suited for the analysis was temperature, distance to shore, and the biomass of crab and shrimp.

The resulting GLM showed the worst results with the lowest R^2 (0.013) and the highest Akaike Information Criterion with correction (AICc) value (323.4), The GAM showed better results with a R^2 value of 0.072 and an AICc value of 313.7, the GWR performed the best with the highest R^2 values of (0.11-0.26) and the lowest AICc value (271.4). The correction to AIC is often used when observational data is small in number. While the R^2 values are poor overall there is an improvement in prediction by using a GWR method over the GLM or GAM methods.

2.3 Transformations and Zero Inflated Data

Transformations can be applied to either the response or predictor variable to improve the performance of data analysis. There are four major reasons for a transformation to be made on a response or predictor variable. The first is to reduce the effect an outlier would have on the dataset. The second is to improve the linearity between a response and predictor variable. The third is to make the dataset closer to a normal distribution. The fourth reason is to stabilize the relationship of the mean and the variance. When extreme observations exist in the response variable a transformation can be used to reduce effect of extreme values. Alternatively, analysis methods that are better equipped at dealing with them such as GLM or GAM that use a Poisson distribution can be used (Zuur et al. 2007).

O'Hara and Kotze (2010) discuss the disadvantages and issues that arise when using transformations for ecological count data. O'Hara and Kotze point out that log-transformation is widely used with count data but there is no significant literature for a reason to use log-transformations over other transformations. They suggest that this could be because log-transformations are one of the first topics discussed in textbooks or that much of the analysis done with ecological data is by biologists rather than statisticians who would better understand the pros and cons of different transformation types.

When collecting ecological count data from methods that include traps or grid searches, the possibility of recording zero observations is possible. While access to absences for a species can be useful in determining its spatial distribution, it can cause issues when determining spatial distribution of abundances due to the possibility of numerous zero observations skewing the data towards lower values. Barry and Welsh (2002) discuss how a dataset that is zero inflated can be difficult to predict from standard error models used with GLM. They state that if the high number of zeros is ignored and the model is applied with a standard Poisson error model, then there can be problems with inference. O'Hara and Kotze also discuss how zero observations are dealt with using a log-transformation by the traditional method which is to add a constant to all values to eliminate zeros. A value of one is commonly used but there is no method standardization that dictates what value should be used for a constant value and the use of different values could alter the fit of the model.

Leathwick and Austin (2001) raised this issue about assuming a Poisson distribution with zero inflated data when modeling competitive interactions between tree species. During their modeling tests, they used transformed data to assume a Gaussian distribution as well as a non-transformed dataset with a Poisson distribution, their results suggested that while the Poisson-

based model was not ideal it was the best available. Barry and Welsh suggest that a better method is to model zero inflated data in two steps using the Zero Inflated Poisson method (ZIP). The first step is to perform a presence/absence model, 0 or 1 value, using all data. Once this model has been created the spatial distribution of the species in question can be determined. The second step is to use only the presence data to model the abundance levels values. If the models were to be extrapolated into unknown areas the abundance value model would only be applied to areas that were determined to be a presence by the first method.

When using the ZIP method, the error of two different models would have to be considered when trying to make decisions from the results. Results from a ZIP method that produced a good distribution model but a poor abundance model would only tell a decision maker where the species most likely was. The model would not be able to predict abundance with significant confidence, leading to limits in the decisions that could be made. A result that produces a poor distribution model but a good abundance model could misidentify which areas a species could be found in.

Barry and Welsh raised the subject that a Poisson based model would not be able to predict a large amount of zeros with a zero dominated dataset. However, this is discussed mainly in association with non-mobile species such as trees, as was the case in the research done by Leathwick and Austin. However, when dealing with mobile species, such as fish, an absence does not necessarily mean that the area is not used by the species but simply was not in the location at the time of sampling. The ability to observe mobile and non-mobile species also plays a role, it is easy to search a grid and determine the exact number and position of a tree species, but the capture of fish species is not as efficient and false zero observations are possible. So predicted values from a Poisson method that does not match observed zeros could be studied and

compared with nearby sampling locations as well as observations from previous years, if available, to determine if the value could be a true absence or if the predicted value is possible.

2.4 Relevance to this Research

Most of the previous research for GLM, GAM, and GWR is based on comparison studies to see which model performs best. Marine ecosystems were also commonly used for the study areas for GLM, GAM, and GWR. This is likely a result of the importance of finding commercial fisheries in such large areas. Much of the research available on fisheries species distribution models focuses on presence and absence data that is better modeled with methods other than GLM, GAM, and GWR (Murphy et al. 2015). ANN modeling with abundance categories also seemed to outperform GLM, GAM, and GWR but no studies could be found that included all these models to predict abundance values.

While past research showed that ANN has performed better than GLM and GAM, they can be more susceptible to overfitting if environmental relationships with the target species are not completely understood. As discussed above, including many environmental variables can also lead to the possibility of overfitting the model. Models that are the result of overfitting could lead to misinterpretation of significant variables and possibly result in poor decision making in conservation or harvesting operations.

Since environmental relationships are not fully understood with the species in this study, ANN was not be used because of this susceptibility. An additional reason ANN was excluded as a method was due to the complexity of the model that is developed.

When comparing GLM, GAM and GWR with each other, GLM has often been seen as the poorest performing method with GAM boosting performance. GWR, when compared to GLM and GAM, showed an increase in performance in both R^2 and AIC values. The previous

research seems to indicate that fisheries data is often better modeled with a nonlinear method (GAM) than a linear method (GLM) when using a global regression. When comparing the performance of a global regression and a local regression, the local method tends to outperform the global method. The previous research also stresses the importance of choosing the modeling method that fits the data distribution (Gaussian or Poisson).

One of the biggest issues that many researchers must overcome is the large number of zero data values. This is the reason that many choose to use observations as presence and absence only. The decision to model with the zeros as is or to implement the ZIP method should depend on whether the researcher is confident that the zeros in the data are either true absences or false absences due to gear malfunction or simply bad timing and luck. This study did not implement the ZIP method since it would add too much complexity to model development, making it difficult to compare base performance between GLM, GAM, and GWR.

The study by McKenna and Castiglione showed what lake wide environmental data was currently available to use in analysis for the Great Lakes region. The environmental data shown to be available by McKenna and Castiglione in combination with the success other methods have had by including biological data, gave needed insight into the predictor variables that are most suitable to use for model development.

2.5 Research Gaps

There are a number of gaps in current research in fisheries species distribution modeling. The most notable of these are:

(1) The use of biotic variables in distribution models is lacking in many studies. The presence and abundance of food or a predator could not only influence the probability of presence but the abundance of the target species as well.

(2) Many of the studies focus on the water surface temperature, most likely due to the readily available sources. While the surface temperature can give an idea of the temperature just below the surface, it would not be relevant to fish further from the surface.

(3) The largest gap in fisheries species distribution models is the focus on presence probability. By modeling for abundance as well as distribution, the models would better locate important habitats indicated by the higher abundances.

This research sought to address these gaps.

2.6 Study Area: Lake Ontario

The study used in the analysis is Lake Ontario. While the observational and environmental attributes are the focus of the analysis, it is also important to take the history of the study area into account. Because it is not possible to obtain a dataset covering the entire time period of the study area, it is wise to know what the state of the study area was prior to the creation of the datasets just as much as the state during the dataset.

2.6.1. Physical characteristics

Lake Ontario is the eastern most of the Great Lakes situated between New York State, USA and the Providence of Ontario, Canada (Figure 1). The lake is located at the lowest elevation of all the other Great Lakes as well as having the smallest length, breadth, land drainage area, water area, total area, and shoreline length. While Lake Ontario might have the lowest values of these features, it does have an average water depth greater than all the other lakes beside Lake Superior. Maximum water depth is also greater than Lake Erie and Lake Michigan. While the smallest of the Great Lakes, it still has more volume than Lake Erie. Lake Ontario's retention time is one of the shortest with only a period of six years. This retention time

is much shorter than lakes Huron, Michigan, and Superior who have retention times over 20 years (Botts & Krushelnicki 1987). Table 1 summarizes all of these details.

Figure 1 Study Area: Lake Ontario (black box) in the Great Lakes region

Table 1 Physical Characteristics of the Great Lakes. Numbers in parenthesis indicate where the lake ranks among the Great Lakes with 1 ranking the highest values. * indicates that the physical characteristic was measured using the low water datum. ** indicates that island shorelines were included in the measurement. Source: Botts & Krushelnicki 1987

Attribute	Ontario	Erie	Huron	Michigan	Superior
Elevation (m)*	74(4)	173(3)	176(2)	176(2)	183(1)
Length (km)	311(5)	388(3)	332(4)	494(2)	563(1)
Breadth (km)	85(5)	92(4)	245(2)	190(3)	257(1)
Average Depth (m)*	86(2)	19(5)	59(4)	85(3)	147(1)
Maximum Depth (m)*	244(3)	64(5)	229(4)	282(2)	406(1)
Volume (km^3)*	1,640(4)	484(5)	3,540(3)	4,920(2)	12,100(1)
Water Area (km^2)	18,960(5)	25,700(4)	59,600(2)	57,800(3)	82,100(1)
Land Drainage Area (km^2)	64,030(5)	78,000(4)	134,100(1)	118,000(3)	127,700(2)
Total Area (km^2)	82,990(5)	103,700(4)	193,700(2)	175,800(3)	209,800(1)
Shoreline Length (km)**	1,146(5)	1,402(4)	6,157(1)	2,633(3)	4,385(2)
Retention Time (years)	6(4)	2.6(5)	22(3)	99(2)	191(1)

2.6.2. Lake History

The history of Lake Ontario is a history of numerous changes. Many of these changes can be contributed to human actions on and around the lake. Settlements along Lake Ontario began to take off in the 1780s. With the settlements came the construction of mills and dams to meet the needs of the settlers. Construction of dams would become barriers for migrating species. As settlements grew, the need for resources to meet the demands and to create profit for the settlements led to an increase in land clearing for lumber and agriculture.

The early 1800s saw increased traffic on Lake Ontario with the introduction of steamships and the connection to the Erie Canal at Buffalo, Oswego, Rochester, and Hamilton. One of the biggest events of the early 1800s was the completion of the Welland Canal that connected Lake Ontario with Lake Erie. Land held by the British navy was released and more timber exploitation around Lake Ontario occurred leading to further landscape transformations around the lake. The 1830s saw the introduction of pound nets, which are trap nets with a fence leading toward the shore that funnels fish into the trap net. The 1850s saw the deployment of offshore fishing and gill nets used for major commercial fishing. The 1860s was the peak of mill

and dam construction in the Lake Ontario basin. Generation of hydro-electric power from the Niagara River started in the 1870s. The 1870s was also the peak period for timber exploration around the lake.

The deployment of cotton gill nets into the lake occurred in the early 1900s, an improvement over older gill nets. The 1950s saw the first signs of concern for high lake levels and shore erosion. The 1960s marked a major event with the opening of the St. Lawrence Seaway to ocean shipping. These vessels would become a key player in the unintentional release of several exotic species. Regulation of outflow from the lake began as well. Eutrophication, which occurred due to the result of nutrient loading, the primary nutrient being phosphorus, became an issue in the 1960s. The 1980s saw additional remedial actions implemented to address increases in toxic substances as well as a growing concern of habitat loss, increased shoreline development and the effects of high lake levels on shore stability.

To combat further eutrophication and to return the lake to a more historic production level, the Canada-US Great Lakes Water Quality Agreement was signed in the 1970s. The intent of this agreement was to reduce nutrient loading in hopes of lowering the productivity of the Great Lakes. A significant decrease has been observed (Sly 1991). The EPA set a trophic state goal of oligomesotrophy, which was considered to have been met in 2012 (Sullivan 2015).

2.6.3. Species invasions and declines

As with most ecosystems that have major a human influence, there have been many cases of exotic species, some intended and others unintended, introduced to Lake Ontario. These exotic populations have occurred at all trophic levels and often include many species that have a primarily marine distribution (Stoermer et al. 1985).

The 1830s saw the collapse of Atlantic Salmon (*Salmo salar*) stocks. Alewife (*Alosa pseudoharengus*) became naturalized within Lake Ontario in the 1870s. From the late 1800s and early 1900s saw Lake Sturgeon (*Acipenser fulvescens*) become very scarce. Sea Lamprey (*Petromyzon marinus*) became abundant in Lake Ontario in the 1920s. The 1940s saw the collapse of Lake Trout (*Salvelinus namaycush*), Burbot (*Lota lota*), Lake Herring (*Coregonus artedii*), and Deepwater Ciscoes (*Coregonus johannae*) stocks. While these species were declining, Rainbow Smelt (*Osmerus mordax*) were rising to become one of the dominate species.

In the 1950s controls were enacted to limit the number of Sea Lamprey in Lake Ontario. The mid 1900s also saw the White Bass (*Morone chrysops*), Blue Pike (*Sander vitreus glaucus*), and Deepwater Sculpin (*Myoxocephalus thompsonii*) disappear from the lake. The 1960s saw the collapse of whitefish stocks in the eastern portion of Lake Ontario. Lake Trout and salmonids saw a rebound in stocks but were highly dependent on human stocking efforts in the 1980s. Whitefish stocks also began to recover in the 1980s (Sly 1991).

The end of the 1980s marked the invasion of Zebra Mussels (*Dreissena polymorpha*) (Griffiths et al. 1991). The 1990s saw the introduction of other invasive species such as the Round Goby (*Neogobius melanostomus*) and Quagga Mussel (*Dreissena bugensis*) (Owens and Dittman 2003, Mills et al. 1993). The Lake Ontario deep water benthic community was historically comprised of the dominate Diporeia spp. as well as sphaeriids, oligochaetes, and chironomids (Cook and Johnson 1974). In the 1990s these deep water benthic organisms began to decline with spread of Zebra and Quagga Mussels (Watkins et al. 2007).

2.7 Summary

This review of research about SDMs developed using fisheries data shows that GAMs often outperform GLM. The studies also indicated that GWR use in fisheries is relatively new

but shows promise modeling improvements over GLM and GAM. A large range of environmental variables have been included in previous fisheries research, which suggests an appropriate set of predictor variables for this research. Research into the history of Lake Ontario revealed that there were numerous occasions when human development heavily impacted the lake. One of the biggest impacts to face the lake was the introduction of a number of invasive species. This suggests that the use of temporal subsets of the fisheries data used in this study that focus on the years before and after the introduction of the biggest threats may provide useful insights. The next chapter explores the data used in this study.

Chapter 3 Data

To develop the models, both environmental and biological datasets were collected to be used for predictor and response variables in the GLM, GAM, and GWR methods. Research into the history of Lake Ontario helped identify which years should be used to create dataset subsets that contained periods with the most meaning. This chapter explores the biological and environmental datasets that were used.

3.1 Biological Data

Benthic trawling surveys were obtained through personal contact with the United States Geological Survey's (USGS) Lake Ontario Biological Station, one of the Great Lake Research Centers, located in Oswego, NY. The primary purpose of these bottom trawls is to assess the status of important prey fishes. Bottom trawls are done in the spring for Alewife, summer for Rainbow Smelt, and in the autumn for Slimy Sculpin (*Cottus cognatus*) and Deepwater Sculpin. While the surveys are targeted at specific species, non-target species were always counted as well. The specific months during which each of these surveys were conducted were chosen following a study undertaken in 1972 when trawls were performed throughout May to October to determine the times of peak catches of the target species (USGS 2012).

Surveys are done with a fixed station sampling design, which is commonly used in the other Great Lakes as well as in northern Europe. The fixed stations used for these surveys are often confined to the same area year after year at the same depths year after year. It is assumed that mean abundance from fixed station surveys will not be biased if the fish population are randomly distributed. The assumption that fish populations are randomly distributed in the geographic area was confirmed with the use of acoustic sampling done during the 2004-2006 Alewife bottom trawl surveys (ICES 2004).

The trawls that occurred at these stations were done at a standard 10 minute bottom drag. While time on the bottom of the lake was roughly similar between events the area swept from surface to bottom was heavily dependent on the depth being sampled. This created a greater area swept for deeper water and less area swept for shallower water with depths ranging from 5 meters to 215 meters. The range of area swept in the dataset used was approximately 1,100 square meters to approximately 25,300 square meters.

Until 1997, bottom trawls were conducted with a Yankee trawl with a 12 meter headrope and flat rectangular trawl doors. In 1997, increased dreissenid densities interfering with trawls led to a change to use of a 3-in-1 bottom trawl with an 18 meter headrope and slotted, cambered V-doors. Yankee bottom trawls were still used for deep waters until 2004 where dreissenid densities were low. In 2011, the 12 meter Yankee trawl once again returned to being the only gear type used.

The trawl dataset obtained from the USGS consisted of a .csv file. Each trawl has an assigned unique identification number and the starting point for each trawl is included as lat/long coordinates. In addition to the number of individuals caught, the Lake Ontario Biological Station included their calculations for area swept and density (fish per square meter) for each fish species in each trawling event. Table 2 shows the columns in the original .csv file.

Table 2 USGS Database Field Descriptions

Field Name	Description	Format
OP_ID	Operation ID	Numeric
YEAR	Year of the operation	Date
OP_DATE	Julian date of operation	Date
LATITUDE	Latitude coordinate for start location	Numeric
LONGITUDE	Longitude coordinate for start location	Numeric
TARGET	Target species code	Coded Value
TARGET_NAME	Target species name	Text
FISHING_TEMP	Temperature at fishing depth	Numeric
FISHING_DEPTH	Fishing depth	Numeric
SPECIES	Species code	Coded Value
COMMON_NAME	Species common name	Text
N	Number of individuals	Numeric
WEIGHT	Weight of individuals	Numeric
areaSampled_m2	Calculated area swept (m^2)	Numeric
Number_Npm2	Calculated density of species (fish/m^2)	Numeric
Weight_gpm2	Calculated weight of species (g/m^2)	Numeric

A list was compiled of the fish species that were observed in at least 5% of the total number of trawls, resulting in ten species to model. The species present in sufficient abundance were Alewife, Round Goby, Johnny Darter (*Etheostoma nigrum*), Lake Trout, Yellow Perch (*Perca flavescens*), Slimy Sculpin, Rainbow Smelt, Spottail Shiner (*Notropis hudsonius*), Threespine Stickleback (*Gasterosteus aculeatus*), and Trout Perch (*Percopsis omiscomaycus*).

Since each species had a record in the original .csv for each unique event, for the purposes of this research, it was necessary to reorganize the table so that there was a single record for each unique trawling event and a column for each species. To create the response variables, a new field was created for each of the ten species to be modeled, labelled using a coded name for each. The individual species response variable fields were populated by multiplying the species square meter density by the average area swept (7,276 m^2) and rounded to the nearest whole number as shown in Table 3. The actual count numbers were not used due to the differences in area swept. The area swept is heavily influenced by the depth being sampled,

so models would be more heavily influenced by the deeper trawls. The original CPUE values were not used due to the units being for fish per square meter, which was populated with very small values (0 - 32) with a mean of approximately 0.1 fish per square meter. Thus, to achieve fish densities more likely to be encountered in an average trawling event, the average area swept was used.

Table 3 Response Variable Fields Added to USGS Database

New Field	Description
ALEW	Alewife density for average trawl length as whole number
GOBY	Round Goby density for average trawl length as whole number
JOHN	Johnny Darter density for average trawl length as whole number
LTRT	Lake Trout density for average trawl length as whole number
PRCH	Yellow Perch density for average trawl length as whole number
SLIM	Slimy Sculpin density for average trawl length as whole number
SMLT	Rainbow Smelt density for average trawl length as whole number
SPOT	Spottail Shiner density for average trawl length as whole number
STK3	Threespine Stickleback density for average trawl length as whole number
TRPR	Trout Perch density for average trawl length as whole number

Species abundances were also used as a predictor variable for other species but required a transformation to produce a better relationship between response and predictor. A square root transformation was chosen to limit the influence of extreme values without over generalizing the values that a Log_{10} transformation could have done. Thus, additional fields were created to be used as the predictor variables with other species that were named with the prefix sqrt and species code and populated with the square root of the density times 7,276 m^2. Equation 1 illustrates the formula for the Alewife species.

$$sqrtALEW = \sqrt{ALEW\ Field} \quad (1)$$

Four datasets were created to address the important temporal ranges identified earlier. These included the entire time period of the database, 1978-2014; the time period before the invasion of dreissenid mussels, 1978-1989; the time period after the invasion of dreissenid

mussels, 1990-2014; and the time period of occurrence for Round Goby to only be used in the modeling of Round Goby, 2004-2014. These trawl event sets are shown in Figures 2 and 3. The 1978-2014 dataset allowed the regressions to model the fish species for long term environmental and biological trends. The 1978-1989 dataset allowed the regression to model fish species for trends before the invasion of dreissenid mussels. The 1999-2014 dataset allowed the regression to model fish species for trends that started after the invasion of dreissenid mussels. The 2004-2014 dataset allowed the regression to model for Round Goby during their entire occurrence of the species during the database extent.

The reorganized and augmented trawling .csv table was used to create a feature class in Esri ArcGIS (version 10.2) named Trawling Events Complete using the trawl start location. This initial feature class was then used to create the data subsets based on years (Trawling Events 1978 - 1989, Trawling Events 1990 - 2014, & Trawling Events 2004 - 2014). All four of these feature classes were used iteratively as input for the GLM, GAM, and GWR tools described in the next chapter.

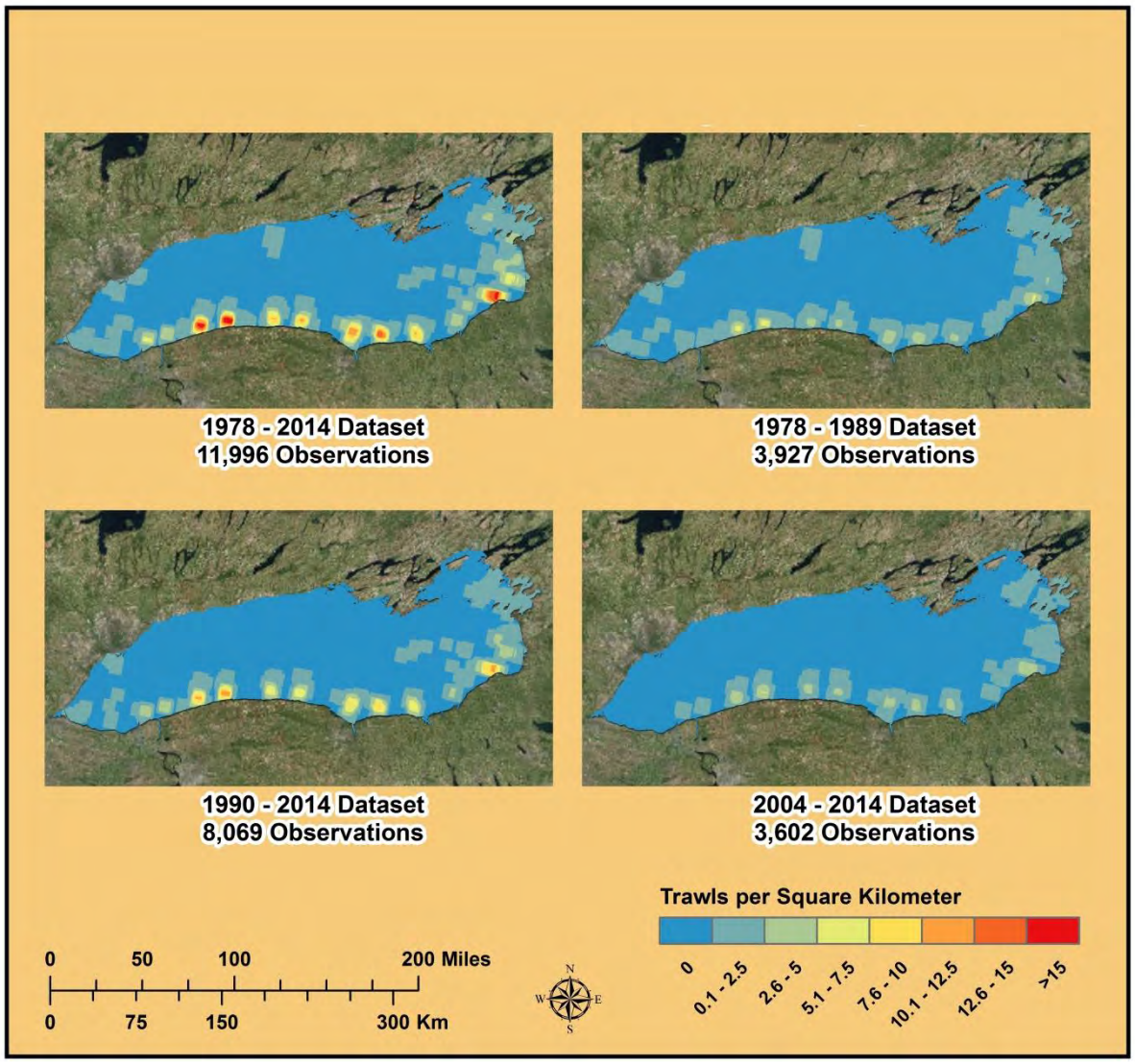

Figure 2 Densities of USGS Benthic Trawling Events per Km^2 for Each Dataset in Lake Ontario

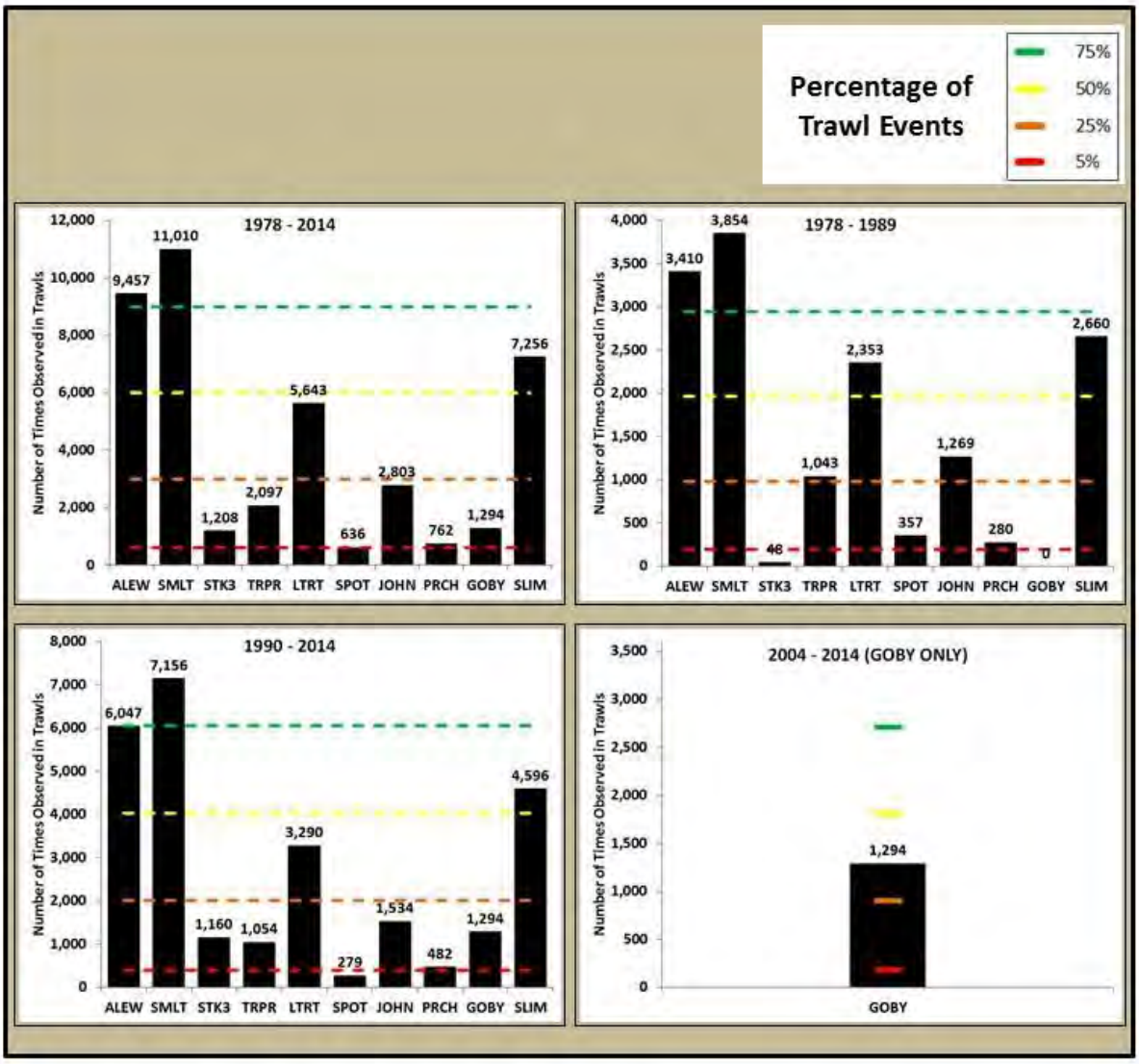

Figure 3 Species Occurrences in USGS Benthic Trawl Surveys for Each Dataset in Lake Ontario. Alewife (ALEW), Rainbow Smelt (SMLT), Threespine Stickleback (STK3), Trout Perch (TRPR), Lake Trout (LTRT), Spottail Shiner (SPOT), Johnny Darter (JOHN), Yellow Perch (PRCH), Round Goby (GOBY), Slimy Sculpin (SLIM)

3.2 Environmental Data

Environmental data used consisted of temperature at fishing depth, depth, effective fetch, distances to major river mouths and wetlands. These conditions contribute to the distribution and abundance of fish species by offering the necessary needs for temperature and structure. While some datasets may offer direct relationships, others could offer indirect relationships, for

example depth plays a role in the amount of light in the water. These indirect relationships were used in place of direct environmental variables for which there are no datasets that match the temporal or spatial scale of the fish database. All predictor variables, excluding fishing depth temperature and species abundance, were obtained from Chris Castiglione, the GIS Coordinator and Fish & Wildlife Biologist at the Lower Great Lakes Fish and Wildlife Conservation Office in Basom, NY. This database of environmental data was created as part of the Great Lakes Aquatic Gap project. The objective of the Great Lakes Aquatic Gap project is to classify aquatic habitats in rivers, streams, and lakes in the Great Lake basin using regionally consistent methods. The project also aims to develop biological databases at state and regional scales as well as mapping actual and predicted occurrences and distributions of aquatic species (Myers et al. 2002). All of the Castiglione data was obtained as 90 m raster files in the ArcGIS Grid format.

The environmental data that were not collected during the trawling events were extracted from these rasters for each trawling point and added to the attribute tables of the four Trawling Event feature classes. The trawl direction is not known so only the value of these rasters at the start location of the trawl was extracted to the feature class point. The variation of these environmental values from the start and end location of any trawl was considered to be small enough that the value at a single point is sufficiently representative. An area average around the point was not done because it would likely include areas not sampled. Another reason for not using an area average was that the environmental variables involved distance measurements that would likely average down to the center tile, which would be the tile occupied by the start location.

3.2.1. Temperature

Temperature (Temp) data was recorded at the time of each trawling event at the fishing depth and was included in the .csv file provided by the USGS. Since fishing depths reach deeper than 200 meters for some events and the average fishing depth is 60 meters, it would not be useful to use surface temperature. Since fishing depth temperature was recorded at the time of the trawling, the records match both the temporal scale and geographic scale.

3.2.2. Depth

While depth was included in the .csv from the USGS, it was commonly rounded to a set value (i.e. 20 m, 25 m, etc.). The depth raster values were extracted to the trawling events feature class so a comparison of values from the USGS .csv file to the depth raster included in the Castiglione data could be done. This comparison was done to see if there were any major differences between the .csv and raster values. Major differences could have indicated that there was a recording error in the trawling events coordinates or fishing depth fields. Since the trawling events are all benthic trawls, fishing depth should have been equal or close to the value of the depth raster. The comparison of depth values showed that the .csv values were within five meters of the depth raster signifying that the fishing depth field closely matched the depth raster. Since the depth raster values offered more variation, these were used instead of the .csv values.

Castiglione created this file from raster data distributed jointly by the National Centers for Environmental Information, National Environmental Satellite, Data, and Information Service, and NOAA (Figure 4). This bathymetry data of Lake Ontario was originally collected for nautical navigation. US data came from the US National Ocean Service's (NOS) Hydrographic Survey Data, the NOS Coast Survey, and the US Army Corps of Engineers. Surveyors collected data at one meter intervals for the first ten meters of depth and two meter intervals at depths

greater than ten meters. The Canadian waters were collected using sounding measurements collected by the Canadian Hydrographic Service for one meter intervals. While the data supplied is not meant to be used for navigation, it is considered high quality for planning and modeling purposes (National Geophysical Data Center 1999).

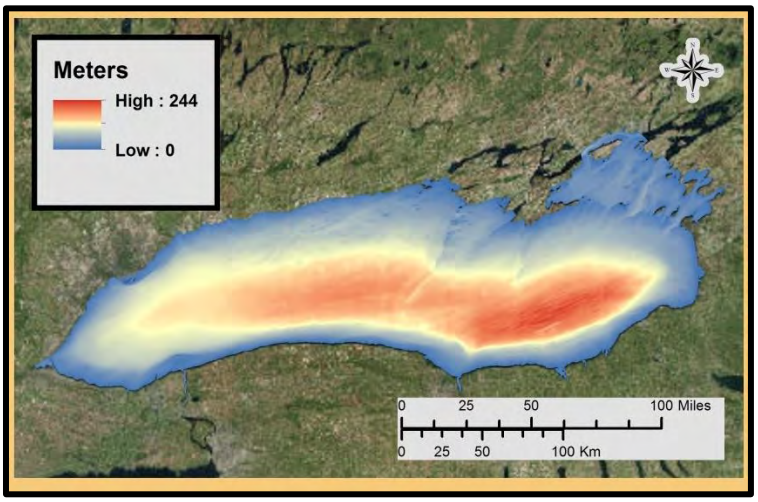

Figure 4 Depth Raster for Lake Ontario

3.2.3. Effective fetch

Effective fetch (Fetch) was included in the Castiglione data and shows the uninterrupted distance that the wind can travel over water (Figure 5). Effective fetch was calculated using the recommended procedure of the Shore Protection Manual (USACE 1984). Fetch is a component in wave action and upwelling. Because of the large range of values in fetch, a $\log_{10}$ transformation (Fetch_Log) was created as an alternative variable to the non-transformed fetch in the modeling process.

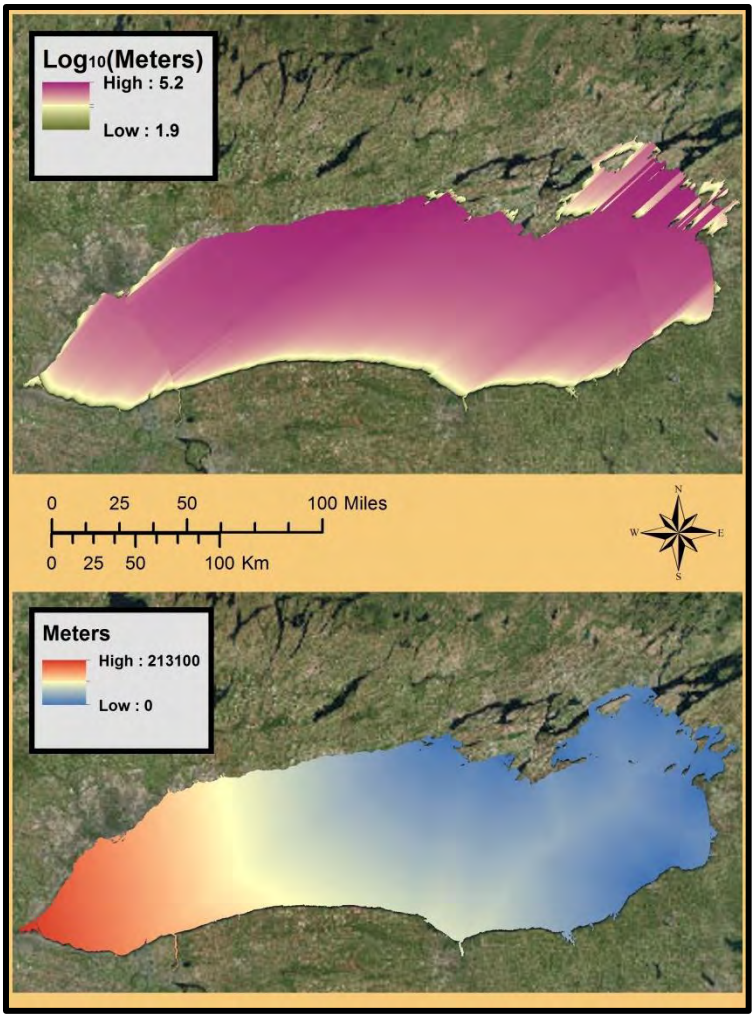

Figure 5 Fetch Raster for Lake Ontario, with and without transformation

3.2.4. Distance to major river mouth

The distance to major river mouth (RivDist) raster was created by calculating Euclidean distance from major river mouths (Figure 6). Major rivers are those of Strahler order 4 or larger (McKenna and Castiglione 2010). Rivers can be the source of nutrients for primary productivity as well as habitat for spawning. Similar to fetch, because of the large range of values in distance to a major river mouth, a $\log_{10}$ transformation (RivDist_Log) was also created as an alternative predictor variable.

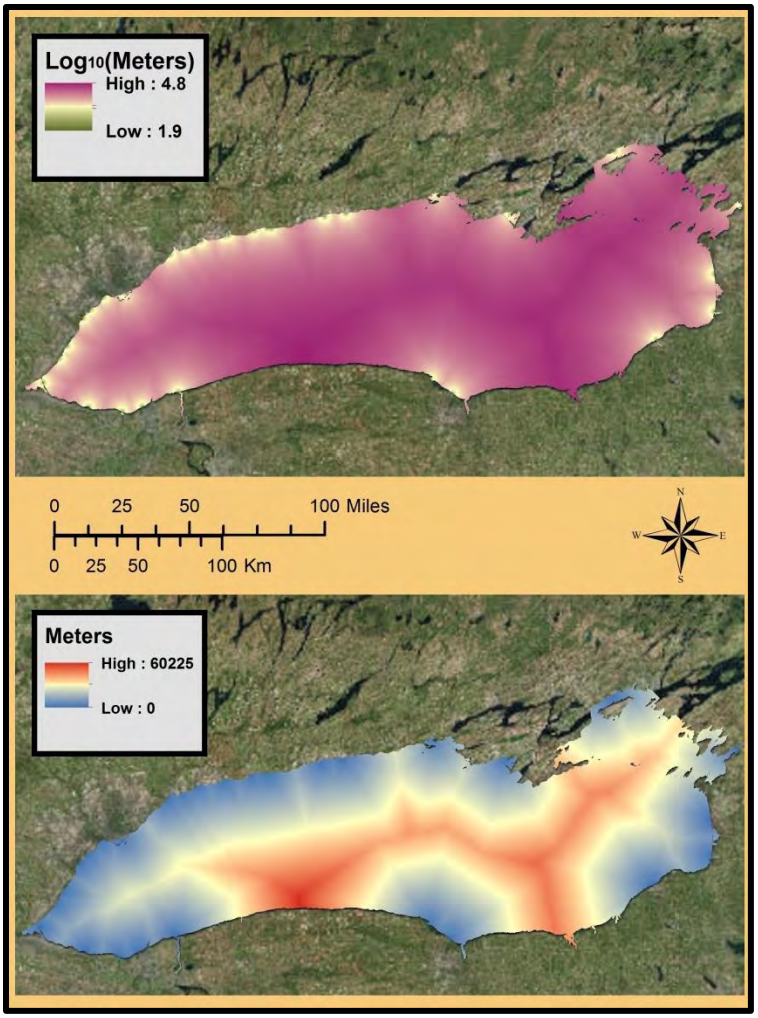

Figure 6 Distance to Major River Mouth Raster for Lake Ontario,
with and without transformation

3.2.5. Distance to delta type wetland

The distance to delta type wetland (DeltaDist) raster was created by calculating Euclidean distance from delta type wetlands (Figure 7). Delta type wetlands are wetlands that extend out into Lake Ontario and are formed primarily of alluvial materials (GLC 2004). Wetlands can play a key role in fish species distribution by offering structural habitat or spawning areas. Because of the large range of values in distance to delta type wetland, a $\log_{10}$ transformation (DeltaDist_Log) again was prepared as an alternative predictor variable to the non-transformed distance to delta type wetland.

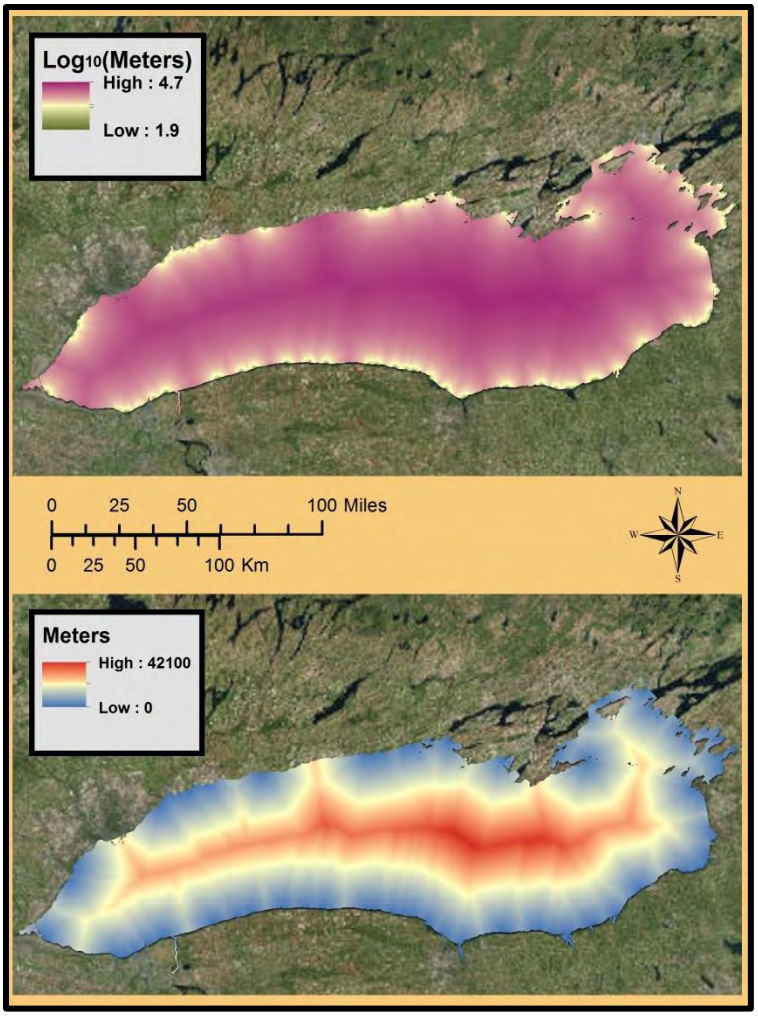

Figure 7 Distance to Delta Type Wetland Raster for Lake Ontario,
with and without transformation

3.2.6. Distance to protected type wetland

The distance to protected type wetland (ProtDist) raster was created by calculating Euclidean distance from protected wetlands (Figure 8). Protected wetlands are wetlands that have increased protection due to bay or sand-spit formations. This increased protection can cause an increase in sediment accumulation making them shallower and heavier in vegetation (GLC 2004). As was done with other variables with large ranges of values, a $\log_{10}$ transformation (ProtDist_Log) was created, to be used as an alternative to the non-transformed distance to protected type wetland.

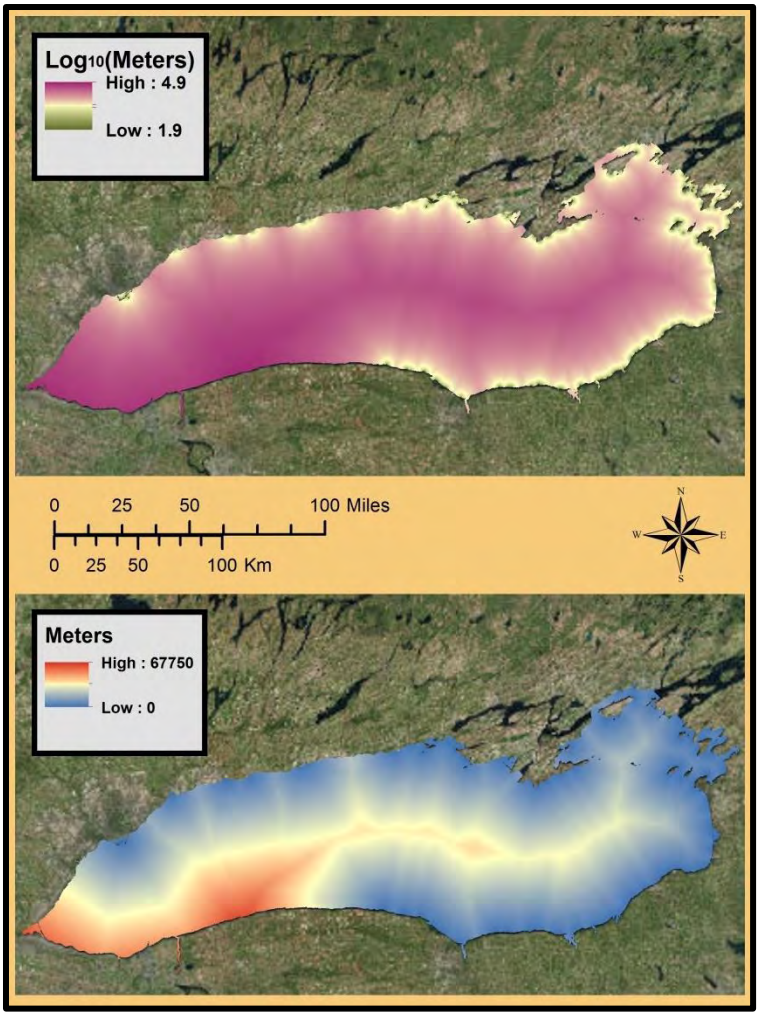

Figure 8 Distance to Protected Type Wetland Raster for Lake Ontario, with and without transformation

3.2.7. Distance to open type wetland

The distance to open type wetland (OpenDist) raster was created by calculating Euclidean distance from open wetlands (Figure 9). Open type wetlands are wetlands that are directly exposed to the nearshore processes, a situation that results in little sediment accumulation as well as scarcer vegetation generally located closer to the shore (GLC 2004). A $\log_{10}$ transformation (OpenDist_Log) alternative to the original distance to open type wetland raster was also created.

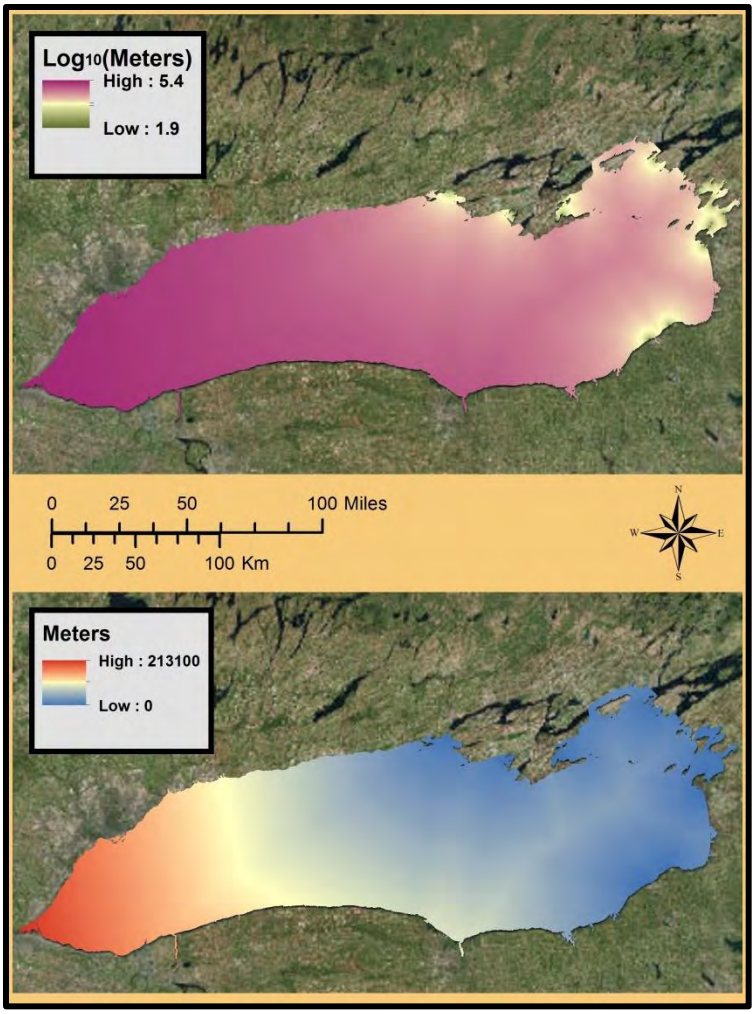

Figure 9 Distance to Open Type Wetland Raster for Lake Ontario,
with and without transformation

3.3 Temporal Data

Year and month variables were included for consideration to account for temporal trends. The YEAR field already existed in the original .csv from the USGS and was included in the created feature class to show the year in which each trawling event occurred. A MONTH field was added to the feature class and the numeric value of the month calculated from the Julian date provided in the original OP_DATE field. The GLM and GAM tools included in the Marine Geospatial Ecology Tools toolbox allow for the use of categorical data so the categorical value of MONTH is sufficient. The GWR tool included in the Spatial Analysis extension, however,

does not allow for categorical data. So, in order to include month into the model, dummy binary fields had to be added for each month (1 if it is the labelled month, 0 if it is not). Month was included so that seasonal movements within the lake could be accounted for. Year was included to help assist in modeling for long term trends that are not yet understood or as a proxy for data that is not available.

3.4 Summary

The inclusion of transformed data as predictor variables allowed for alternatives to be used in the case that the non-transformed variables does not perform well in a GLM, GAM, or GWR. Transformations were only done with species densities and habitat variables that had values that varied by magnitudes. Variables like temperature and depth were not transformed because the ranges of values were small. As is shown in the results discussed later, the choice of which transformed and non-transformed variables were used was dependent on the species and dataset modeled. Having outlined the variables used in the models, the next chapter turns to a detailed discussion of the modeling methods used and the statistics employed to assess their effectiveness.

Chapter 4 Methods

This study used the Marine Geospatial Ecology Tools (MGET) suite developed by Duke University to incorporate the R statistical program into ArcGIS so that GLM and GAM could be performed. The Exploratory Regression and Geographically Weighted Regression tools from Esri's ArcGIS Spatial Analysis extension were used to determine the best predictors and development of the GWR model. This chapter describes the procedure used to run the GLM, GAM, and GWR model tools.

4.1 Software Requirements

To run all necessary analysis, the statistical R (version 2.15.2) program, MGET (version 0.8a60), and ArcGIS were required. While MGET is used as the interface that utilizes R within ArcGIS, the statistical program must be installed on the computer. ArcGIS Advanced license version 10.2.2 was used with the Spatial Analysis extension activated.

4.2 Procedure

The procedure for this study required multiple GLM, GAM, and GWR models to be developed for each species using different datasets. The use of four different datasets created from different year ranges allowed for an investigation to see if a specific time period modeled better than others. Another key model component that was varied was the choice of Gaussian or Poisson distributions. This section discusses each of the modeling methods separately and outlines the required inputs for the tools to run as well as the outputs that are created.

4.2.1. Generalized Linear Model

For a valid GLM model to be developed the following characteristics must be met:

1) The relationship is defined by the link function between the mean of the response and the linear combination of predictor variables.
2) Data is assumed to belong to one of several distribution families including normal, binomial, Poisson, negative binomial, or gamma.
3) The predictor variables used to develop the model will be statistically significant.
4) Coefficients of predictor variables reflect an expected or justifiable relationship with the response variable.
5) There will be no redundancy in the predictor variables.

The GLM models for each of the ten fish species were developed using the MGET tool called Fit GLM. This tool requires an input table that contains the response variable, continuous predictors, and categorical predictors. The tool also requires that the response variable distribution be assigned. The Fit GLM tool also has a feature for automated model selection. By choosing the stepwise backward option, the tool runs iteratively, dropping terms from the original selection and calculating an AIC value for each result. The model with the lowest AIC value is determined to be the best model.

For each run, the input table was one of the four datasets created and described above. The response variable used was the one of the species coded fields described in the data chapter. The response variable distribution was run using the Gaussian method initially with all the variables included. The stepwise backward feature was used to reduce the number of variables used to five. The Fit GLM tool was redone for each species and dataset using the Poisson distribution. The Gaussian distribution used an identity link and the Poisson distribution used a log function link. The output for each Fit GLM tool run was a .Rdata file that contains the

regression equation and was used to calculate prediction values in the next tool. The .Rdata file is R statistics program file that can also be used with the MGET tool Predict GLM From Table.

After all of the models were run, a new blank field was created in each of the four time period feature classes named to hold the prediction values for each of the models that were developed. The fields were named with the species code and the suffix _GLMG (e.g. ALEW_GLMG) for models using the Gaussian distribution and _GLMP (e.g. ALEW_GLMP) for models using the Poisson distribution.

Finally, the Predict GLM From Table was used to populate the newly created fields with the species code and suffix _GLMP and _GLMG (i.e. ALEW_GLMP). This was done by using the .Rdata files created from the Fit GLM tool mentioned above and the attribute table from the trawling event feature classes that were used to develop the model. The species response variable field found in the trawling event feature class attribute table was used as the observed field and the corresponding _GLMG or _GLMP field, depending on the distribution used, was set as the field to receive the predicted value for each trawling event in the feature class. The output of this tool was populated prediction field and R^2 value displayed in the geoprocessing window. This was done for each model run in each of the time period datasets.

4.2.2. Generalized Additive Models

For a valid GAM model to be developed the following characteristics must be met:

1) Functions are additive and the components are smooth.
2) No redundancy in predictor variables.

The GAM models for each of the ten fish species were done using the MGET tool Fit GAM. Like the GLM tool, this tool required an input table that contained the response variable, continuous predictors, and categorical predictors. The tool also requires that the response

variable distribution be assigned as well as the R package that is to be used. Unlike the Fit GLM tool that used a single R package for GLM, the Fit GAM tool had two packages to choose from. One is an older R package called gam, which allowed splines or loess functions to be used, and a newer R package called mgcv, which allowed for a variety of splines to be used.

For each run, the input table was one of the four datasets created and described above. The response variable used was the one of the species coded fields described in the data chapter. The response variable distribution was run using the Gaussian method initially with the variables used in that species corresponding GLM model. All continuous predictor variables were set to use a thin plate regression spline with shrinkage and a maximum of three degrees of freedom. The shrinkage function would reduce any variable that was found to be insignificant to zero to prevent it from heavily impacting the model. The maximum number of degrees of freedom for the spline states how many curves the function is allowed. The setting of three degrees of freedom allowed for up to two curves.

Then the Fit GAM tool was redone for each species and dataset using the Poisson distribution. The Gaussian distribution used an identity link and the Poisson distribution used a log function link. The output for each Fit GLM tool run was a .Rdata file that contains the regression equation and was used to calculate prediction values. The .Rdata file is R statistics program file that can also be used with the MGET tool Predict GLM From Table. The same smoothing function and maximum degrees of freedom were used as the GAM with Gaussian distribution. The output for each Fit GAM tool was a .Rdata file that contains the regression equation and was used to calculate prediction values in the next tool. The .Rdata file is R statistics program file that can also be used with the MGET tool Predict GAM From Table. Again, new blank fields had to be created so prediction values could be added to the trawling

event feature classes. The fields were named with the species code and the suffix _GAMG (e.g. ALEW_GAMG) for models using the Gaussian distribution and _GAMP (e.g. ALEW_GAMP) for models using the Poisson distribution.

Finally, the Predict GAM From Table was used to populate the newly created fields with the species code and suffix _GAMP and _GAMG (i.e. ALEW_GAMP). This was done by using the .Rdata files created from the Fit GAM tool mentioned above and the attribute table from the trawling event feature classes that were used to develop the model. The species response variable field found in the trawling event feature class attribute table was used as the observed field and the corresponding _GAMG or _GAMP field, depending on the distribution used, was set as the field to receive the predicted value for each trawling event in the feature class. The output of this tool was populated prediction field and R^2 value displayed in the geoprocessing window. This was done for each model run in each of the time period datasets.

4.2.3. Geographically Weighted Regression

For a valid GWR model to be developed the following characteristics must be met:

1) Statistically significant coefficients of predictor variables.
2) No redundancy in predictor variables.
3) Normally distributed residuals.
4) Over and under predictions are randomly distributed spatially.

To determine which variables would be best to use for a GWR, the Spatial Analysis extension tool Exploratory Regression was used to determine the best OLS model that had a significant Koenker Statistic indicating that the response is non-stationary and could do better with a GWR. When the Exploratory Regression was finished running, the best models were

identified and they were then rerun using the Spatial Analysis extension tool Geographically Weighted Regression.

The Geographically Weighted Regression tool requires an input shapefile that has both the response and predictor variables. A kernel type was set to adaptive so that when feature distribution is dense the spatial context would be smaller, and if the feature distribution is not as dense the spatial context would be larger. The bandwidth method was set to AICc so that the kernel extent would be determined by AICc. The optional parameter to save the intercept and coefficients for variables as surface rasters was set. While the GLM and GAM tools had the ability to change the distribution family, the GWR tool included in ArcGIS only allows for a Gaussian distribution.

The output for the Geographically Weighted Regression tool is a table with the number of neighbors used for the GWR, the residual squares, effective number, sigma value, AICc, R^2, adjusted R^2, the response variable used, and the list of predictor variables used. A feature class is also created with fields that display the observed response value, condition number, local R^2, predicted value, coefficient fields for intercept and predictor variables, residual values, standardized residuals, standard error, and standard error fields for intercept and predictor variables. As noted above, optional surface rasters can be created for the extent of the feature class for the intercept and predictor variable coefficients.

4.2.4. Predictor variable significance and multicollinearity

Predictor variables were checked during model development to determine if the modeling tool found them significant. If a variable received a p-value of 0.05 or less, the variable was determined to be significant. The GAM tools offered a variety of smoothing functions, which

were selected to be used, that limited insignificant variables, if present, by reducing the coefficient value to zero.

Multicollinearity was determined by the variance inflation factor (VIF) value that was calculated by each tool. The cutoff value was set to 7.5, which falls within the suggested five to ten range of cutoff values (Craney & Surles 2002).

4.2.5. Model comparison

Several methods were used to compare models and assess their success. All models were compared to each other using Pearson's adjusted R^2. The R^2 value is a statistical measure of how observed data fits the regression line. For this analysis, a R^2 value for the fit between the predicted and observed data informed how well the model was able to predict actual values. However, R^2 does not take into account the number of independent (predictor) variables in the creation of the model. Hence, adjusted R^2 which does take the number of independent variables into account is used here. While it is arbitrary, a value of 0.70 or better was the mark set here to indicate a high adjusted R^2 value. The 0.70 value was selected based on the work of previous researchers. Nishida and Chen (2004) is one of the few studies that predicted for abundances rather than just presence and absence. They concluded that their successful models had adjusted R^2 higher than 0.70.

While AIC has commonly been used in past studies to compare the results of different regressions against each other, there is little research on whether it is appropriate to use AIC to compare models that use different distribution families. Because of this lack of past research, AIC comparison was not chosen to be a major indictor to distinguish model performance between methods. While the equation for AIC is the same for both Gaussian and Poisson, the method by which log-likelihood is calculated is different and could lead to the inability to

compare Gaussian models with Poisson models (Steenbergen 2012). Even though not a major indicator, AIC was calculated using R and the GWR tool from ArcGIS to compare the models that used the same distribution family. Unlike R^2, the value of AIC cannot be interpreted directly because of the variation in constants as well as the influence of the sample size. So rather than use the actual value, the change in AIC (ΔAIC) was used (Burnham and Anderson 2004).

The models were also compared using Cohen's Kappa values for different abundance categories. Cohen's Kappa is a statistic that measures the agreement of two categorical items. In this study the agreement that was measured was between the observed and predicted values for a variety of abundance classes. Cohen's Kappa was used because a model that has the accuracy to predict the exact number of fish to be found at a location is highly unlikely. What is more likely is the development of a model that can accurately predict a value that falls within an abundance category. The abundance categories that were used to determine Cohen's Kappa values was presence and absence (0 or >0), low abundance (1 - 10 individuals or not), moderate abundance (10 - 1,000 individuals or not), and high abundance (>1,000 individuals or not).

Cohen's Kappa was calculated using a cross verification table for each abundance class (Table 4). The table was populated with observed values in columns and predicted values in the rows. The frequency on whether the predicted or observed values fell within the abundance class range was compared between each other.

Table 4 Cohen's Kappa cross verification table

Abundance Class		Observed Abundance		
		Yes	No	
Predicted Abundance	Yes	Both fell within abundance class frequency ($Agree_{Yes}$)	Only predicted fell within abundance class frequency ($Disagree_{YesNo}$)	Total number of times predicted fell within abundance class ($Pred_{YesTotal}$)
	No	Only observed fell within abundance class frequency ($Disagree_{NoYes)}$	Neither fell within abundance class frequency ($Agree_{No}$)	Total number of times predicted did not fall within abundance class ($Pred_{NoTotal}$)
		Total number of times observed fell within abundance class ($Obs_{YesTotal}$)	Total number of times observed did not fall within abundance class ($Obs_{NoTotal}$)	Total number of events ($Events_{Total}$)

The cross verification table was used to calculate the observed agreement as seen in Equation 2. Observed agreement is the proportion where observed and predicted abundance values both either fall within the abundance class range or that neither of them falls within the abundance class range. Expected agreement, which is the proportion of agreement that is expected to occur by chance, is then calculated by using Equation 3. Finally, Cohen's Kappa was calculated using Equation 4.

$$Observed\ Agreement = \frac{(Agree_{Yes} + Agree_{No})}{Events_{Total}} \tag{2}$$

$$Expected\ Agreement = \left[\frac{Pred_{YesTotal}}{Events_{Total}} \times \frac{Obs_{YesTotal}}{Events_{Total}}\right] + \left[\frac{Pred_{NoTotal}}{Events_{Total}} \times \frac{Obs_{NoTotal}}{Events_{Total}}\right] \tag{3}$$

$$Cohen's\ Kappa = \frac{Observed\ Agreement - Expected\ Agreement}{1 - Expected\ Agreement} \tag{4}$$

Cohen's Kappa values were used to determine if a specific model was better at predicting one abundance category over another. While there is no general standard for assessing the significance of Cohen's Kappa values, many authors have used the categorization that is displayed in Table 5 (see for example, Landis and Koch 1977, McKenna and Castiglione 2014). The ranking system displayed in Table 5 was used for determining Cohen's Kappa significance in this study.

Table 5 Cohen's Kappa value agreement ranking

Cohen's Kappa Value	Agreement Ranking
<0.01	No agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
>0.80	Almost perfect agreement

For this study, a good model is one that obtained an adjusted R^2 value of 0.70 or greater with Cohen's Kappa values for moderate or better agreement in each abundance category. An adjusted R^2 of 0.70 or better would show high correlation between the observed and predicted abundances. A moderate agreement in each abundance category would indicate a model that can moderately predict any abundance class, rather than be biased to predicting only lower or higher abundances. Additional comparisons using AIC values and a qualitative assessment of model structure added further insight.

4.3 Summary

Using four different input datasets, the MGET GLM and GAM tools were used to develop 120 models with the best combination of five predictor variables for each of the two distribution types. The Esri tools for developing the 30 GWR models were able to determine the best combination of predictor variables to be used. Indicators that were used to compare the models were adjusted R^2 and Cohen's Kappa values. Additional insights were obtained by examining AIC values and model structure. The model results are discussed and compared in the next chapter.

Chapter 5 Model Results

The MGET tools were able to run without any errors using the all the datasets and for all the species. However, as described below, the Geographically Weighted Regression tool produced errors when attempting to model Threespine Stickleback for any of the datasets. The success of the GLM and GAM tools and the majority of the successes for the GWR tool produced enough results for comparisons to be made for the three modeling methods.

5.1 Generalized Linear Model

The MGET GLM toolset was able to produce results for each species and all the datasets. Overall results were poor for each model, the highest achieving only an adjusted R^2 value of 0.63 for Spottail Shiner in the 1978 - 1989 dataset. The Poisson distribution was able to achieve a better R^2 value for a majority of the models. Cohen's Kappa values were dependent on the distribution type used for the model. The Gaussian distribution performed better at predicting presence and absence whereas the Poisson distribution performed better with the moderate and high abundance classes. Neither distribution type was able to get any Kappa value agreement ranking between observed and predicted values above slight agreement.

5.1.1. 1978 - 2014 dataset

The 1978 - 2014 dataset yielded overall poor results for both the Gaussian and Poisson distribution GLMs (Table 6). Differences between the two distribution types were small, >0.1, except for Trout Perch which saw a decrease, of 0.12, in adjusted R^2 value when run with a Poisson distribution. Trout Perch (0.26) was the only species modeled with GLM that could explain at least 25% of the response. The Trout Perch GLM used Log_{10} transformed distance to

open type wetland, effective fetch, and the square root of Rainbow Smelt, Yellow Perch, and Spottail Shiner abundances as predictor variables.

Table 6 Adjusted R^2 values for GLMs for each species (1978-2014). GLMG uses Gaussian distribution and GLMP uses Poisson distribution. **Bolded** values are the higher between Gaussian and Poisson.

Species	GLMG Adj R^2	GLMP Adj R^2	Difference (GLMG-GLMP)
ALEW	**0.09**	0.07	0.02
GOBY	0.05	**0.10**	-0.05
JOHN	0.04	**0.04**	0
LTRT	**0.10**	0.08	0.02
PRCH	0.08	**0.14**	-0.06
SLIM	0.15	**0.21**	-0.06
SMLT	0.12	**0.13**	-0.01
SPOT	**0.18**	0.17	0.01
STK3	0.01	**0.03**	-0.02
TRPR	**0.26**	0.14	0.12

The results of the Cohen's Kappa showed that the Gaussian distribution was better at determining presence and absence but not as well as for moderate and high abundances compared to the Poisson distribution (Table 7). Neither distribution type was able to get a fair or higher agreement ranking between the observed and predicted values for the low abundance class. Trout Perch which had the highest adjusted R^2 using the Gaussian distribution was able to get a fair agreement ranking for the presence and absence classification as well as a moderate agreement ranking for the high abundance class, but was only able to get a slight agreement ranking for the low and moderate abundance classes.

The standardized residuals versus the fitted values for the Trout Perch GLM using a Gaussian distribution showed that there is a coned shaped pattern indicating that there is not homogeneity of the variance. There is more dispersal of standardized residuals for the higher values with lower values being more clustered. The QQ - Plot also indicates that the model did not meet the assumptions (Figure 10). When the Trout Perch standardized residuals were mapped

the highest (≥ 1.5) and lowest (≤ -1.5) deviation from the mean were more heavily distributed in the eastern portion of the lake (Figure 11).

Table 7 Cohen's Kappa values for GLMs for each species (1978-2014). G = Gaussian, P = Poisson, (*) denotes fair agreement, (**) denotes moderate agreement. **Bolded** values are the higher between Gaussian and Poisson.

Species	Presence		Low Abundance		Moderate Abundance		High Abundance		When Species Present	
	G	P	G	P	G	P	G	P	Range of abundances	Average abundance
ALEW	**0.33***	<0.01	<0.01	<0.01	**0.05**	<0.01	0.24*	**0.31***	1 - 124,648	2,438
GOBY	**0.15**	0.11	<0.01	<0.01	0.10	**0.27***	<0.01	**0.25***	1 - 13,076	215
JOHN	**0.12**	0.07	**0.03**	<0.01	0.14	**0.26***	<0.01	<0.01	1 - 10,935	71
LTRT	**0.16**	0.01	**0.06**	<0.01	**0.30***	0.29*	NA	NA	1 - 732	9
PRCH	0.08	**0.22***	<0.01	**0.08**	0.21*	**0.43****	<0.01	<0.01	1 - 2,664	56
SLIM	**0.35***	0.03	**0.02**	<0.01	**0.23***	0.06	<0.01	**0.30***	1 - 11,595	224
SMLT	**0.06**	<0.01	<0.01	<0.01	0.10	**0.16**	0.32*	**0.41****	1 - 181,082	1,391
SPOT	0.09	**0.14**	<0.01	<0.01	0.17	**0.31***	0.25*	**0.42****	1 - 12,055	246
STK3	**0.05**	0.04	<0.01	<0.01	0.03	**0.13**	<0.01	**0.1**	1 - 16,701	138
TRPR	**0.21***	<0.01	**0.01**	<0.01	**0.14**	0.12	**0.60****	0.41**	1 - 23,917	358

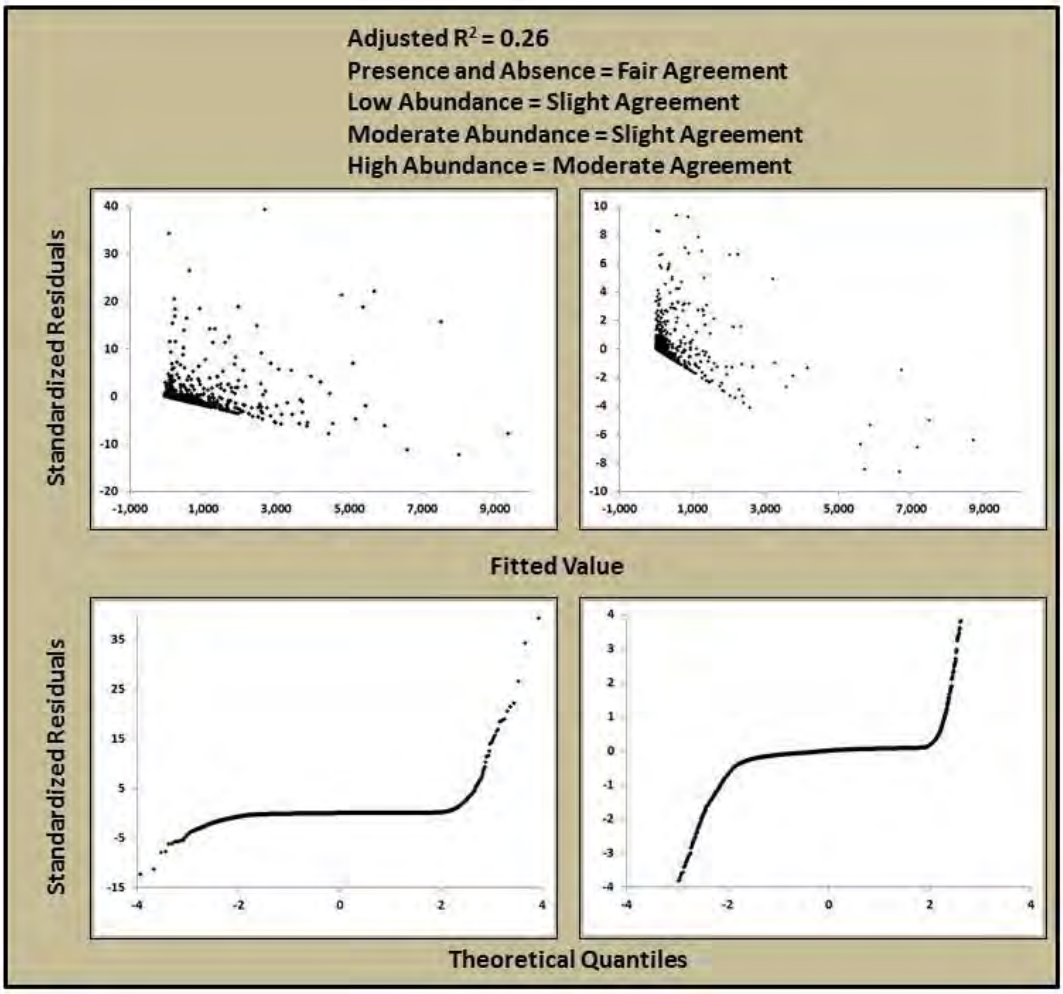

Figure 10 Standardized Residual versus Fitted Value and QQ Plot for Trout Perch GLM (1978 - 2014), with Gaussian Distribution

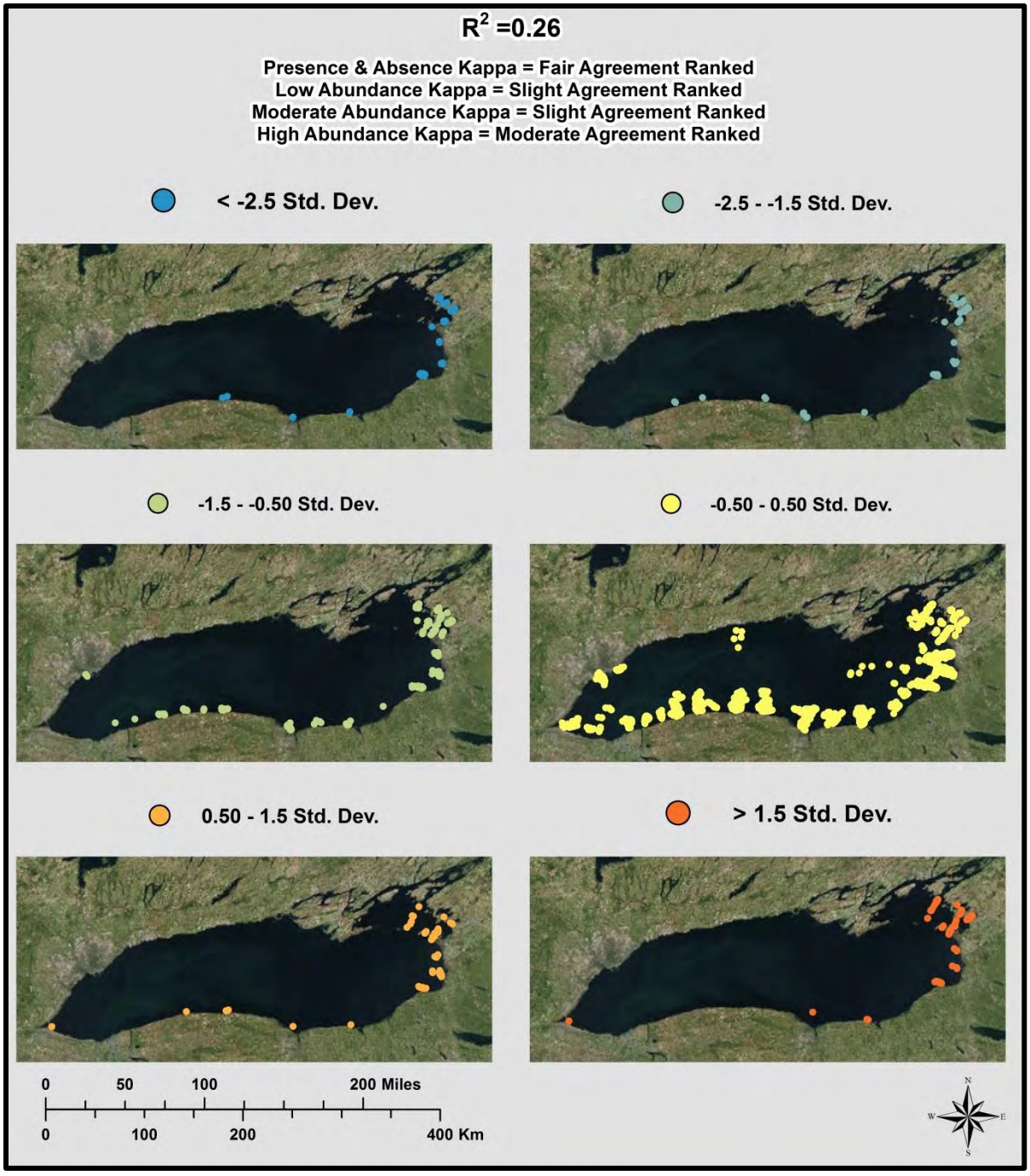

Figure 11 Spatial Distribution of Standardized Residuals for Trout Perch GLM (1978 - 2014), with Gaussian Distribution

5.1.2. 1978 - 1989 dataset

The 1978 - 1989 dataset saw poor adjusted R^2 values, <0.25, for a majority of the species with both distributions (Table 8). The GLMs with Gaussian distribution produced an adjusted R^2

value that explained at least 25% of the response for Spottail Shiner (0.30) and Trout Perch (0.43). The GLMs with Poisson distribution produced higher adjusted R^2 values than the GLMs with Gaussian distribution for Slimy Sculpin (0.36) and Spottail Shiner (0.63). The GLM with Poisson distribution caused a slight decrease in adjusted R^2 for Trout Perch (0.40). The shift from Gaussian to Poisson distribution showed at least a 0.10 increase in adjusted R^2 for three species. The change in distribution more than doubled the adjusted R^2 for Spottail Shiner. The predictor variables used for the Spottail Shiner model was the Log_{10} transformed distance to open type wetland, month, depth, and the square root of Trout Perch and Johnny Darter abundances. The change in distribution family wasn't as large for Johnny Darter or Slimy Sculpin, but the increase did allow for more than 25% response explanation for Slimy Sculpin.

Table 8 Adjusted R^2 values for GLM models for each species (1978-1989). GLMG uses Gaussian distribution and GLMP uses Poisson distribution. **Bolded** values are the higher between Gaussian and Poisson.

Species	GLMG Adj R^2	GLMP Adj R^2	Difference (GLMG-GLMP)
ALEW	0.13	**0.17**	-0.04
JOHN	0.04	**0.18**	-0.14
LTRT	**0.11**	0.08	0.03
PRCH	0.12	**0.16**	-0.04
SLIM	0.20	**0.36**	-0.16
SMLT	0.17	**0.22**	-0.05
SPOT	0.30	**0.63**	-0.33
STK3	0.01	**0.04**	-0.03
TRPR	**0.43**	0.40	0.03

The results of the Cohen's Kappa showed that the GLMs with Gaussian distribution was better at determining presence and absence but not as good as for moderate and high abundances compared to the GLMs with Poisson distribution (Table 9). Neither distribution type was able to get a fair or higher agreement ranking between the observed and predicted values for the low abundance class. Spottail Shiner had the highest adjusted R^2 using the Poisson distribution was

able to get a slight agreement ranking between observed and predicted values for the presence and absence classification as well as a fair agreement ranking for the moderate abundance class, but could not get any agreement for the low abundance classes. The highest rank of agreement was moderate for the high abundance class.

Table 9 Cohen's Kappa values for GLM models for each species (1978-1989). G = Gaussian, P = Poisson, (*) denotes fair agreement, (**) denotes moderate agreement. **Bolded** values are the higher between Gaussian and Poisson.

Species	Presence		Low Abundance		Moderate Abundance		High Abundance		When Species Present	
	G	P	G	P	G	P	G	P	Range of abundances	Average abundance
ALEW	**0.34***	<0.01	0.01	0.01	**0.05**	0.01	0.32*	**0.42****	1 - 114,693	2,745
JOHN	**0.10**	0.01	0.01	0.01	0.15	**0.23***	<0.01	**0.20**	1 - 10,935	64
LTRT	**0.15**	<0.01	<0.01	<0.01	**0.33***	0.31	NA	NA	1 - 431	13
PRCH	0.11	**0.25***	0.03	**0.07**	0.31*	**0.49****	<0.01	<0.01	1 - 2,336	63
SLIM	**0.24***	0.14	**0.02**	<0.01	0.16	**0.19**	0.24*	**0.46****	1 - 8,528	344
SMLT	**0.10**	<0.01	<0.01	**0.01**	0.14	**0.26***	0.35*	**0.44****	1 - 72,261	2,313
SPOT	**0.28***	0.13	**0.01**	<0.01	0.25	**0.38***	0.35*	**0.56****	1 - 7,002	262
STK3	0.02	**0.08**	<0.01	**0.02**	<0.01	<0.01	NA	NA	1 - 91	6
TRPR	**0.19**	0.09	**0.01**	<0.01	0.13	**0.20**	**0.57****	0.53**	1 - 17,612	431

The standardized residuals versus the fitted values for the Spottail Shiner GLM using a Poisson distribution showed that there was a coned shaped pattern indicating that there is not homogeneity of the variance. There was more dispersal of standardized residuals for the higher values with lower values being more clustered. The QQ - Plot also indicates that the model did not meet the assumption of homogeneity of the variance (Figure 12). When the Spottail Shiner standardized residuals were mapped, the lowest (≤ -1.5) deviations below the mean were isolated to the eastern portion of the lake (Figure 13). The highest (≥ 1.5) deviations above the mean were more frequently located in the eastern portion of the lake, but with a few isolated events in the central portion of the lake.

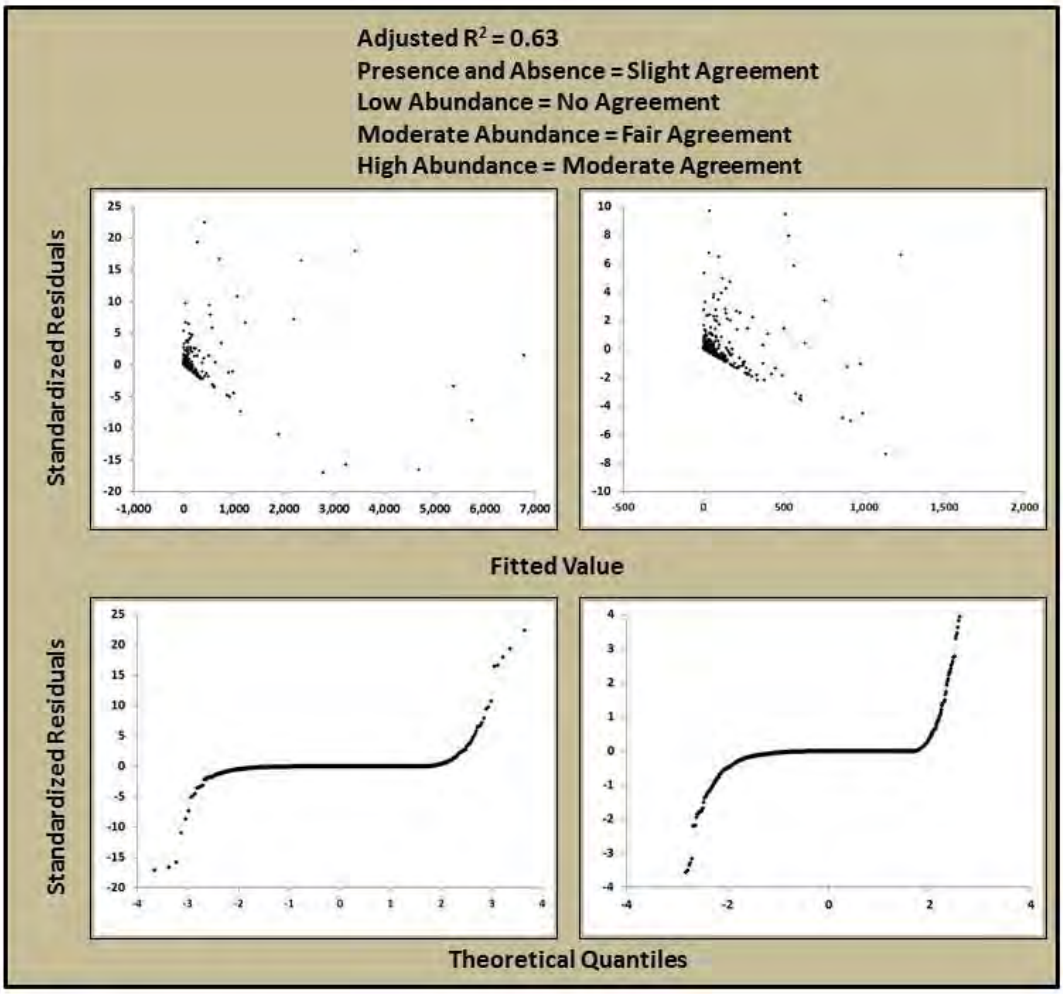

Figure 12 Standardized Residual versus Fitted Value and QQ Plot for Spottail Shiner GLM (1978 - 1989), with Poisson Distribution

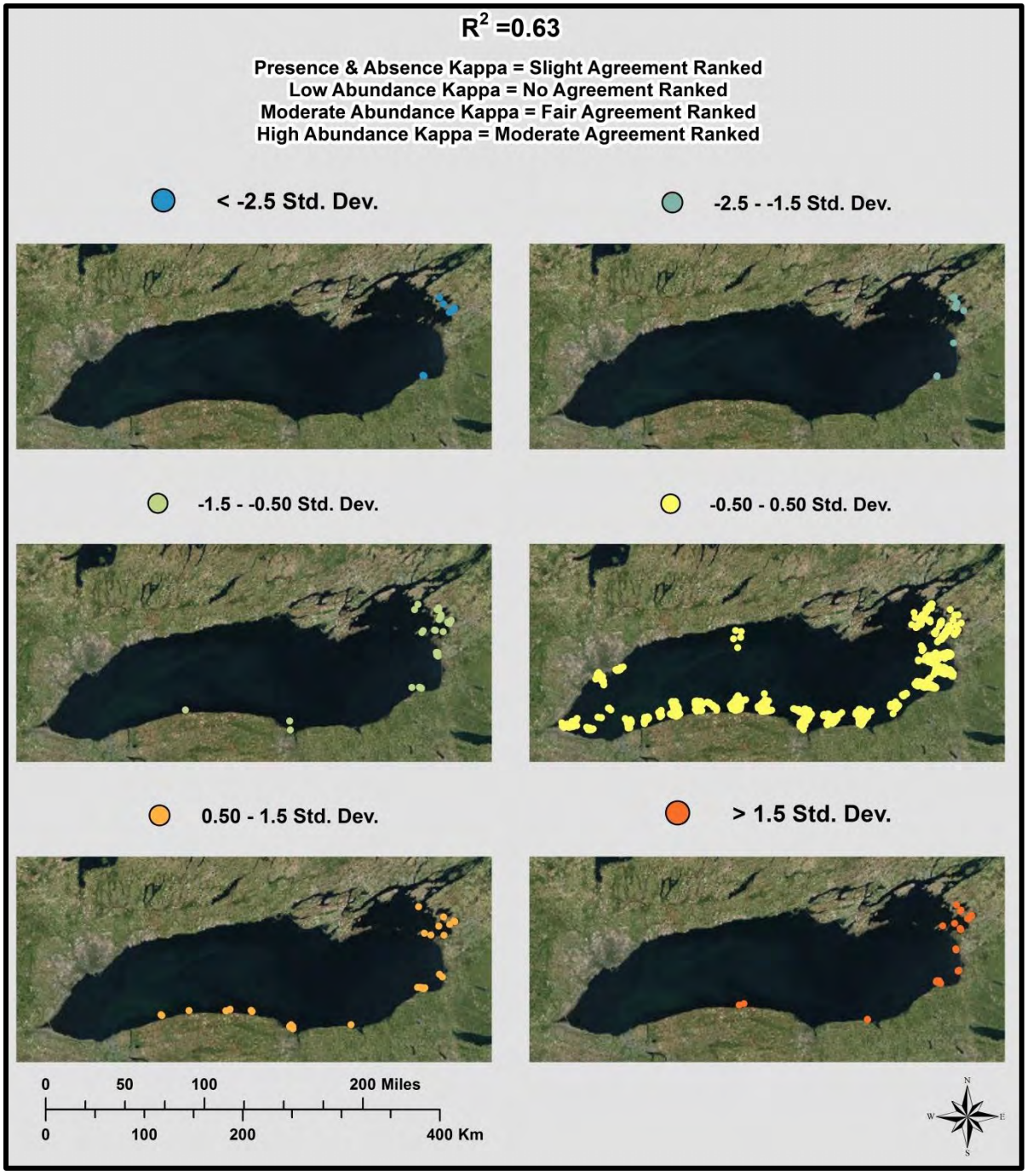

Figure 13 Spatial Distribution of Standardized Residuals for Spottail Shiner GLM (1978 - 1989), with Poisson Distribution

5.1.3. 1990 - 2014 dataset

The 1990 - 2014 dataset saw poor adjusted R^2 values, <0.25, for all but one of the species with both distributions (Table 10). The GLMs with the Gaussian distribution didn't get a single species to produce an adjusted R^2 higher than 0.14. The GLMs with the Poisson distribution was able to produce higher adjusted R^2 values than the Gaussian distribution but only Spottail Shiner was able to get a value greater than 0.25. The shift from Gaussian to Poisson distribution more than tripled the adjusted R^2 value for Spottail Shiner. The predictor variables used for the Spottail Shiner model development was Log_{10} transformed distance to delta type wetland, distance to protected type wetland, month, and the square root of Yellow Perch and Trout Perch abundances.

Table 10 Adjusted R^2 values for GLM models for each species (1990-2014). GLMG uses Gaussian distribution and GLMP uses Poisson distribution. **Bolded** values are the higher between Gaussian and Poisson.

Species	GLMG Adj R^2	GLMP Adj R^2	Difference (GLMG-GLMP)
ALEW	**0.07**	0.06	0.01
GOBY	0.06	**0.10**	-0.04
JOHN	0.04	0.04	0
LTRT	**0.05**	0.04	0.01
PRCH	0.06	**0.15**	-0.09
SLIM	0.14	**0.21**	-0.07
SMLT	**0.07**	0.06	0.01
SPOT	0.09	**0.44**	-0.35
STK3	0.02	**0.04**	-0.02
TRPR	0.12	**0.17**	-0.05

The results of the Cohen's Kappa showed that the Gaussian distribution was only slightly better at determining presence and absence but not as well as for moderate and high abundances compared to the Poisson distribution (Table 11). Neither distribution type was able to get a fair or higher agreement ranking between the observed and predicted values for the low abundance class. Spottail Shiner, which had the highest adjusted R^2 using the Poisson distribution, had

higher Kappa values than the Gaussian distribution for each category. There was a slight agreement ranking for the presence and absence classification and the low abundance class. The moderate abundance class was able to achieve a fair agreement ranking. The highest rank of agreement was for moderate agreement for the high abundance class.

Table 11 Cohen's Kappa values for GLM models for each species (1990-2014). G = Gaussian, P = Poisson, (*) denotes fair agreement, (**) denotes moderate agreement. **Bolded** values are the higher between Gaussian and Poisson.

Species	Presence		Low Abundance		Moderate Abundance		High Abundance		When Species Present	
	G	P	G	P	G	P	G	P	Range of abundances	Average abundance
ALEW	**0.39***	<0.01	0.01	0.01	**0.09**	<0.01	0.23*	**0.27***	1 - 124,648	2,265
GOBY	**0.18**	0.05	<0.01	<0.01	0.12	**0.22***	<0.01	**0.25***	1 - 13,076	215
JOHN	**0.14**	0.13	**0.03**	<0.01	0.17	**0.34***	<0.01	<0.01	1 - 9,103	77
LTRT	**0.11**	0.02	**0.07**	0.01	0.09	**0.15**	NA	NA	1 - 732	6
PRCH	0.07	**0.21***	<0.01	**0.10**	0.14	**0.40***	<0.01	<0.01	1 - 2,664	52
SLIM	**0.39***	0.01	**0.03**	<0.01	**0.27***	0.19	<0.01	**0.19**	1 - 11,595	155
SMLT	**0.06**	<0.01	**0.01**	<0.01	0.07	**0.08**	0.34*	**0.40***	1 - 181,082	895
SPOT	0.04	**0.14**	<0.01	**0.02**	0.08	**0.30***	0.17	**0.44****	1 - 12,055	227
STK3	0.05	**0.09**	<0.01	**0.02**	0.05	**0.15**	<0.01	**0.09**	1 - 16,701	144
TRPR	**0.16**	0.13	<0.01	<0.01	0.11	**0.23***	0.36*	**0.46****	1 - 23,917	287

The standardized residuals versus the fitted values for the Spottail Shiner GLM using a Poisson distribution showed that there is a non-random pattern indicating that there was not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the assumption of homogeneity of the variance (Figure 14). When the Spottail Shiner standardized residuals were mapped the highest (≥ 1.5) deviations above the mean was more heavily distributed in the eastern portion of the lake, with a few isolated locations in the central portion of the lake (Figure 15). The lowest (≤ -1.5) deviations below the mean were also heavily distributed in the eastern portion of the lake with an isolated event in the western portion of the lake.

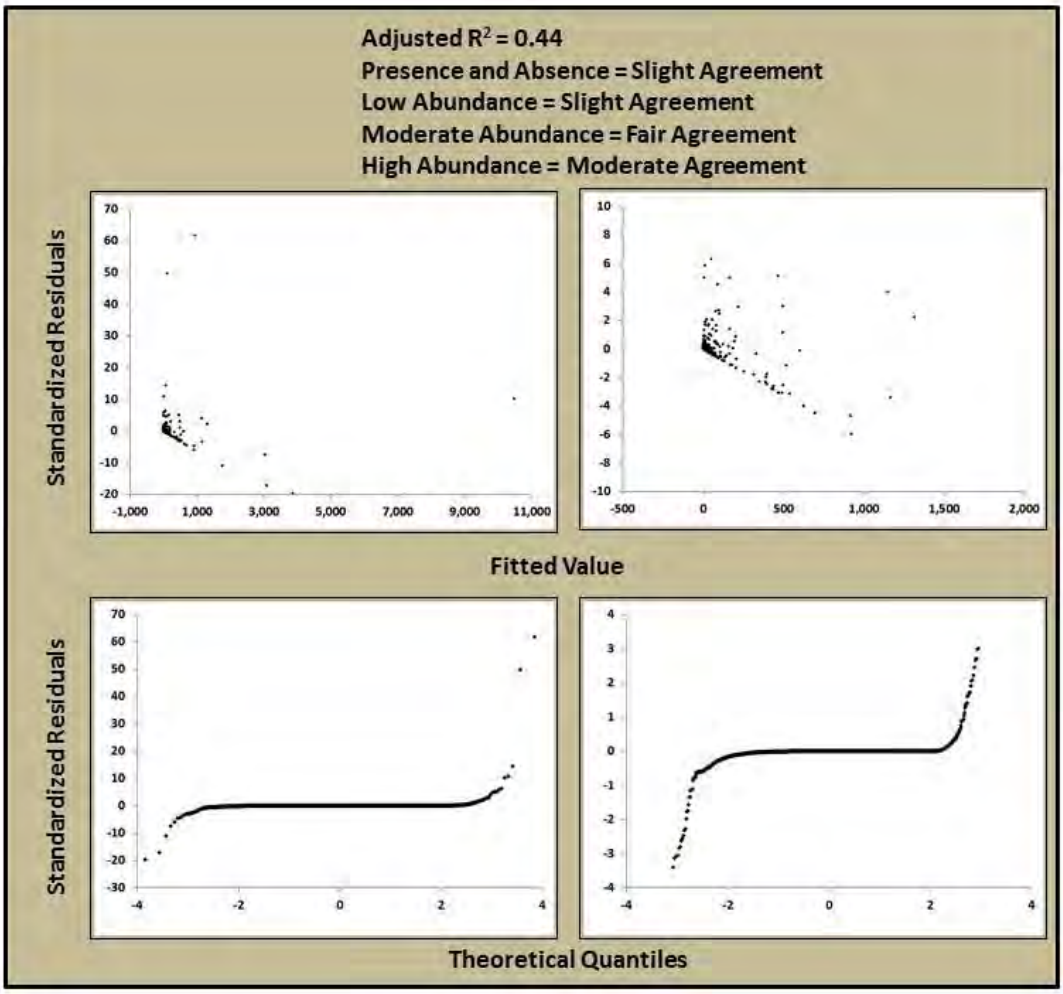

Figure 14 Standardized Residual versus Fitted Value and QQ Plot for Spottail Shiner GLM (1990 - 2014), with Poisson Distribution

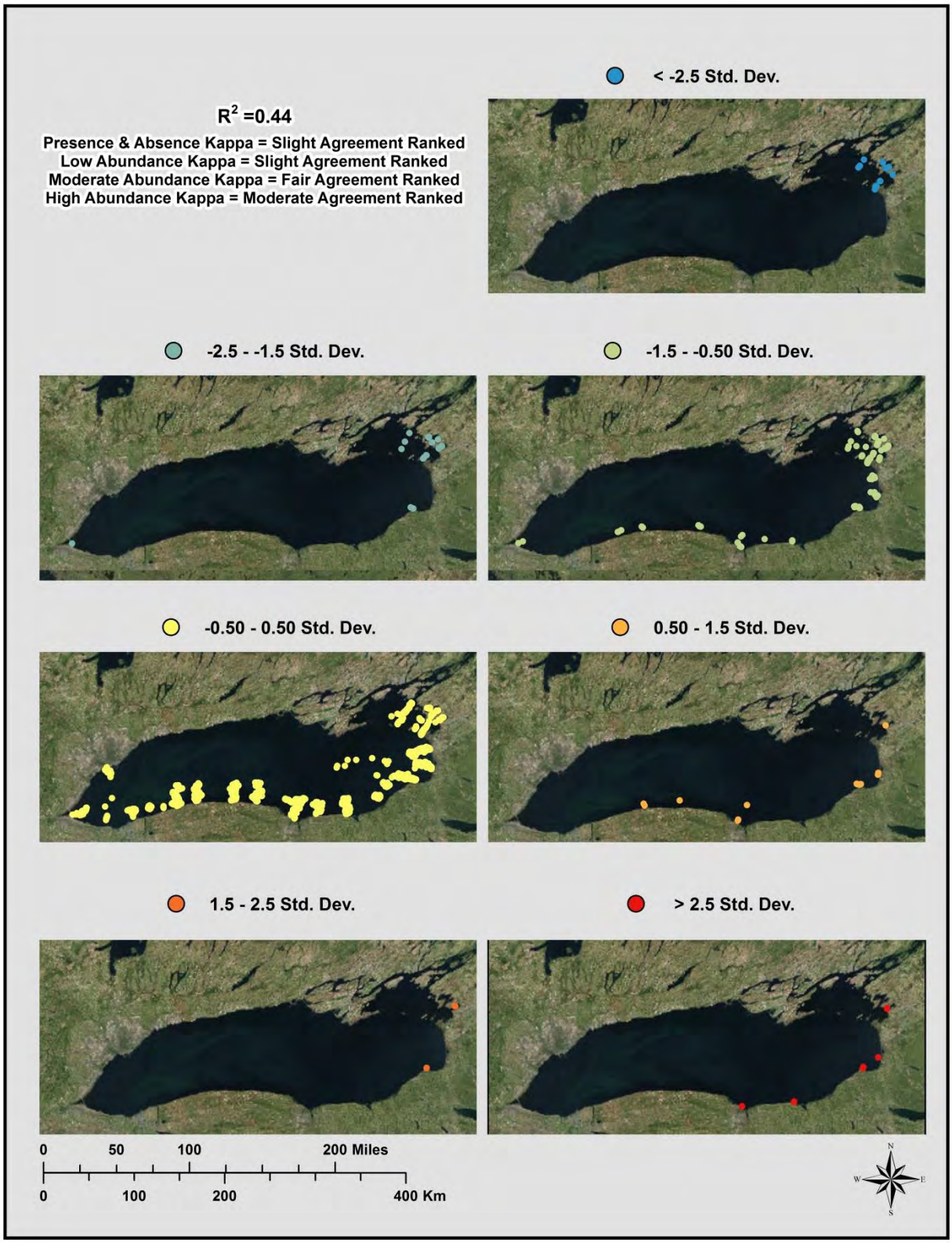

Figure 15 Spatial Distribution of Standardized Residuals for Spottail Shiner GLM (1990 - 2014), with Poisson Distribution

5.1.4. 2004 - 2014 dataset

The 2004 - 2014 dataset was used to model only for Round Goby. The GLM with Gaussian distribution saw a poor adjusted R^2 values while the Poisson distribution was able to produce a higher adjusted R^2 value that more than doubled the adjusted R^2 of the Gaussian distribution (Table 12). The predictor variables used for the Round Goby model development was depth, temperature at fishing depth, distance to protected type wetland, month, and the square root of Alewife abundance. The results of the Cohen's Kappa showed that the GLM with Gaussian distribution was better at predicting presence and absence (Table 13). The GLM with Gaussian distribution could not get any agreement between the observed and predicted values for the high abundances class compared to the GLM with Poisson distribution which was able to account for a fair agreement ranking. Neither distribution was able to get any agreement for the low abundance class.

Table 12 Adjusted R^2 values for GLM models for Round Goby (2004-2014). GLMG uses Gaussian distribution and GLMP uses Poisson distribution. Bolded values are the higher between Gaussian and Poisson.

Species	GLMG Adj R^2	GLMP Adj R^2	Difference (GLMG-GLMP)
GOBY	0.11	**0.26**	-0.15

Table 13 Cohen's Kappa values for GLM models for Round Goby (2004-2014). G = Gaussian, P = Poisson, (*) denotes fair agreement, (**) denotes moderate agreement. Bolded values are the higher between Gaussian and Poisson.

Species	Presence		Low Abundance		Moderate Abundance		High Abundance		When Species Present	
	G	P	G	P	G	P	G	P	Range of abundances	Average abundance
GOBY	**0.21***	0.01	<0.01	<0.01	0.16	**0.18**	<0.01	**0.37***	1 - 13,076	215

The standardized residuals versus the fitted values for the Round Goby GLM using a Poisson distribution showed that there was a non-random pattern indicating that there is not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the assumption of homogeneity of the variance (Figure 16). When the Round Goby standardized residuals were mapped, the highest (≥ 1.5) deviations above the mean were distributed throughout the central and eastern portion of the lake (Figure 17). The lowest (≤ -1.5) deviations below the mean were heavily distributed in the central portion of the lake.

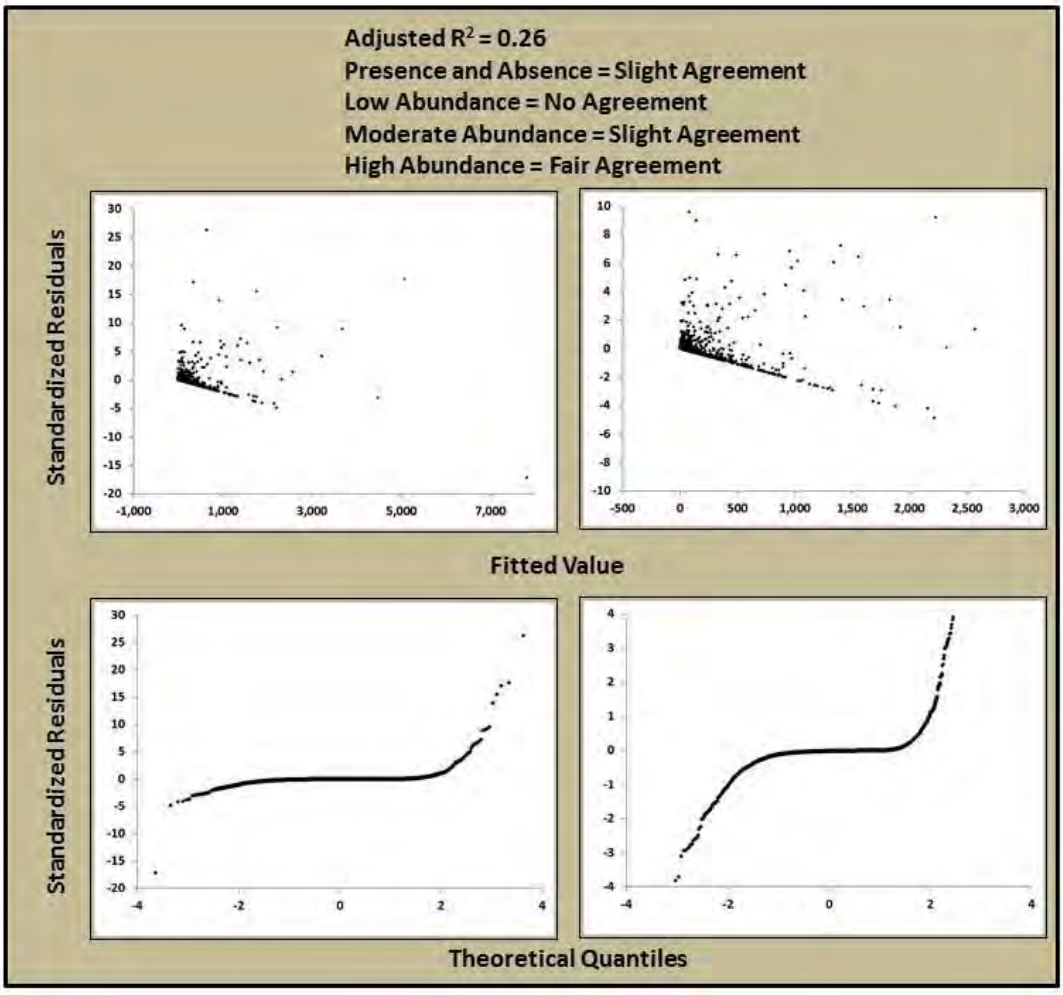

Figure 16 Standardized Residual versus Fitted Value and QQ Plot for Round Goby GLM (2004 - 2014), with Poisson Distribution

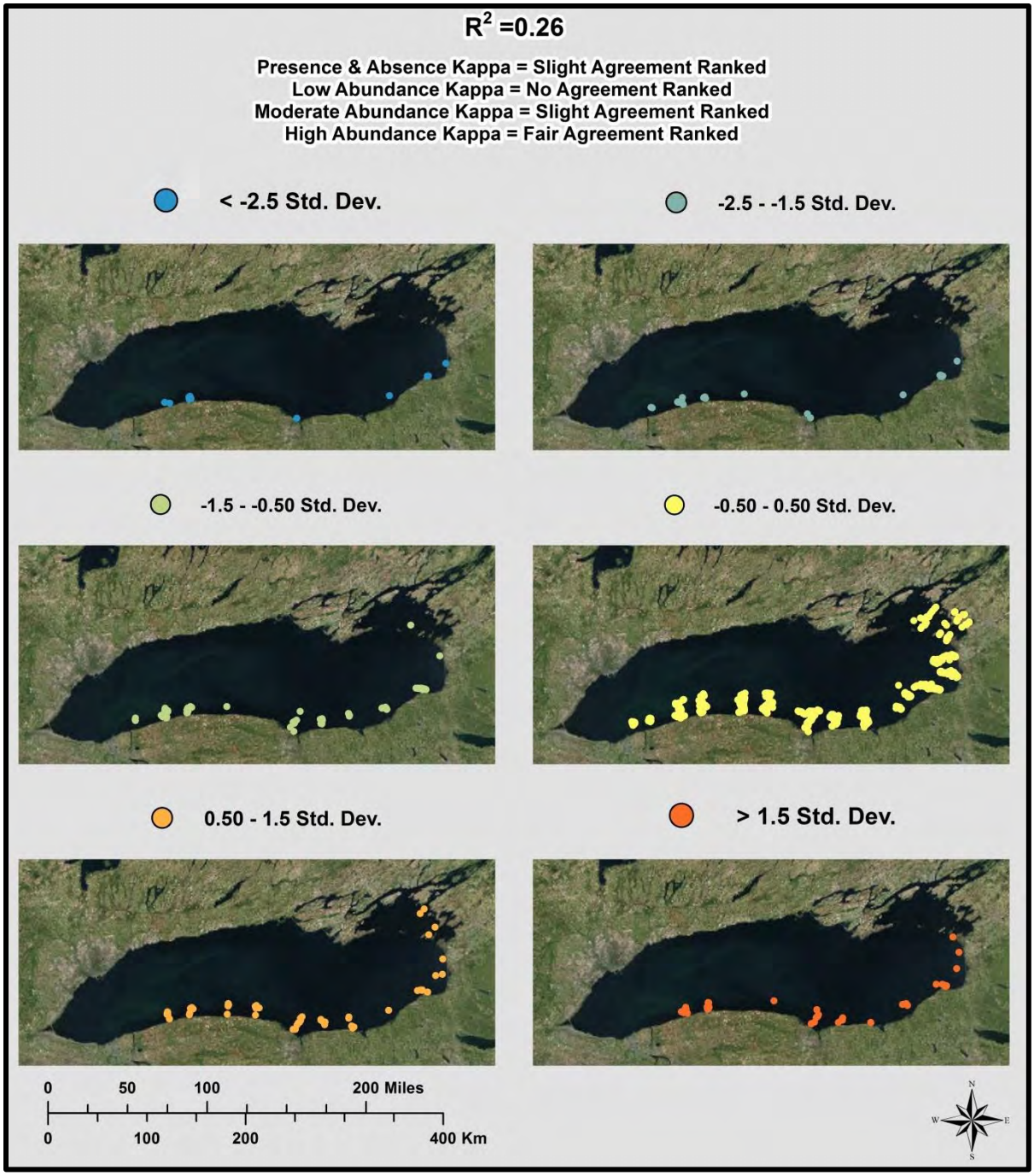

Figure 17 Spatial Distribution of Standardized Residuals for Round Goby GLM (2004-2014), with Poisson Distribution

5.2 Generalized Additive Model

The MGET GAM toolset was able to produce results for each species and all the datasets. Overall results were poor for each model, the highest achieving only an adjusted R^2 value of 0.74

for Spottail Shiner in the 1978 - 1989 dataset. The Poisson distribution was able to achieve better adjusted R^2 values for a majority of the models. Cohen's Kappa values were dependent on the distribution type used for the model. The Gaussian distribution performed better at predicting presence and absence whereas the Poisson distribution performed better with the moderate and high abundance classes. Neither distribution type was able to get any Kappa value agreement ranking between the observed and predicted values above slight agreement.

5.2.1. 1978 - 2014 dataset

The 1978 - 2014 dataset yielded mostly poor results for both the Gaussian and Poisson distribution GAMs (Table 14). Differences between the adjusted R^2 for the two distributions were >0.1 for five of the ten species. The GAMs with Poisson distribution did outperform all GAMs with Gaussian distribution models expect for Trout Perch which had the same adjusted R^2 values for both distributions. The highest adjusted R^2 value was for Round Goby (0.48), with Slimy Sculpin (0.40), Spottail Shiner (0.36), and Trout Perch (0.29) having at least 25% of the response explained. The predictor variables that were used to develop the Spottail Shiner model was the Log_{10} transformed distance to delta type wetlands, depth, month, and the square root of Trout Perch and Johnny Darter. The predictor variables used for the Round Goby model development was depth, temperature at fishing depth, distance to protected type wetland, year, and month.

Table 14 Adjusted R^2 values for GAMs for each species (1978-2014). GAMG uses Gaussian distribution and GAMP uses Poisson distribution. **Bolded** values are the higher between Gaussian and Poisson.

Species	GAMG Adj R^2	GAMP Adj R^2	Difference (GAMG-GAMP)
ALEW	0.10	**0.12**	-0.02
GOBY	0.06	**0.48**	-0.42
JOHN	0.05	**0.17**	-0.12
LTRT	0.11	**0.12**	-0.01
PRCH	0.10	**0.16**	-0.06
SLIM	0.19	**0.40**	-0.21
SMLT	0.14	**0.16**	-0.02
SPOT	0.21	**0.36**	-0.15
STK3	0.02	**0.15**	-0.13
TRPR	0.29	0.29	0

The results of the Cohen's Kappa showed that the Poisson distribution outperformed the Gaussian distribution in determining presence and absence by producing fair agreement ranks for three species and a single moderate agreement ranking for one (Table 15). Neither distribution type was able to get a fair or higher agreement ranking between the observed and predicted values for the low abundance class. The GAM with Poisson distribution did outperform the Gam with Gaussian for the moderate and high abundance classes by producing more fair and moderate agreement rankings between the observed and predicted values. Round Goby had the highest adjusted R^2 also had better Kappa values for each abundance category with the Poisson distribution over the Gaussian distribution. Spottail Shiner also had a higher adjusted R^2 as well as an improvement in all abundance categories using the Poisson distribution.

Table 15 Cohen's Kappa values for GAMs for each species (1978-2014). G = Gaussian, P = Poisson, (*) denotes fair agreement, (**) denotes moderate agreement. **Bolded** values are the higher between Gaussian and Poisson.

Species	Presence		Low Abundance		Moderate Abundance		High Abundance		When Species Present	
	G	P	G	P	G	P	G	P	Range of abundances	Average abundance
ALEW	**0.30***	<0.01	<0.01	<0.01	**0.05**	<0.01	0.23*	**0.33***	1 - 124,648	2,438
GOBY	0.18	**0.42****	<0.01	**0.12**	0.11	**0.37***	<0.01	**0.52****	1 - 13,076	215
JOHN	0.17	**0.18**	0.02	**0.03**	0.19	**0.38***	<0.01	**0.20**	1 - 10,935	71
LTRT	**0.21***	0.03	**0.09**	<0.01	0.35*	**0.37***	NA	NA	1 - 732	9
PRCH	0.03	**0.23***	<0.01	**0.09**	0.19	**0.51****	<0.01	<0.01	1 - 2,664	56
SLIM	**0.30***	0.10	**0.02**	0.01	0.22*	**0.31***	<0.01	**0.35***	1 - 11,595	224
SMLT	**0.07**	<0.01	<0.01	**0.04**	0.10	**0.16**	0.36*	**0.47****	1 - 181,082	1,391
SPOT	0.09	**0.23***	<0.01	**0.04**	0.16	**0.44****	0.26*	**0.44****	1 - 12,055	246
STK3	0.12	**0.29***	<0.01	**0.10**	0.05	**0.29***	<0.01	**0.22***	1 - 16,701	138
TRPR	**0.21***	0.15	**0.01**	<0.01	0.16	**0.25***	0.46**	**0.55****	1 - 23,917	358

The standardized residuals versus the fitted values for the Spottail Shiner GAM using a Poisson distribution showed that there was a non-random pattern indicating that there was not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the assumptions of homogeneity of the variance (Figure 18). When the Spottail Shiner standardized residuals were mapped the highest (≥ 1.5) and lowest (≤ -1.5) deviations from the mean is primarily distributed in the eastern portion of the lake with some occurring in the central portion of the lake (Figure 19). The standardized residuals versus the fitted values and QQ - Plot for the Round Goby showed similar results as the Spottail Shiner of not having homogeneity of the variance (Figure 20). When the Round Goby standardized residuals were mapped, the lowest (≤ -1.5) deviations below the mean were primarily distributed in the central portion of the lake and the highest (≥ 1.5) deviations above the mean were more dispersed between the central and eastern portion of the lake (Figure 21).

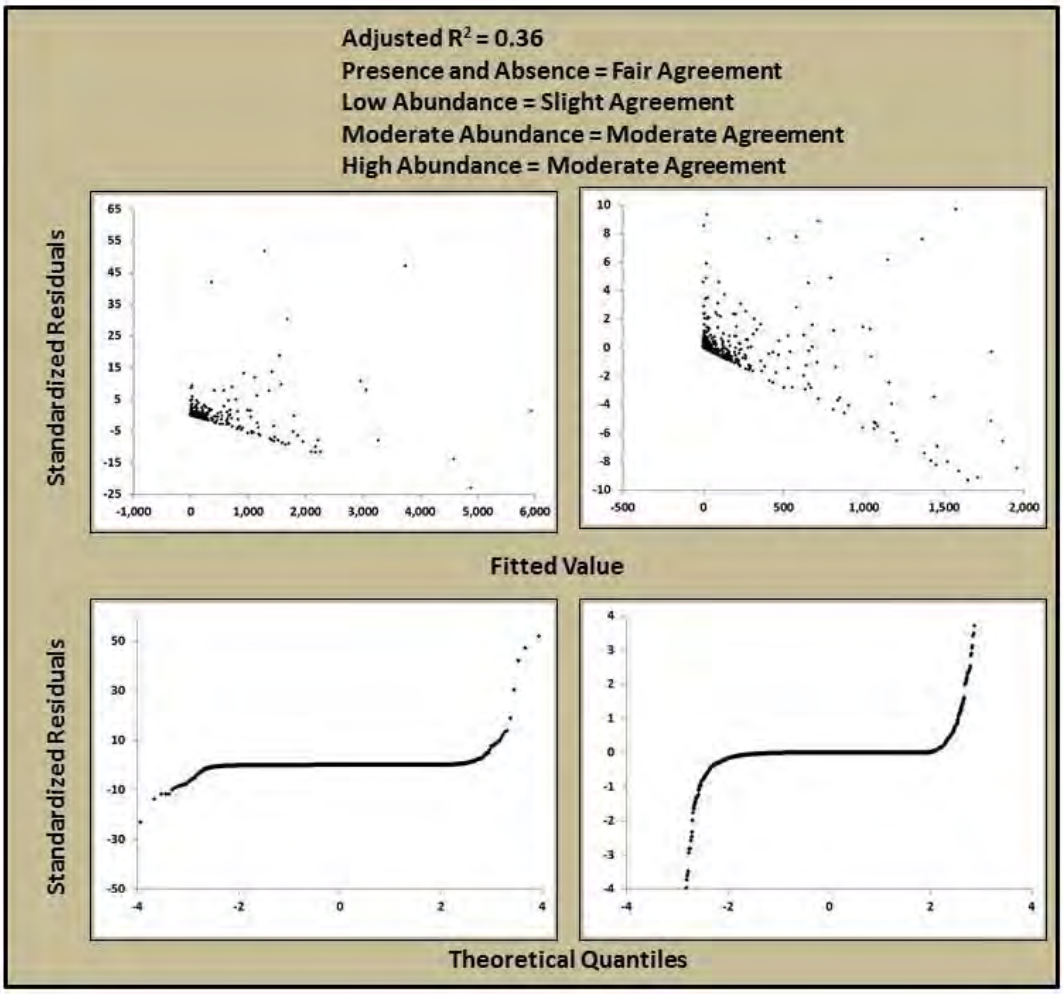

Figure 18 Standardized Residual versus Fitted Value and QQ Plot for Spottail Shiner GAM (1978 - 2014), with Poisson Distribution

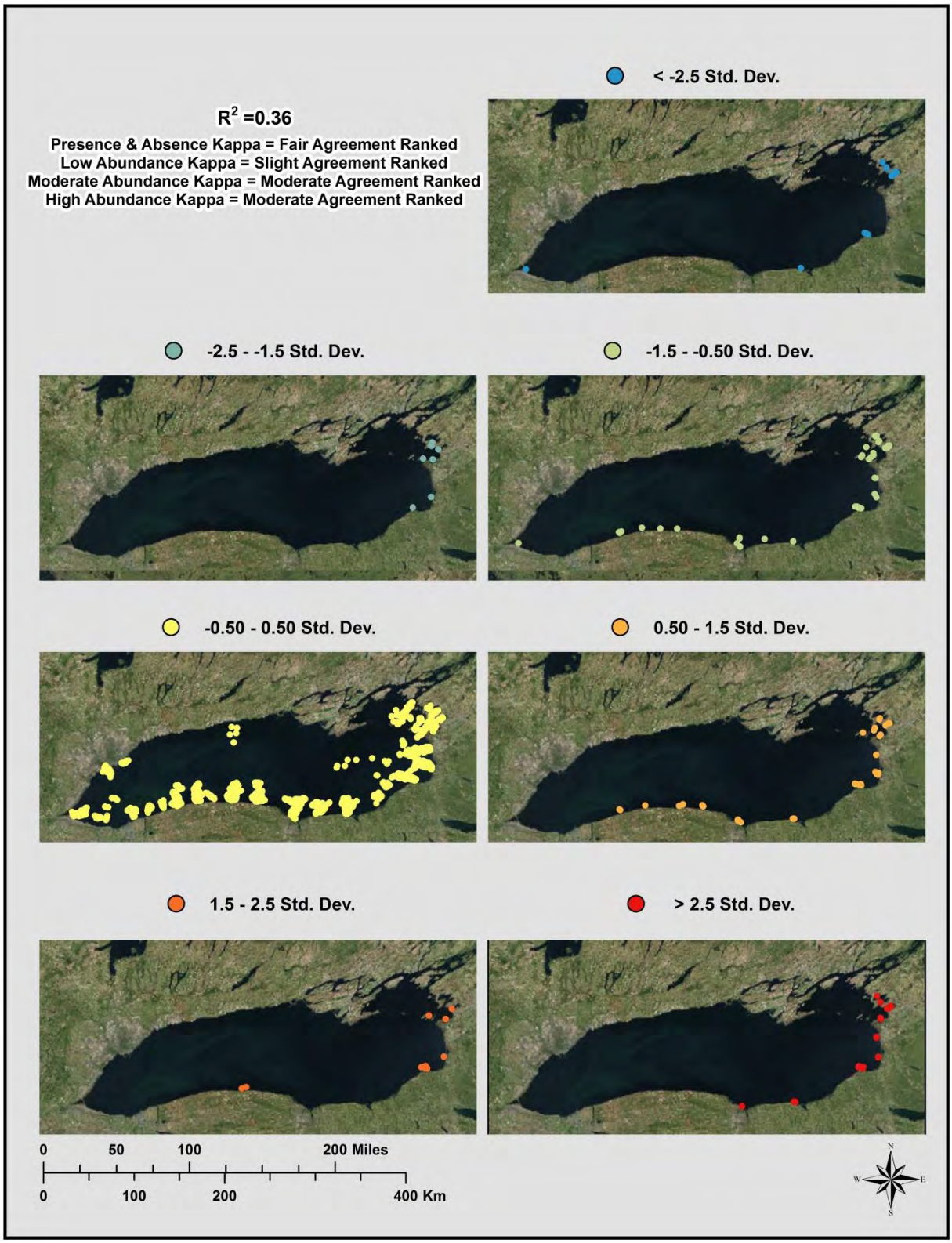

Figure 19 Spatial Distribution of Standardized Residuals for Spottail Shiner GAM (1978 - 2014), with Poisson Distribution

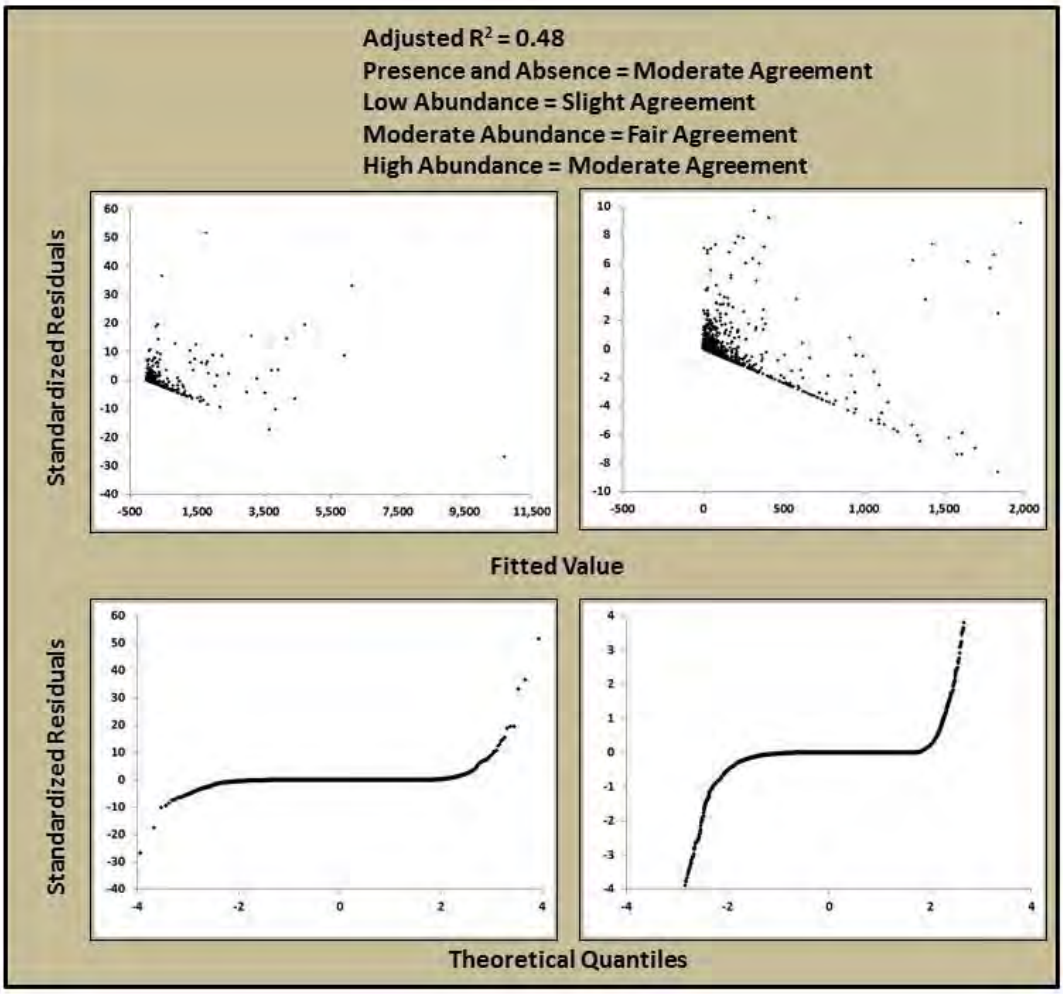

Figure 20 Standardized Residual versus Fitted Value and QQ Plot for Round Goby GAM (1978 - 2014), with Poisson Distribution

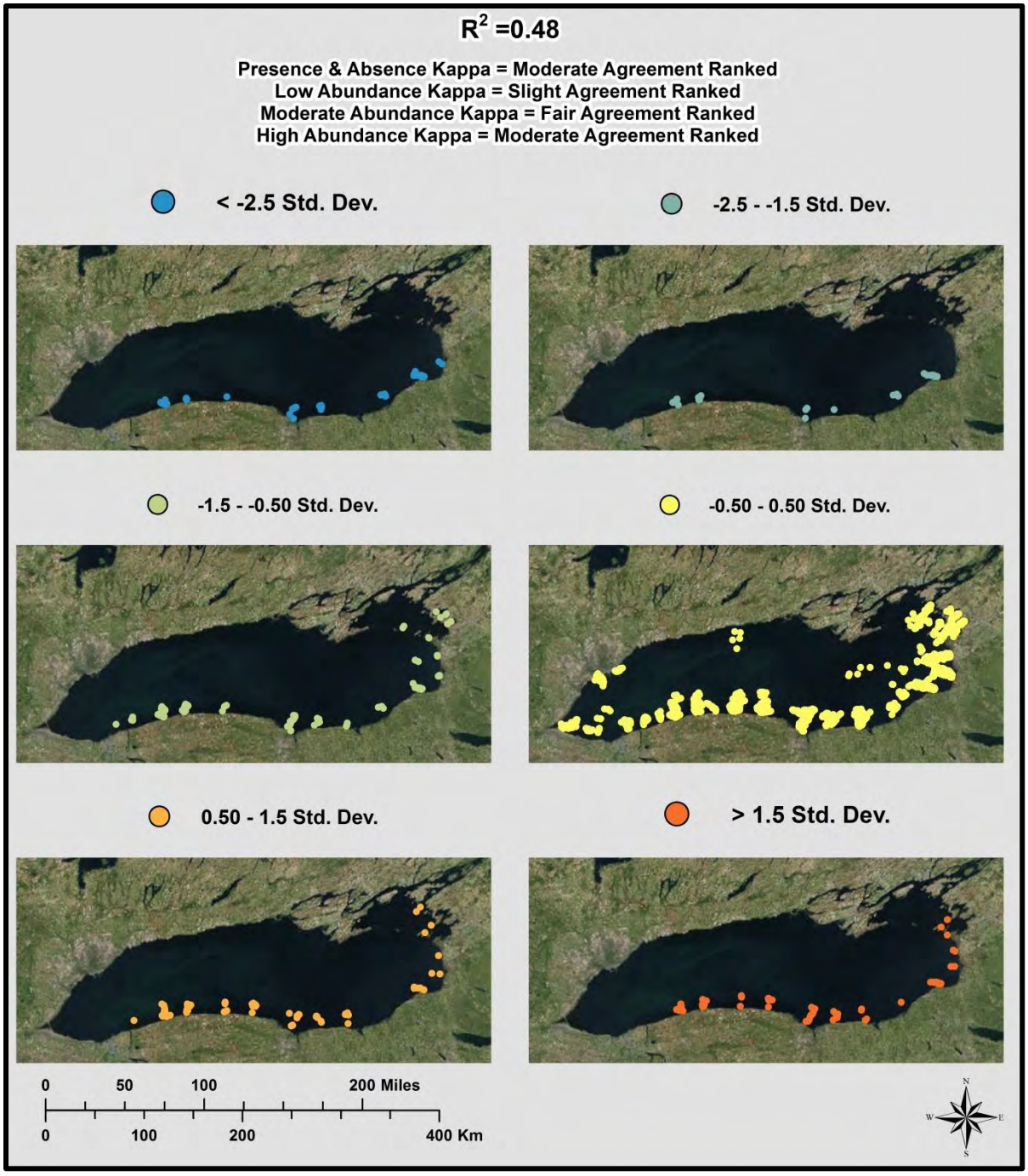

Figure 21 Spatial Distribution of Standardized Residuals for Round Goby GAM (1978 - 2014), with Poisson Distribution

5.2.2. 1978 - 1989 dataset

The 1978 - 1989 dataset saw poor adjusted R^2 values, <0.25, for five of the ten species using either of the distribution types (Table 16). Slimy Sculpin, Spottail Shiner, and Trout Perch

were able to achieve an adjusted R^2value greater than 0.25 using either distribution. Spottail Shiner and Slimy Sculpin saw an increase in adjusted R^2 with the Poisson distribution, but Trout Perch saw a decrease. Spottail Shiner saw the largest difference between the two distributions as well as obtaining the highest adjusted R^2 value. The predictor variables used for the Spottail Shiner model was the Log_{10} transformed distance to open type wetland, month, depth, and the square root of Trout Perch and Johnny Darter abundances.

Table 16 Adjusted R^2 values for GAMs for each species (1978-1989). GAMG uses Gaussian distribution and GAMP uses Poisson distribution. **Bolded** values are the higher between Gaussian and Poisson.

Species	GAMG Adj R^2	GAMP Adj R^2	Difference (GAMG-GAMP)
ALEW	0.14	**0.21**	-0.07
JOHN	0.12	**0.25**	-0.13
LTRT	**0.14**	0.13	.01
PRCH	0.15	**0.24**	-0.09
SLIM	0.28	**0.37**	-0.09
SMLT	0.20	**0.24**	-0.04
SPOT	0.41	**0.74**	-0.33
STK3	0.01	**0.07**	-0.06
TRPR	**0.47**	0.40	0.07

The results of the Cohen's Kappa showed that the Gaussian distribution was better at predicting presence and absence but not as well as for moderate and high abundances compared to the Poisson distribution (Table 17). Neither distribution type was able to get a fair or higher agreement ranking between the observed and predicted values for the low abundance class. Spottail Shiner had the highest adjusted R^2 using the Poisson distribution and was able to get a fair agreement ranking for the presence and absence classification. The GAM with Poisson distribution for Spottail Shiner was also able to get a moderate agreement ranking between observed and predicted values for the moderate and high abundance classes.

Table 17 Cohen's Kappa values for GAMs for each species (1978-1989). G = Gaussian, P = Poisson, (*) denotes fair agreement, (**) denotes moderate agreement. **Bolded** values are the higher between Gaussian and Poisson.

Species	Presence		Low Abundance		Moderate Abundance		High Abundance		When Species Present	
	G	P	G	P	G	P	G	P	Range of abundances	Average abundance
ALEW	**0.30***	<0.01	<0.01	**0.01**	0.03	**0.05**	0.31*	**0.45****	1 - 114,693	2,745
JOHN	**0.13**	0.02	**0.01**	<0.01	0.16	**0.28***	0.20	**0.36***	1 - 10,935	64
LTRT	**0.31***	<0.01	**0.08**	<0.01	0.37*	**0.39***	NA	NA	1 - 431	13
PRCH	0.03	**0.32***	<0.01	**0.12**	0.27*	**0.49****	<0.01	<0.01	1 - 2,336	63
SLIM	**0.30***	0.11	**0.01**	<0.01	**0.27***	0.19	0.41**	**0.46****	1 - 8,528	344
SMLT	**0.06**	<0.01	<0.01	**0.18**	0.15	**0.31***	0.39*	**0.47****	1 - 72,261	2,313
SPOT	0.07	**0.21***	<0.01	**0.01**	0.29*	**0.45****	0.44**	**0.48****	1 - 7,002	262
STK3	0.02	**0.10**	<0.01	**0.05**	<0.01	<0.01	NA	NA	1 - 91	6
TRPR	**0.18**	0.09	**0.02**	<0.01	0.14	**0.27***	0.59**	0.59**	1 - 17,612	431

The standardized residuals versus the fitted values for the Spottail Shiner GAM using a Poisson distribution showed that there was a coned shaped pattern indicating that there is not homogeneity of the variance. There is more dispersal of standardized residuals for the higher values, with lower values being more clustered. The QQ - Plot also indicates that the model did not meet the assumption of homogeneity of the variance (Figure 22). When the Spottail Shiner standardized residuals were mapped, the lowest (≤ -1.5) deviations below the mean were primarily distributed in the eastern portion of the lake (Figure 23). The highest (≥ 1.5) deviations above the mean were dispersed in both the eastern and central portions of the lake.

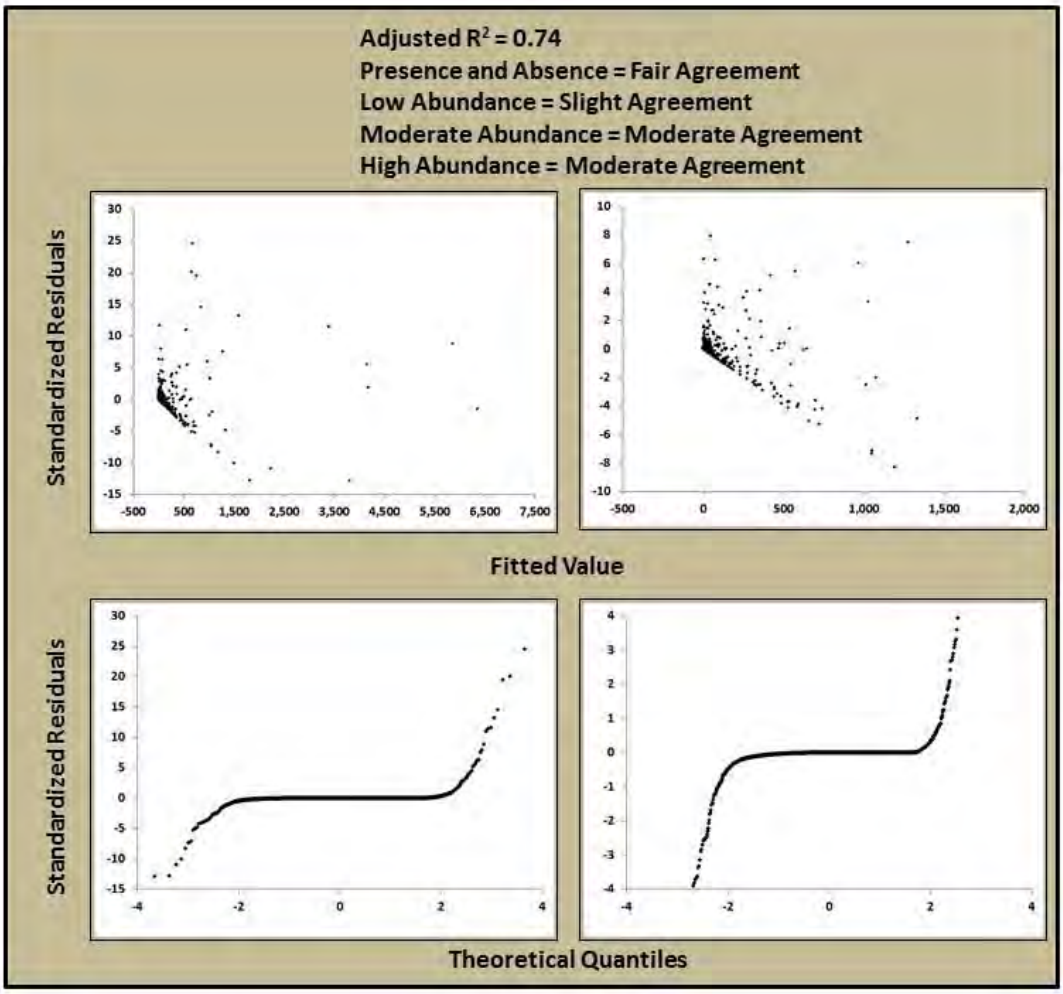

Figure 22 Standardized Residual versus Fitted Value and QQ Plot for Spottail Shiner GAM (1978-1989), with Poisson Distribution

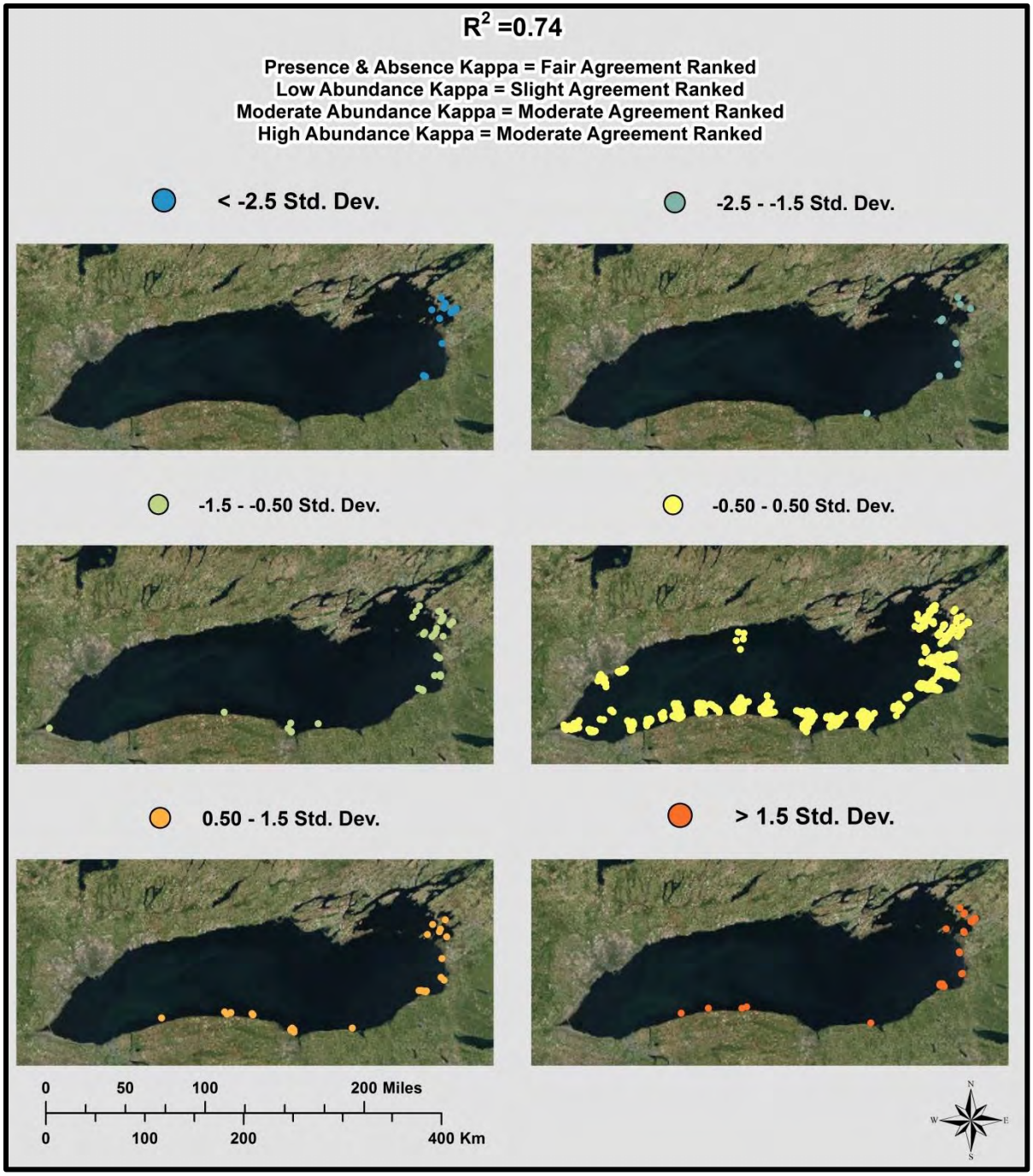

Figure 23 Spatial Distribution of Standardized Residuals for
Spottail Shiner GAM (1978-1989), with Poisson Distribution

5.2.3. 1990 - 2014 dataset

The 1990 - 2014 dataset saw poor adjusted R^2 values, <0.25, for the GAMs with Gaussian distribution (Table 18). The use of the Poisson distribution increased adjusted R^2 saw

values exceed 0.25 for six of the ten species. Spottail Shiner GAM with a Poisson distribution obtained the highest adjusted R^2 value of 0.71. The predictor variables used for the Spottail Shiner model development was Log_{10} transformed distance to delta type wetland, distance to protected type wetland, month, and the square root of Yellow Perch and Trout Perch abundances.

Table 18 Adjusted R^2 values for GAMs for each species (1990-2014). GAMG uses Gaussian distribution and GAMP uses Poisson distribution. **Bolded** values are the higher between Gaussian and Poisson.

Species	GAMG Adj R^2	GAMP Adj R^2	Difference (GAMG-GAMP)
ALEW	0.09	**0.13**	-0.04
GOBY	0.08	**0.48**	-0.40
JOHN	0.06	**0.32**	-0.26
LTRT	0.05	**0.07**	-0.02
PRCH	0.10	**0.28**	-0.18
SLIM	0.17	**0.31**	-0.14
SMLT	0.09	**0.12**	-0.03
SPOT	0.14	**0.71**	-0.57
STK3	0.03	**0.06**	-0.03
TRPR	0.16	**0.35**	-0.19

The Cohen's Kappa showed that the GAM with Poisson distribution was able to get more values in or above the fair agreement rank between observed and predicted values for the presence and absence category as well as the moderate and high abundance classes than the Gaussian distribution (Table 19). Neither distribution type was able to get a fair or higher agreement ranking for the low abundance class. Spottail Shiner which had the highest adjusted R^2 using the Poisson distribution was able to get a fair agreement ranking for the presence and absence classification as well as a moderate agreement ranking for the moderate and high abundance classes, but could only get a slight agreement rank for the low abundance classes.

Table 19 Cohen's Kappa values for GAMs for each species (1990-2014). G = Gaussian, P = Poisson, (*) denotes fair agreement, (**) denotes moderate agreement. **Bolded** values are the higher between Gaussian and Poisson.

Species	Presence		Low Abundance		Moderate Abundance		High Abundance		When Species Present	
	G	P	G	P	G	P	G	P	Range of abundances	Average Abundance
ALEW	**0.35***	<0.01	<0.01	**0.01**	**0.05**	<0.01	0.22*	**0.31***	1 - 124,648	2,265
GOBY	0.21*	**0.36***	0.04	**0.08**	0.13	**0.35***	<0.01	**0.52****	1 - 13,076	215
JOHN	0.17	**0.38***	0.02	**0.11**	0.17	**0.48****	<0.01	**0.38***	1 - 9,103	77
LTRT	**0.14**	0.04	**0.10**	0.02	0.08	**0.27***	NA	NA	1 - 732	6
PRCH	0.02	**0.22***	<0.01	**0.09**	0.12	**0.47****	<0.01	**0.29***	1 - 2,664	52
SLIM	**0.39***	0.16	0.02	0.02	0.27*	**0.32***	<0.01	**0.22***	1 - 11,595	155
SMLT	**0.06**	<0.01	<0.01	**0.03**	0.07	**0.10**	0.36*	**0.44****	1 - 181,082	895
SPOT	0.05	**0.28***	<0.01	**0.09**	0.17	**0.47****	0.31*	**0.50****	1 - 12,055	227
STK3	0.06	**0.09**	0.01	**0.02**	0.07	**0.17**	<0.01	**0.13**	1 - 16,701	144
TRPR	0.17	**0.19**	<0.01	**0.01**	0.12	**0.24***	0.40*	**0.48****	1 - 23,917	287

The standardized residuals versus the fitted values for the Spottail Shiner GAM using a Poisson distribution showed that there was an almost cone like pattern indicating that there is not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the assumption of homogeneity of the variance (Figure 24). When the Spottail Shiner standardized residuals were mapped the highest (≥ 1.5) and lowest (≤ -1.5) deviations from the mean were heavily distributed in the eastern portion of the lake, with few isolated locations in the central portion of the lake (Figure 25).

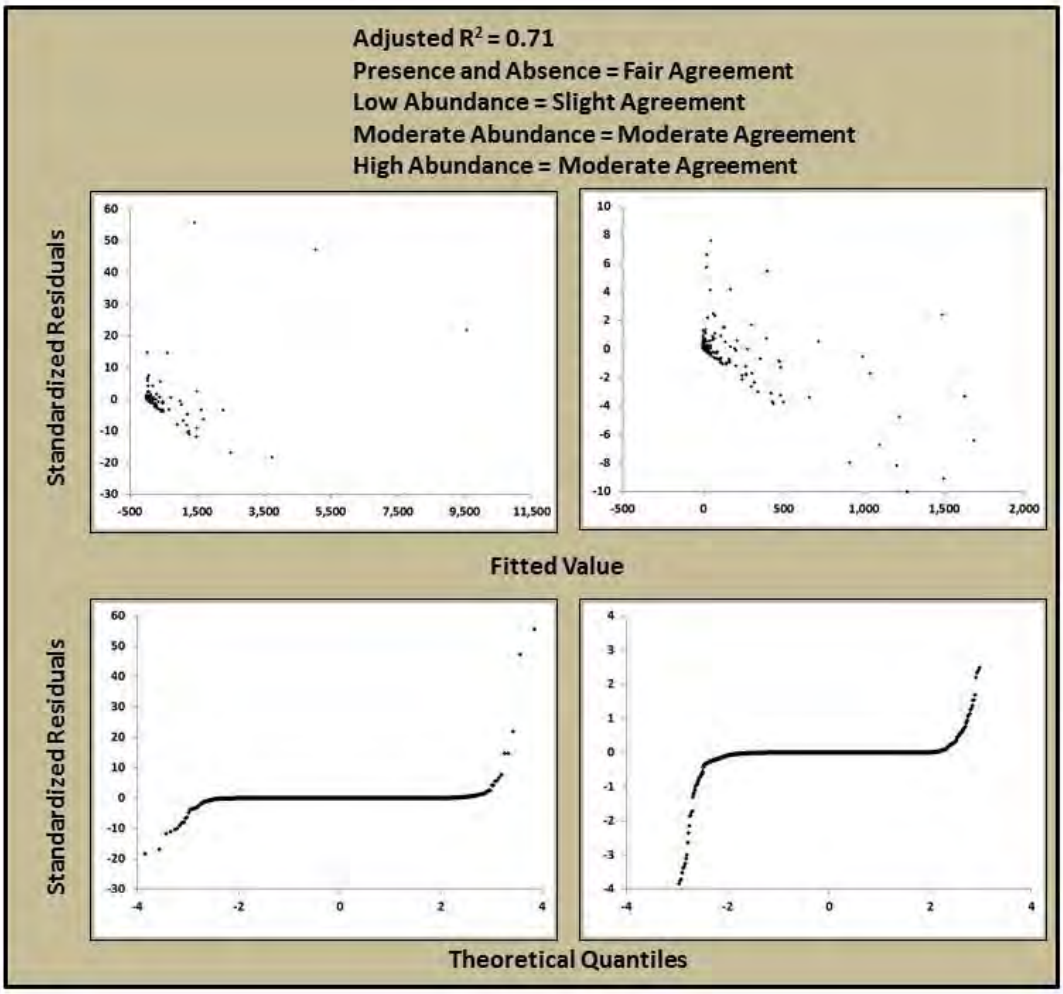

Figure 24 Standardized Residual versus Fitted Value and QQ Plot for Spottail Shiner GAM (1990-2014), with Poisson Distribution

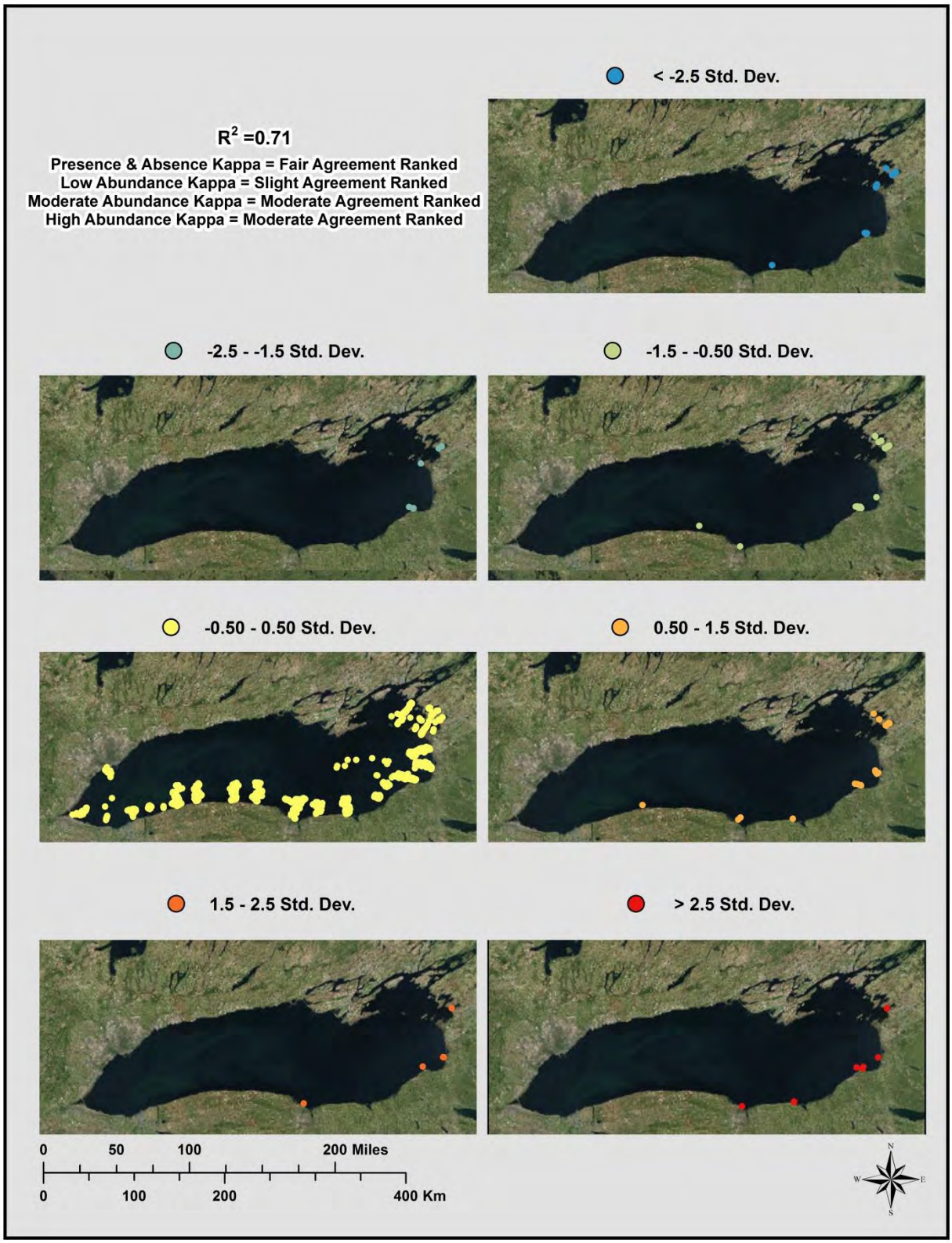

Figure 25 Spatial Distribution of Standardized Residuals for Spottail Shiner GAM (1990-2014), with Poisson Distribution

5.2.4. 2004 - 2014 dataset

The 2004 - 2014 dataset was used to model only for Round Goby and the GAM with Gaussian distribution saw a poor adjusted R^2 values (Table 20). The GAM with Poisson distribution was able to produce a much higher adjusted R^2 values than the Gaussian distribution that explained 49% of the response. The predictor variables used for the Round Goby model development was depth, temperature at fishing depth, distance to protected type wetland, month, and the square root of Alewife abundance. The results of the Cohen's Kappa showed that the Gaussian distribution was better at predicting presence and absence than any other abundance category (Table 21). The Gaussian distribution could only get a slight agreement ranking for the moderate and high abundance classes and no agreement for the low abundance class. The Poisson distribution could only get a slight agreement ranking for presence and absence and a no agreement ranking between the observed and predicted values for low abundances. The Poisson distribution was able to get a fair agreement ranking for the moderate abundance class and a moderate agreement ranking for high abundance class.

Table 20 Adjusted R^2 values of GAMs for Round Goby (2004-2014). GAMG uses Gaussian distribution and GAMP uses Poisson distribution. **Bolded** values are the higher between Gaussian and Poisson.

Species	GAMG Adj R^2	GAMP Adj R^2	Difference (GLMG-GLMP)
GOBY	0.14	**0.49**	-0.35

Table 21 Cohen's Kappa values of GAMs for Round Goby (2004-2014). G = Gaussian, P = Poisson, (*) denotes fair agreement, (**) denotes moderate agreement. **Bolded** values are the higher between Gaussian and Poisson.

Species	Presence		Low Abundance		Moderate Abundance		High Abundance		When Species Present	
	G	P	G	P	G	P	G	P	Range of abundances	Average abundance
GOBY	**0.24***	0.01	<0.01	<0.01	0.19	**0.22***	0.18	**0.51****	1 - 13,076	215

The standardized residuals versus the fitted values for the Round Goby GLM using a Poisson distribution showed that there was a non-random pattern indicating that there was not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the assumption of homogeneity of the variance (Figure 26). When the Round Goby standardized residuals were mapped, the lowest (≤ -1.5) deviations below the mean were distributed in the central portion of the lake and the highest (≥ 1.5) deviations above the mean were dispersed in both the eastern and central portion of the lake (Figure 27).

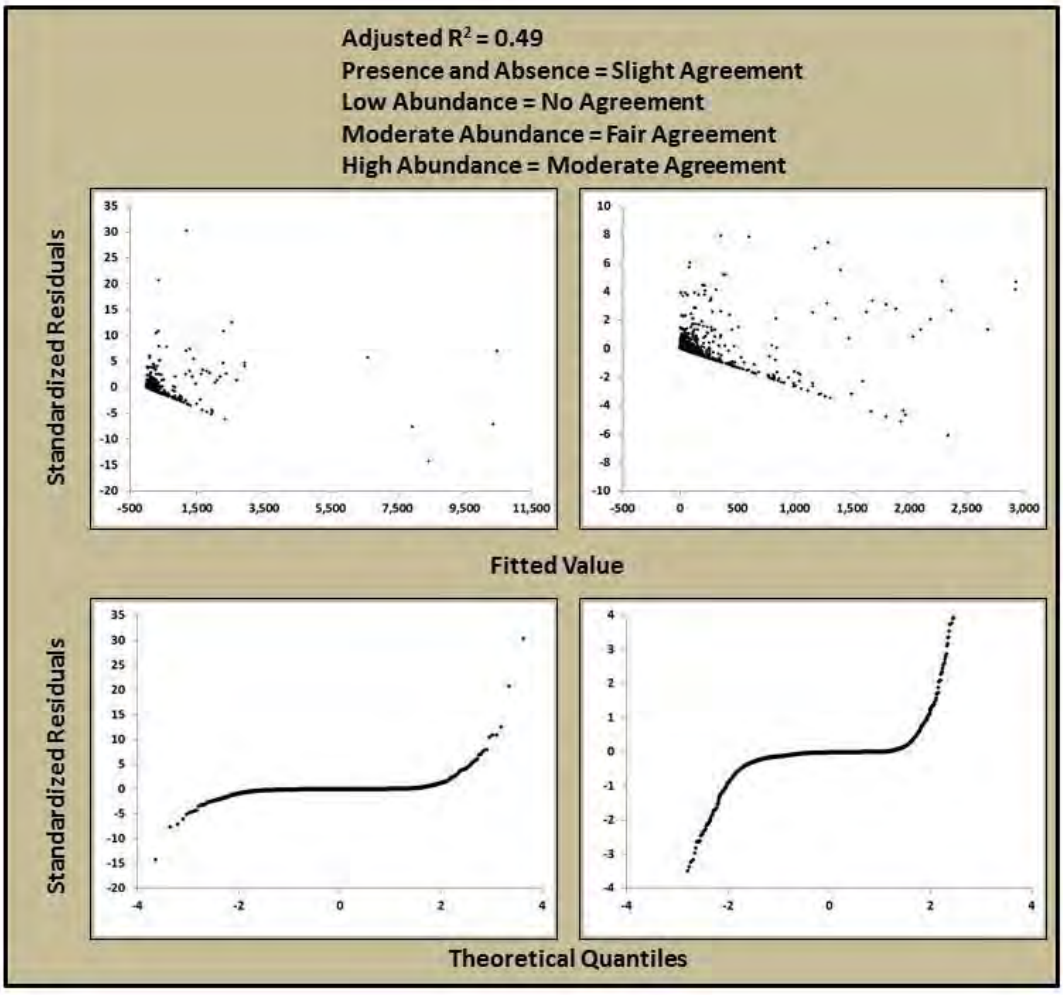

Figure 26 Standardized Residual versus Fitted Value and QQ Plot for Round Goby GAM (2004-2014), with Poisson Distribution

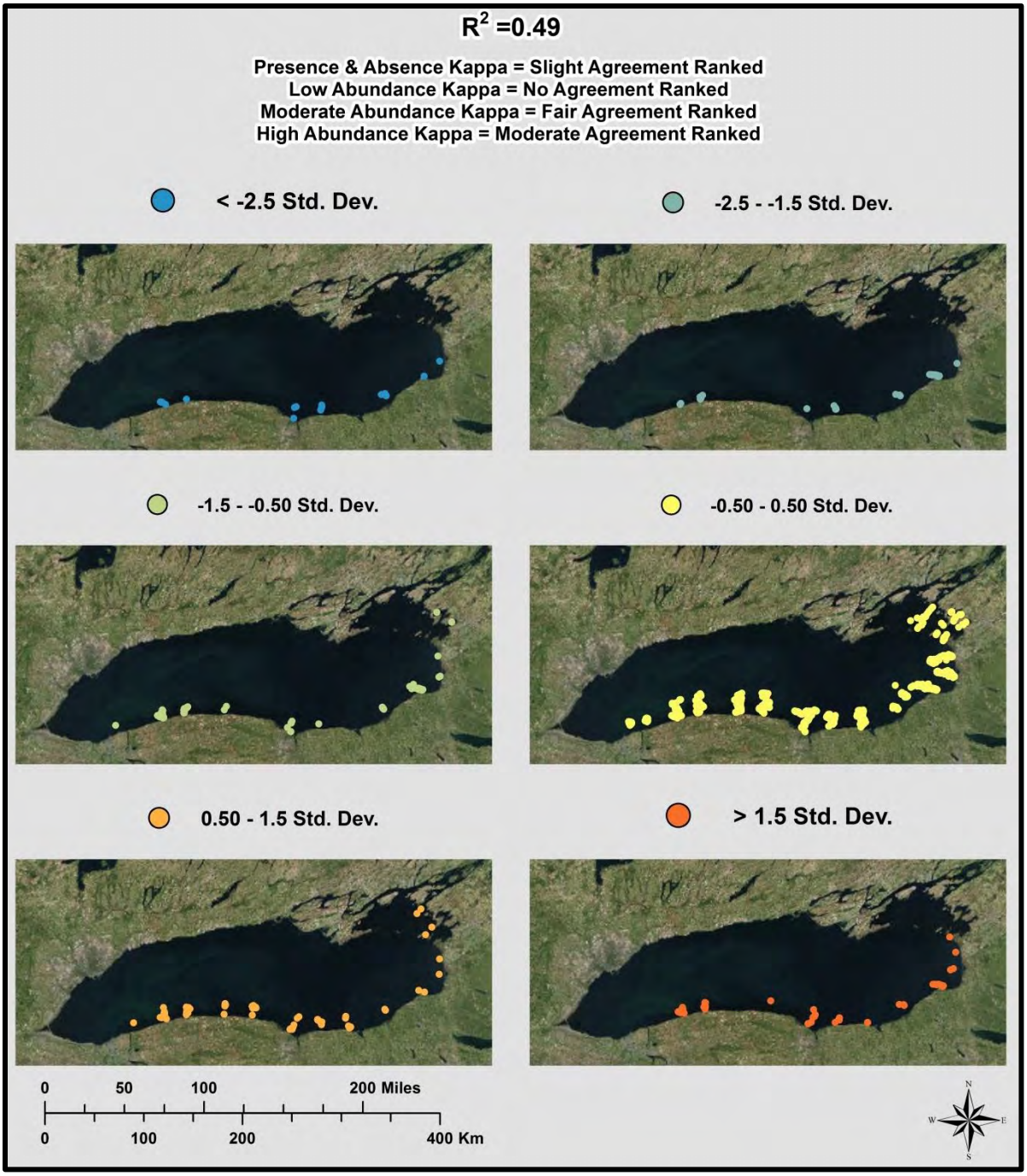

Figure 27 Spatial Distribution of Standardized Residuals for Round Goby GAM (2004-2014), with Poisson Distribution

5.3 Geographically Weighted Regression

The Esri tool used to develop the GWR models was able to produce results for all species and all the datasets, except for Threespine Stickleback. Overall results were poor for each model,

the highest achieving only an adjusted R^2 value of 0.48 for Spottail Shiner in the 1978 - 1989 dataset. Cohen's Kappa values varied from species to species as well as dataset to dataset. The values would range from no agreement to moderate agreement for the presence and absence category, the moderate abundance class, and the high abundance class. A Kappa value for fair agreement between the observed and predicted values for low abundance was achieved for Lake Trout in the 1978 - 2014 and 1990 - 2014 datasets.

5.3.1. 1978 - 2014 dataset

The 1978 - 2014 dataset saw poor adjusted R^2 values, <0.25, for a majority of the species and the local R^2 values range was very broad for all species (Table 22). Round Goby, Lake Trout, Spottail Shiner and Trout Perch were the only species to get an adjusted R^2 value greater than 0.25. The number of neighboring observations needed to produce the models varied species to species. The number of variables also had to be reduced in many cases due to local multicollinearity. Threespine Stickleback could not be modeled at all, with tool warnings of severe model issues. Round Goby had the highest adjusted R^2 value using 151 neighbors and only a single variable of temperature at fishing depth.

Table 22 Adjusted R^2 values for GWRs for each species (1978 - 2014). Local R^2 values, number of neighbors to produce the model, and number of variables used.

Species	Adjusted R^2	Local R^2 Range	Neighbors	# of Variables
ALEW	0.13	0.02 - 0.31	896	4
GOBY	0.33	<0.01 - 0.78	151	1
JOHN	0.13	<0.01 - 0.52	314	2
LTRT	0.31	<0.01 - 0.98	47	1
PRCH	0.15	<0.01 - 0.16	601	1
SLIM	0.24	<0.01 - 0.29	766	3
SMLT	0.15	<0.01 - 0.72	490	2
SPOT	0.27	<0.01 - 0.85	613	2
STK3	--	--	--	--
TRPR	0.29	<0.01 - 0.29	437	3

The results of the Cohen's Kappa showed that Round Goby, which had the highest adjusted R^2 value, could only obtain a fair agreement ranking between the observed and predicted values for the high abundance class. All other abundance categories for Round Goby received a no or slight agreement ranking (Table 23). Lake Trout which had the second highest adjusted R^2 value was able to obtain fair agreement rankings for the presence and absence category as well as the low abundance class. The moderate abundance class was able to get a moderate agreement ranking. Lake Trout did not have values that fell within the range of high abundance, so this category was not available for this species.

Table 23 Cohen's Kappa values for GWRs for each species (1978-2014).
(*) denotes fair agreement, (**) denotes moderate agreement.

Species	Presence	Low Abundance	Moderate Abundance	High Abundance	When Species Present	
					Range of abundances	Average abundance
ALEW	0.24*	<0.01	0.06	0.31*	1 - 124,648	2,438
GOBY	0.07	<0.01	0.12	0.27*	1 - 13,076	215
JOHN	0.17	0.01	0.31*	0.20	1 - 10,935	71
LTRT	0.36*	0.25*	0.53**	NA	1 - 732	9
PRCH	0.39*	0.13	0.45**	<0.01	1 - 2,664	56
SLIM	0.20	<0.01	0.20	0.32*	1 - 11,595	224
SMLT	0.05	0.03	0.23*	0.40*	1 - 181,082	1,391
SPOT	0.29*	0.05	0.43**	0.50**	1 - 12,055	246
STK3	--	--	--	--	1 - 16,701	138
TRPR	0.28*	0.04	0.35*	0.54**	1 - 23,917	358

The standardized residuals versus the fitted values for the Lake Trout GWR showed that there was a non-random pattern indicating that there was not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the homogeneity of the variance assumption (Figure 28). When the Lake Trout standardized residuals were mapped the highest (≥ 1.5) and lowest (≤ -1.5) deviations from the mean were dispersed throughout the eastern and central portions of the lake (Figure 29).

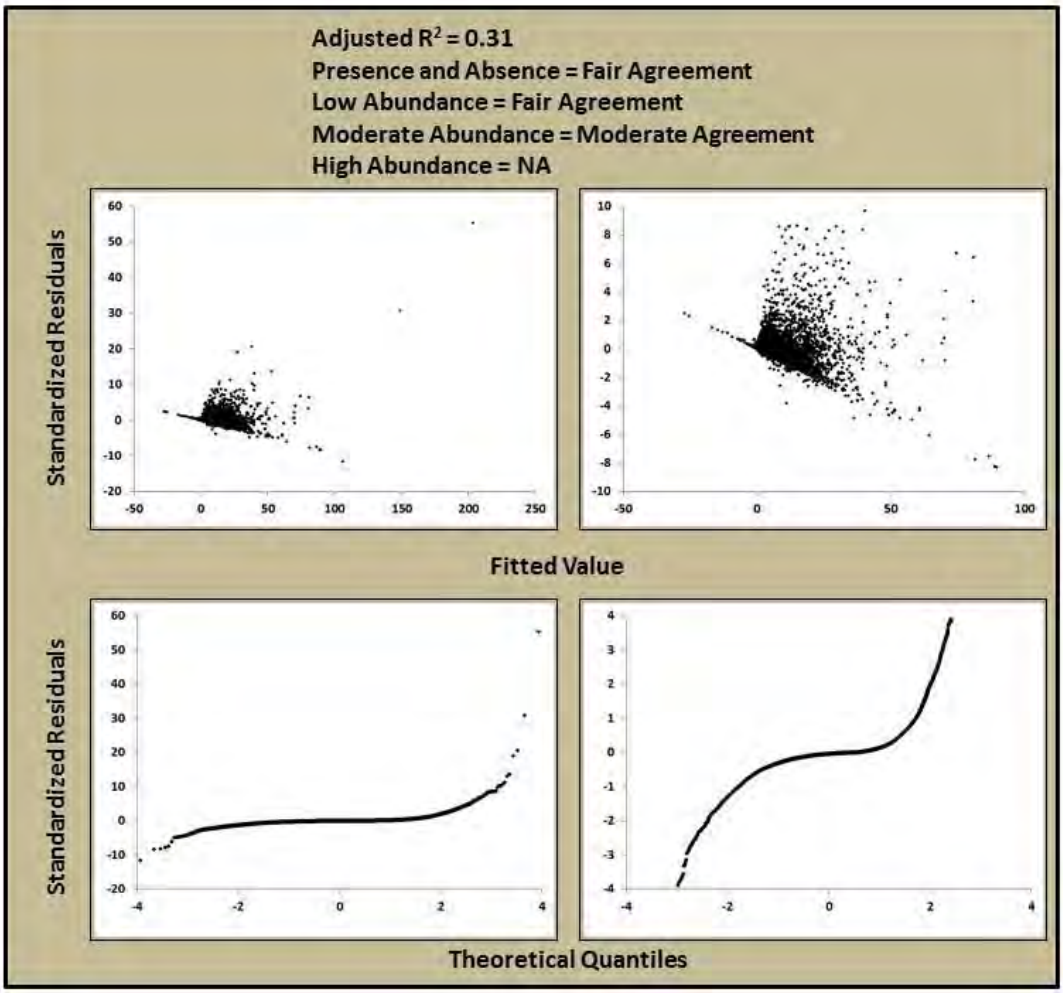

Figure 28 Standardized Residual versus Fitted Value and QQ Plot for Lake Trout GWR (1978-2014)

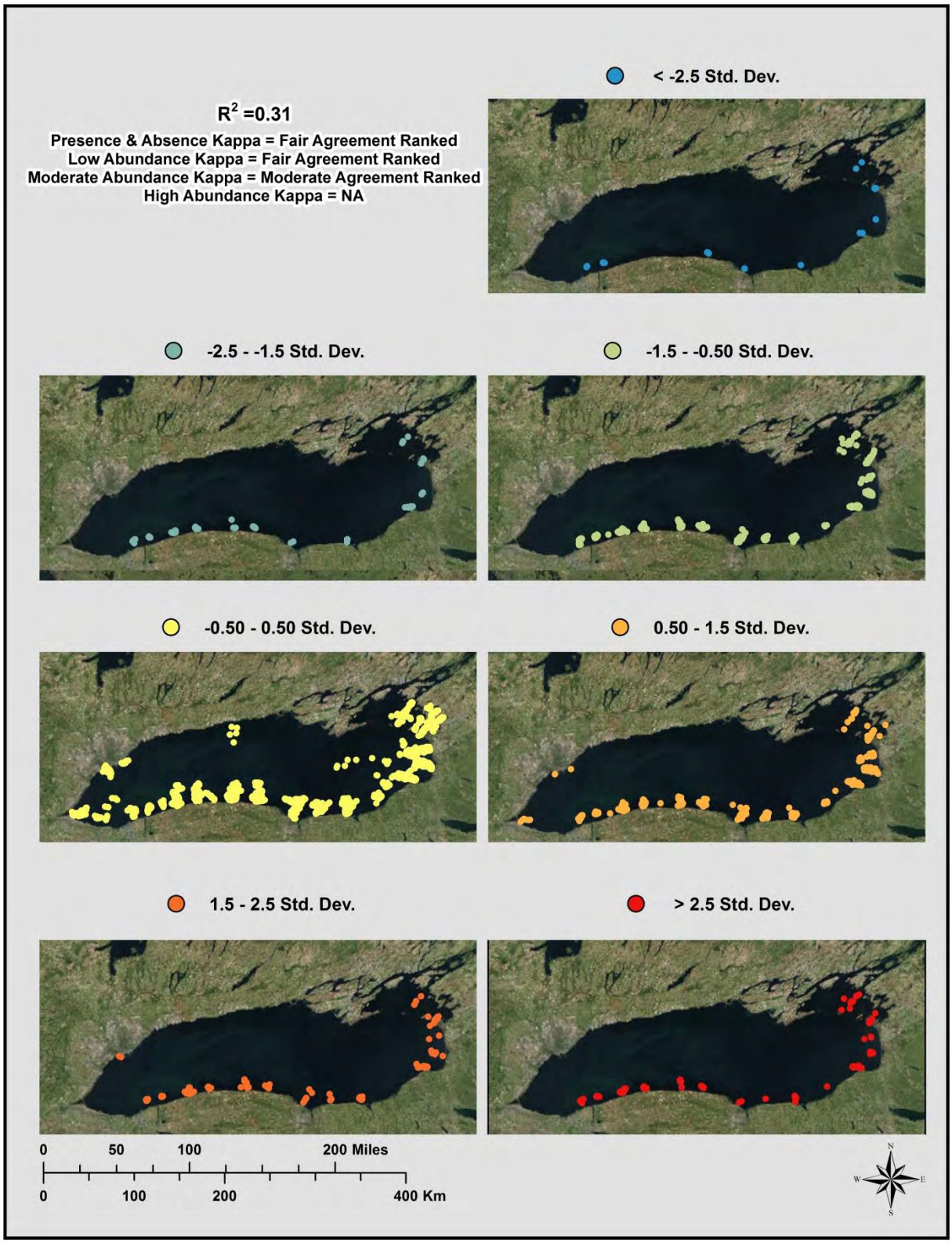

Figure 29 Spatial Distribution of Standardized Residuals for Lake Trout GWR (1978-2014)

The local R^2 values for Lake Trout ranged from less than 0.01 to 0.98 and were dispersed throughout the lake (Figure 30). The coefficients for the Lake Trout model are shown in Figure 31. The coefficient for the square root of abundance for Rainbow Smelt had a small range of values from -1.2 to 1.8. The highest values are indicated with the black boxes labeled A in Figure 31. The majority of the areas have lower values with small moderate values scattered along the US side of the lake. The coefficient for the intercept had a large range of values from -43.1 to 101.2. The lowest values were few and located in the black boxes labeled B in Figure 31. The majority of areas have moderate values with the lowest values located in a small area in the black box labeled B in Figure 31 The highest values were located in the western and eastern portion of the lake in the black boxes labeled C in Figure 31.

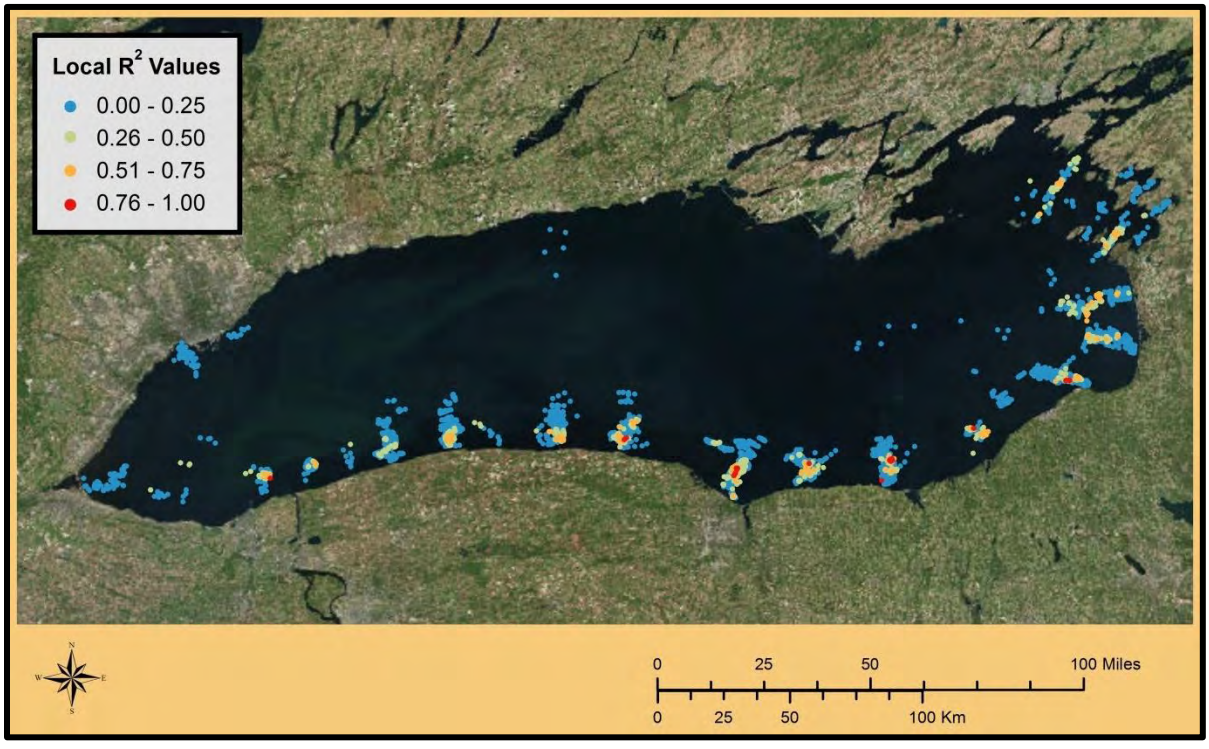

Figure 30 GWR Local R^2 Values for Lake Trout (1978 - 2014)

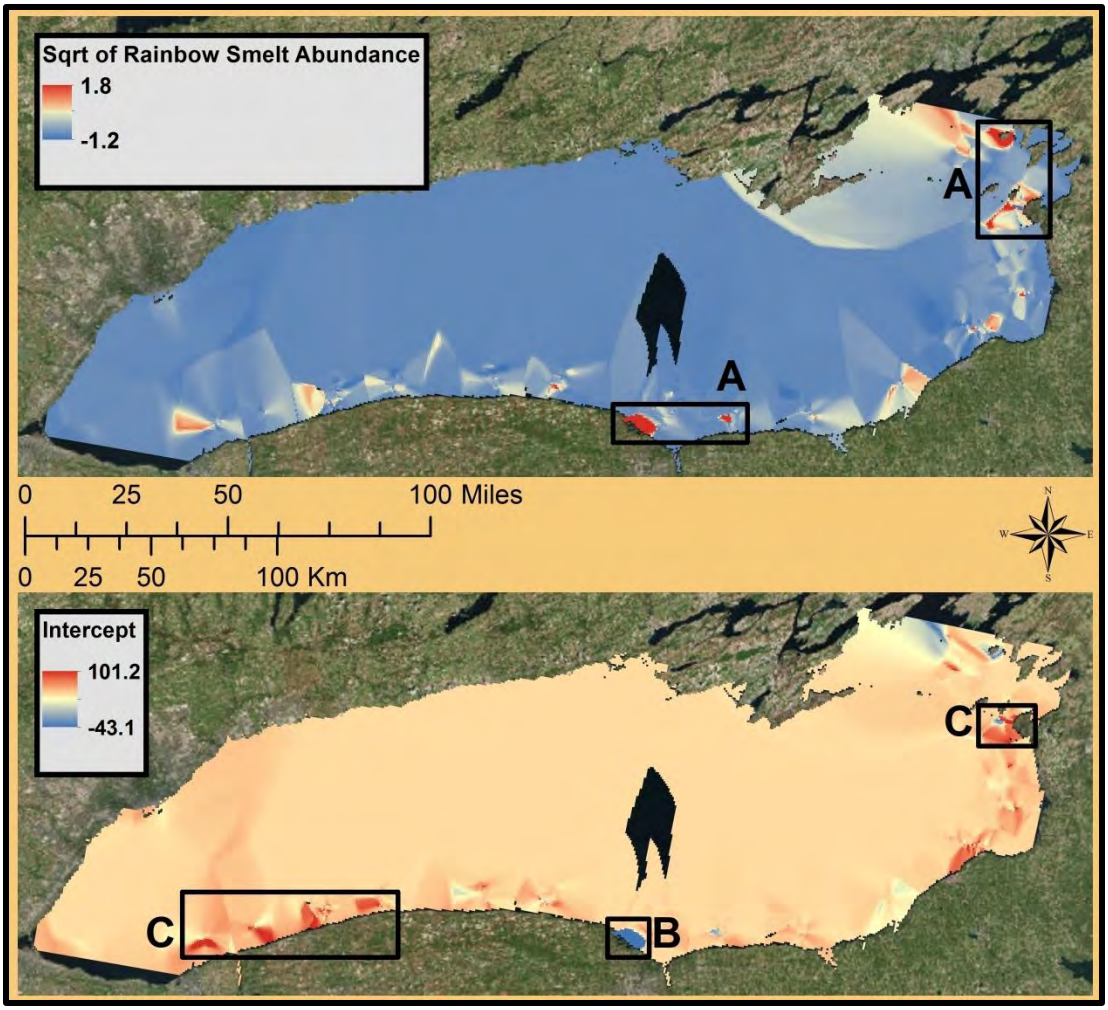

Figure 31 Lake Trout (1978-2014) GWR Coefficient Rasters

5.3.2. 1978 - 1989 dataset

The 1978 - 1989 dataset saw poor adjusted R^2 values, <0.25, for a majority of the species and the local R^2 values range was very broad for all species (Table 24). Slimy Sculpin, Spottail Shiner, and Trout Perch are the only species to get an adjusted R^2 value greater than 0.25. The number of neighboring observations needed to produce the results varied species to species. The number of variables also had to be reduced in many cases due to local multicollinearity. Threespine Stickleback could not be modeled at all due to sever model issues. Trout Perch had

the highest adjusted R^2 value using 450 neighbors and only two variables, the square root of Spottial Shiner and Yellow Perch abundances.

Table 24 Adjusted R^2 values of GWRs for each species (1978 - 1989). Local R^2 values, number of neighbors to produce the model, and number of variables used.

Species	Adjusted R^2	Local R^2 Range	Neighbors	# of Variables
ALEW	0.16	0.08 - 0.57	552	4
JOHN	0.10	<0.01 - 0.71	359	2
LTRT	0.13	0.02 - 0.16	984	4
PRCH	0.20	<0.01 - 0.19	201	1
SLIM	0.46	<0.01 - 0.45	354	2
SMLT	0.18	0.02 - 0.18	943	2
SPOT	0.48	<0.01 - 0.82	199	1
STK3	--	--	--	--
TRPR	0.52	<0.01 - 0.88	450	2

The results of the Cohen's Kappa showed that Trout Perch, which had the highest adjusted R^2 value, only obtained a fair agreement ranking for the moderate abundance class and a moderate agreement ranking for the high abundance class. All other categories received a no or slight agreement ranking for the Trout Perch GWR model (Table 25). Slimy Sculpin and Spottail Shiner had similar adjusted R^2 values and obtained fair agreement rankings between the observed and predicted values for the presence and absence category. Slimy Sculpin and Spottail Shiner both had fair or moderate agreement rankings for the moderate and high abundance classes.

Table 25 Cohen's Kappa values for GWRs for each species (1978-1989).
(*) denotes fair agreement, (**) denotes moderate agreement.

Species	Presence	Low Abundance	Moderate Abundance	High Abundance	When Species Present	
					Range of abundances	Average abundance
ALEW	0.08	0.01	0.02	0.34*	1 - 114,693	2,745
JOHN	0.07	<0.01	0.16	0.29*	1 - 10,935	64
LTRT	0.17	0.08	0.39*	NA	1 - 431	13
PRCH	0.42**	0.12	0.51**	<0.01	1 - 2,336	63
SLIM	0.29*	0.01	0.31*	0.57**	1 - 8,528	344
SMLT	0.13	<0.01	0.15	0.33*	1 - 72,261	2,313
SPOT	0.24*	0.03	0.43**	0.57**	1 - 7,002	262
STK3	--	--	--	--	1 - 91	6
TRPR	0.03	<0.01	0.38*	0.60**	1 - 17,612	431

The standardized residuals versus the fitted values for the Slimy Sculpin GWR show that there is a cone like, pattern indicating that there is not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the required assumptions (Figure 32). When the Slimy Sculpin standardized residuals were mapped, the lowest (≤ -1.5) deviations below the mean were isolated in a small area in the eastern portion of the lake, and the highest (≥ 1.5) deviations above the mean were more heavily distributed in the central portion of the lake with some also in the eastern portion of the lake (Figure 33).

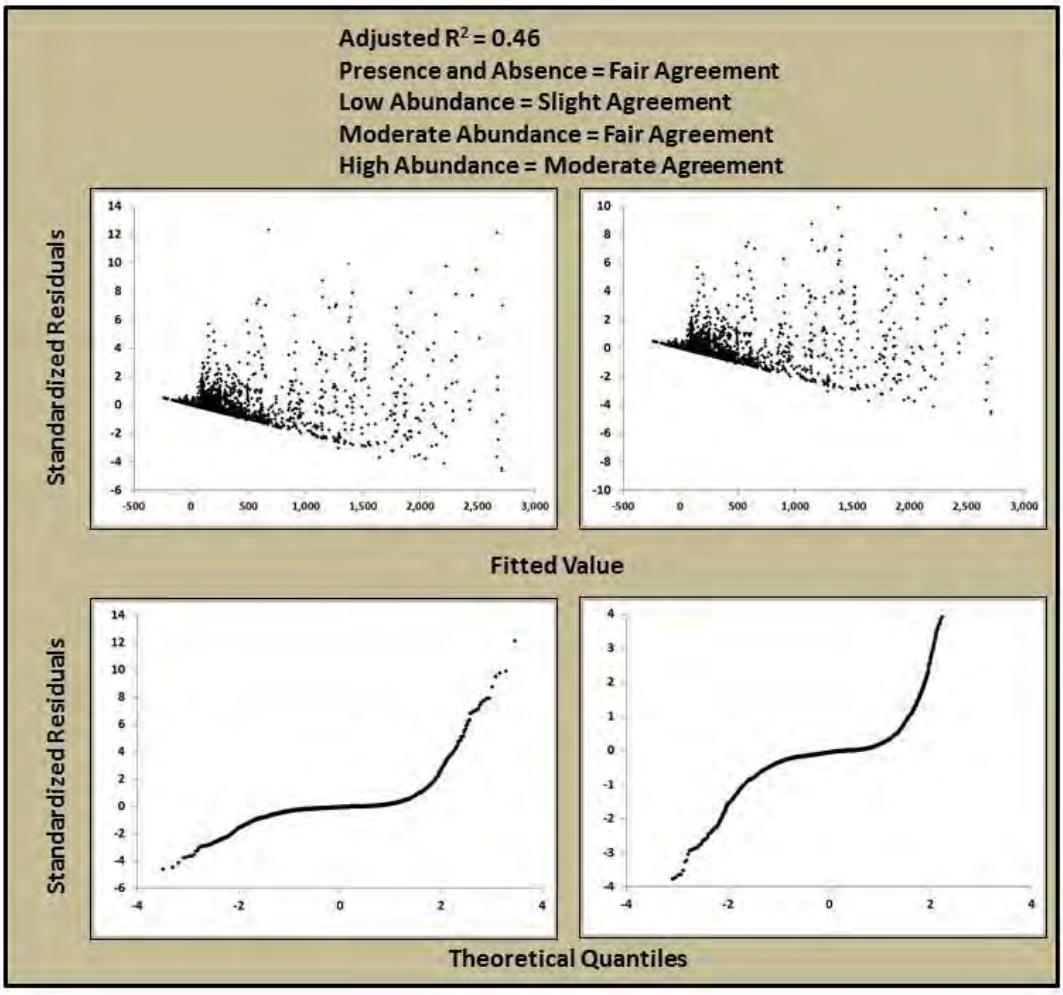

Figure 32 Standardized Residual versus Fitted QQ Plot for Slimy Sculpin GWR (1978-1989)

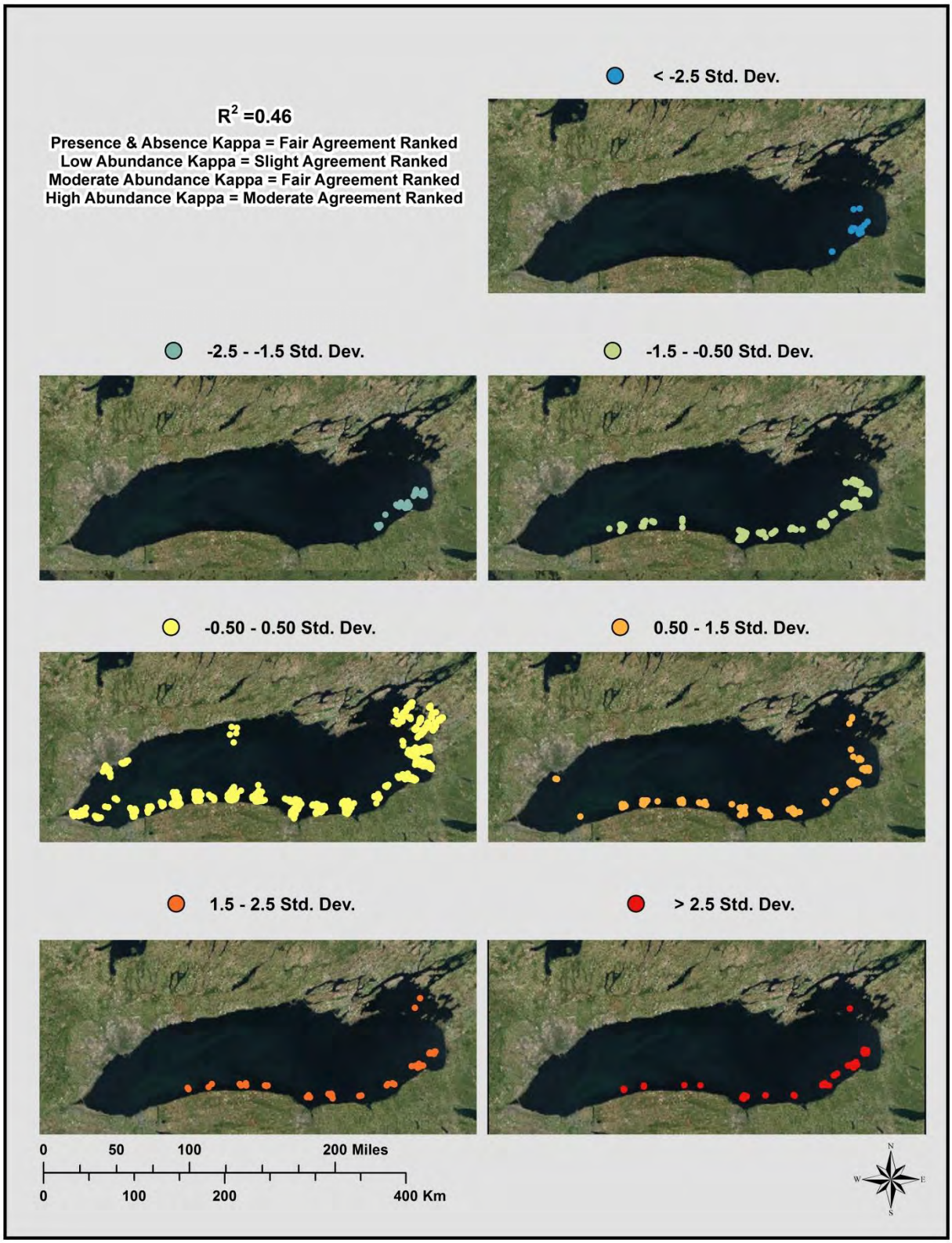

Figure 33 Spatial Distribution of Standardized Residuals for Slimy Sculpin GWR (1978-1989)

The local R^2 values for Slimy Sculpin ranged from less than 0.01 to 0.45. The higher values were isolated in the eastern portion of the lake as well as a small area in the western portion of the lake (Figure 34-A). The rest of the lake has local R^2 values of ≤ 0.25.

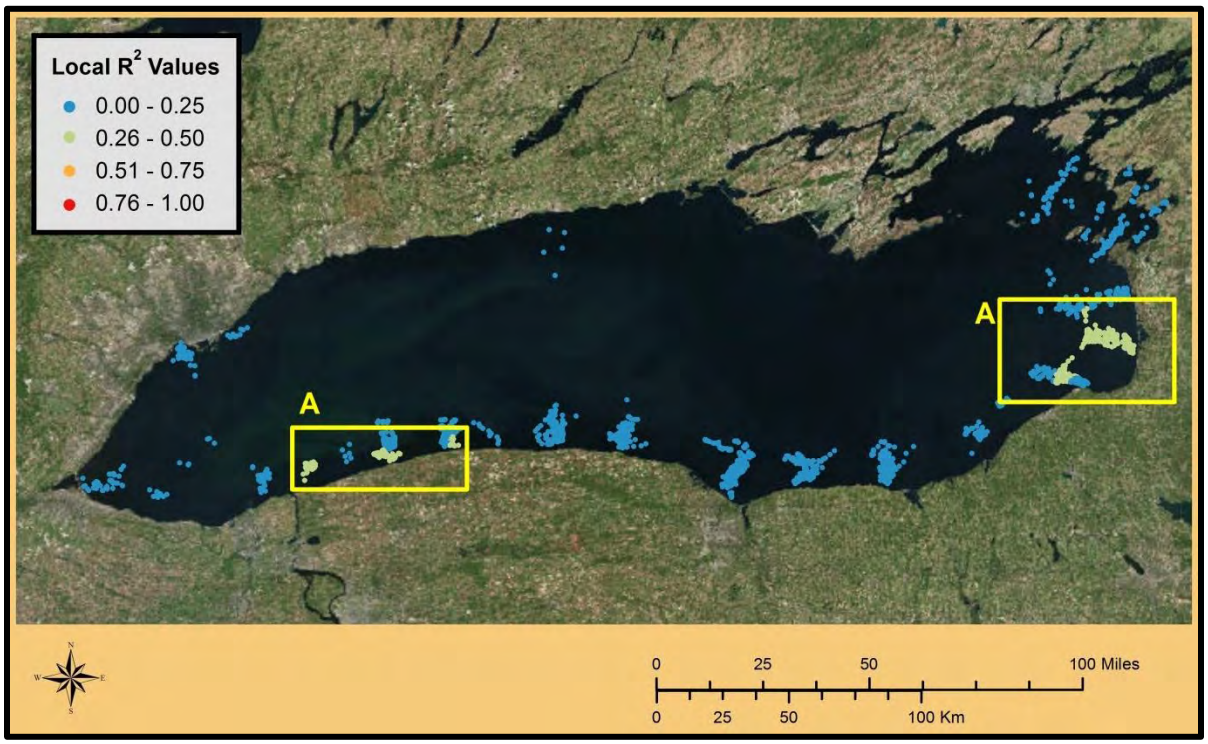

Figure 34 GWR Local R^2 Values for Slimy Sculpin (1978 - 1989)

The coefficients for Slimy Sculpin are shown in Figure 35. The coefficient for the influence of trawling in October had a large range of values from -238 to 1,659. The highest values were isolated to the eastern portion of the lake. The majority of the areas have lower values with few moderate values scattered across the lake (Figure 35-A). The coefficient for depth had smaller values ranging from -27.6 to 64.3. The lowest and highest values for the depth coefficient were also located in the eastern portion of the lake, with moderate values covering the majority of the lake (Figure 35-B). The coefficient for the intercept had a large range of values from -2,653 to 3,707. The lowest and highest values like the other coefficients were heavily

located in the eastern portion of the lake (Figure 35-C). The majority of areas had low to moderate values.

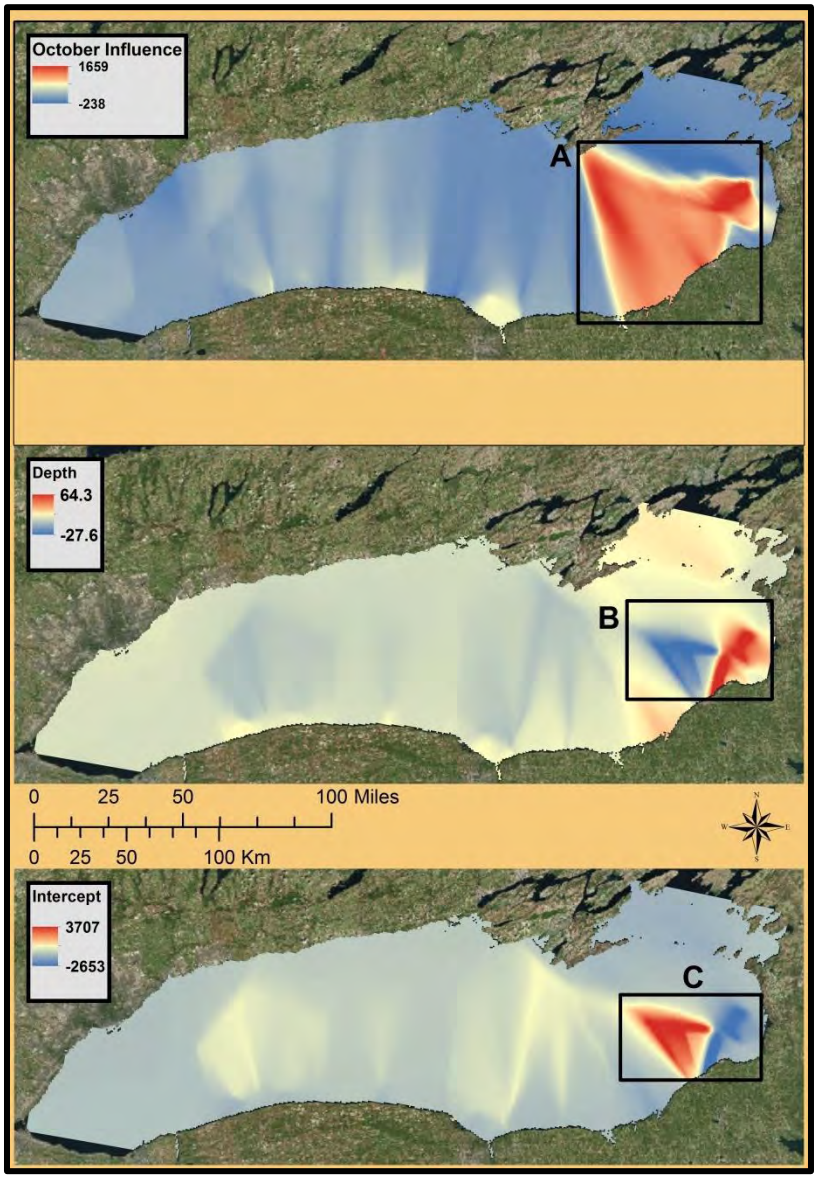

Figure 35 Slimy Sculpin (1978-1989) GWR Coefficient Rasters

The standardized residuals versus the fitted values for the Spottail Shiner GWR showed that there is a cone like pattern indicating that there is not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the homogeneity of the variance0 assumption (Figure 36). When the Spottail Shiner standardized residuals were mapped the highest (≥1.5) and

lowest (≤ -1.5) deviations from the mean were heavily distributed in the eastern portion of the lake, with a few isolated locations in the central portion of the lake (Figure 37).

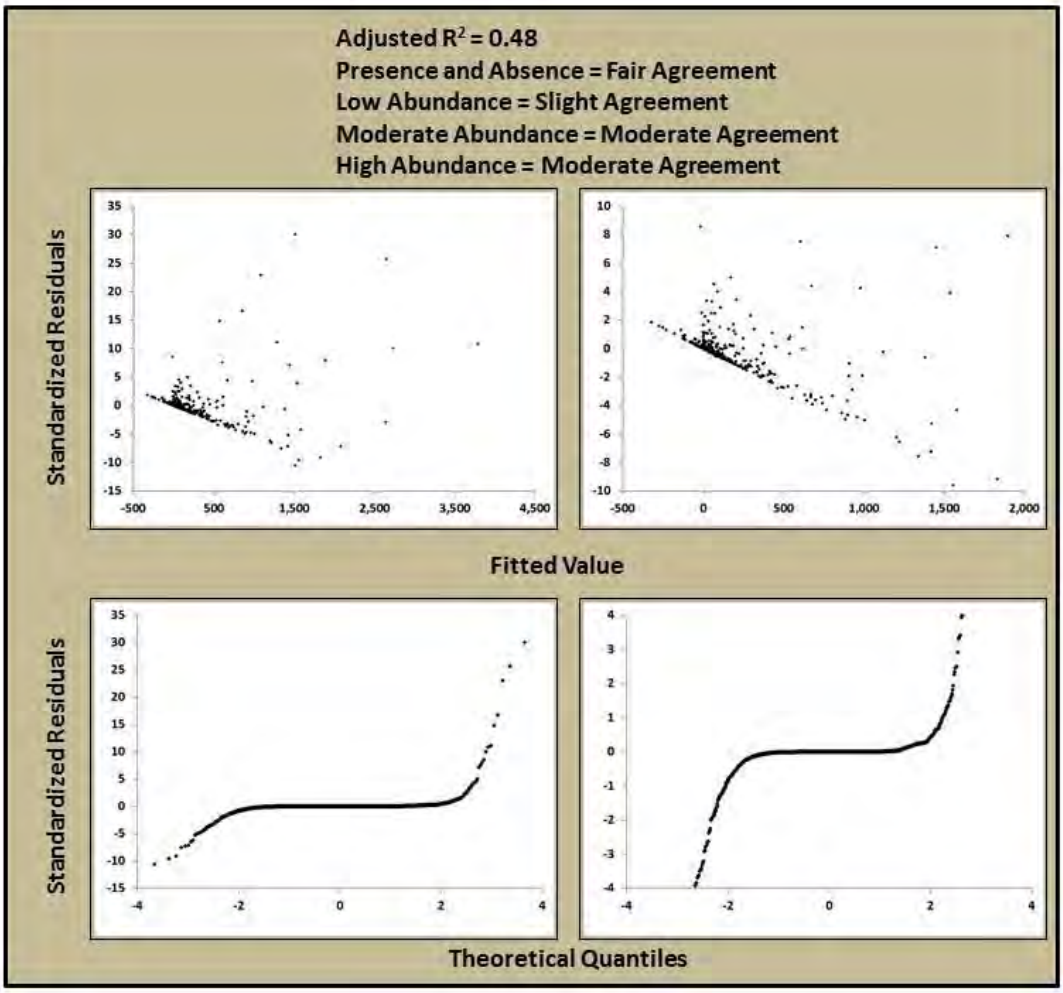

Figure 36 Standardized Residual versus Fitted Value and QQ Plot for Spottail Shiner GWR (1978-1989)

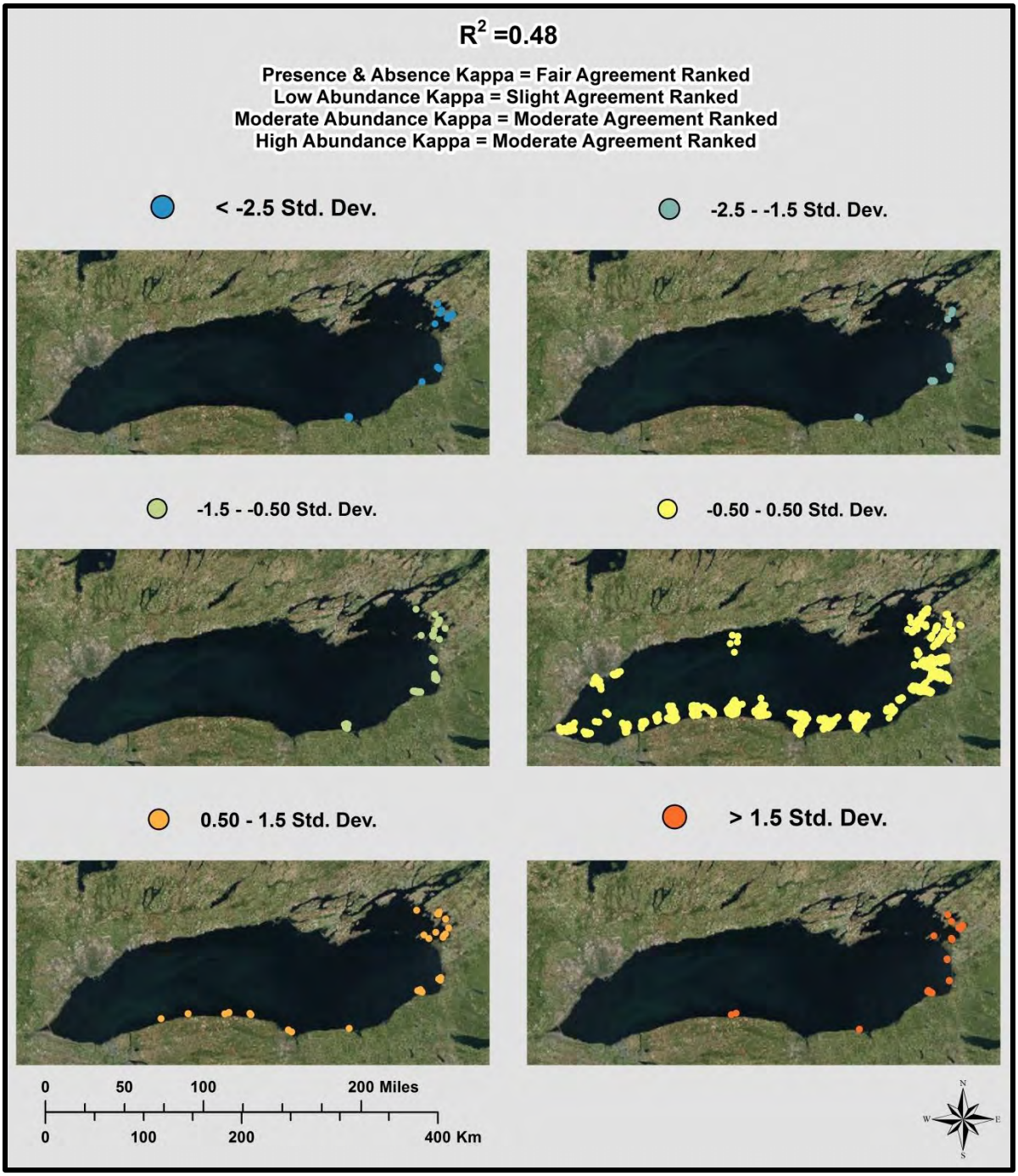

Figure 37 Spatial Distribution of Standardized Residuals for Spottail Shiner GWR (1978-1989)

The local R^2 values for Slimy Sculpin ranged from less than 0.01 to 0.82. The higher values were isolated heavily in three areas enclosed in the yellow boxes labeled A (Figure 38). The rest of the lake had local R^2 values of ≤ 0.25 with a few higher values scattered among them.

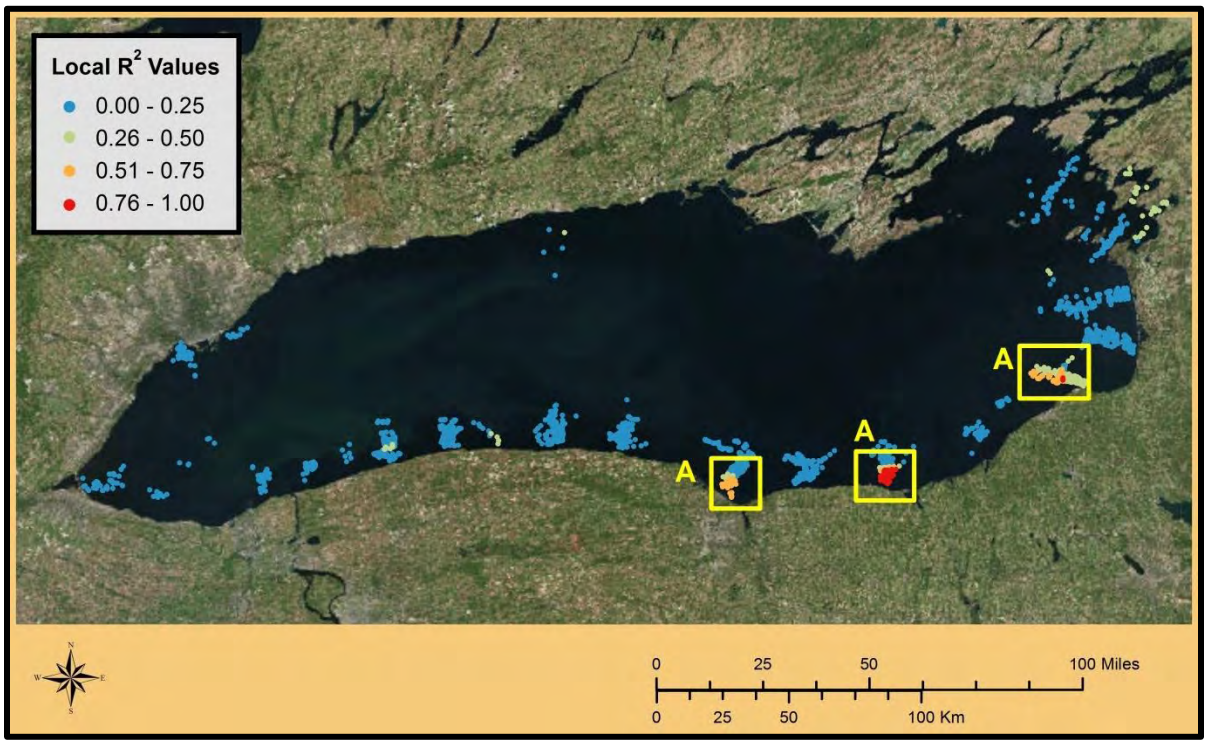

Figure 38 GWR Local R^2 Values for Spottail Shiner (1978 - 1989)

The coefficients for Spottail Shiner are shown in Figure 39. The coefficient for the square root of Trout Perch abundance had a small range of values from -0.5 to 49. The highest values were isolated mainly to the eastern portion of the lake (Figure 39-A). The majority of the areas had lower values across the lake. The coefficient for the intercept had a large range of values from -394.3 to 52.4. The highest values were primarily isolated in the eastern portion of the lake (Figure 39-B). The majority of areas had higher values with a few areas with moderate values located in the eastern portion of the lake.

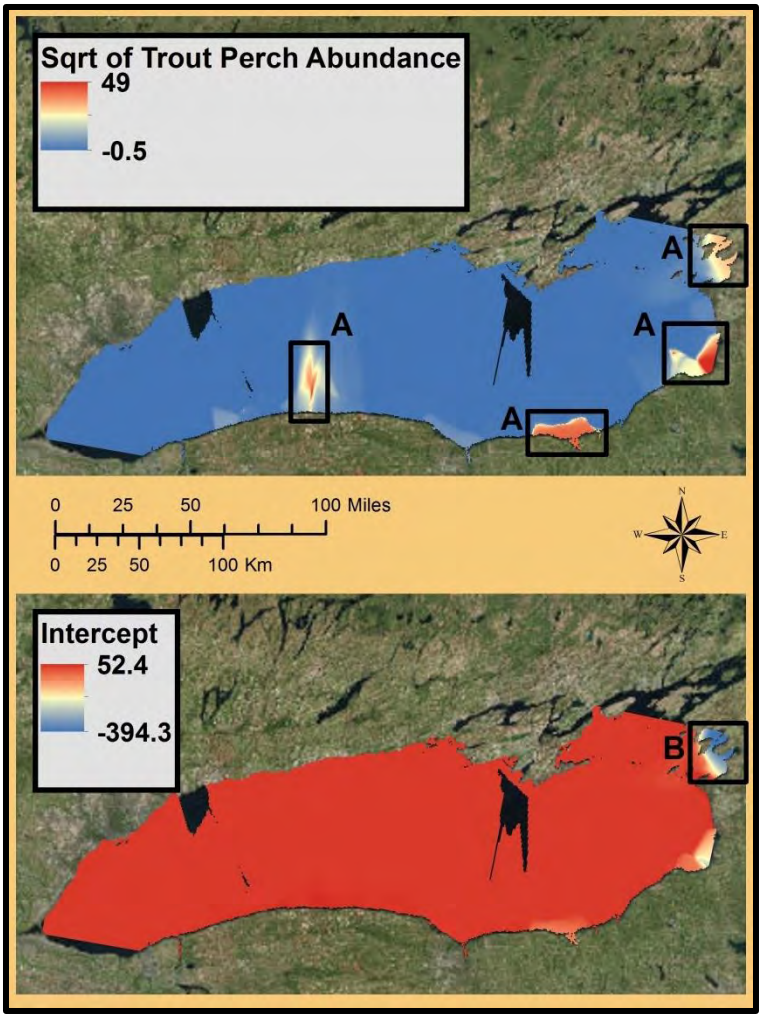

Figure 39 Spottail Shiner (1978-1989) GWR Coefficient Rasters

5.3.3. 1990 - 2014 dataset

The 1990 - 2014 dataset saw poor adjusted R^2 values, <0.25, for all but Round Goby and Lake Trout. The local R^2 value range was very broad for all species (Table 26). The number of neighboring observations needed to develop the model varied species to species. The number of variables also had to be reduced in many cases due to local multicollinearity. Threespine Stickleback could not be modeled at all. Round Goby had the highest adjusted R^2 value using 148 neighbors and only a single variable of temperature at fishing depth.

Table 26 Adjusted R^2 values of GWRs for each species (1990 - 2014). Local R^2 values, number of neighbors to produce the model, and number of variables used.

Species	Adjusted R^2	Local R^2 Range	Neighbors	# of Variables
ALEW	0.11	0.01 - 0.26	960	5
GOBY	0.35	<0.01 - 0.76	148	1
JOHN	0.22	<0.01 - 0.71	211	2
LTRT	0.34	<0.01 - 0.92	41	1
PRCH	0.15	<0.01 - 0.27	220	1
SLIM	0.16	<0.01 - 0.19	766	2
SMLT	0.16	<0.01 - 0.72	503	2
SPOT	0.16	<0.01 - 0.85	549	2
STK3	--	--	--	--
TRPR	0.20	<0.01 - 0.59	990	2

The results of the Cohen's Kappa showed that Round Goby, which had the highest adjusted R^2 value, obtained a fair agreement ranking between observed and predicted values for the high abundance class and a slight agreement for all the other abundance classes (Table 27). Lake Trout, which had a similar adjusted R^2 value of 0.34, obtained a fair or moderate agreement ranking for every abundance class that had representation. Lake Trout had no abundances that few within the high abundance category.

Table 27 Cohen's Kappa values for GWRs for each species (1990-2014). (*) denotes fair agreement, (**) denotes moderate agreement.

Species	Presence	Low Abundance	Moderate Abundance	High Abundance	When Species Present	
					Range of abundances	Average abundance
ALEW	0.32*	0.01	0.07	0.29*	1 - 124,648	2,265
GOBY	0.07	0.01	0.12	0.26*	1 - 13,076	215
JOHN	0.21*	0.05	0.33*	0.25*	1 - 9,103	77
LTRT	0.39*	0.31*	0.41**	NA	1 - 732	6
PRCH	0.39*	0.20	0.55**	<0.01	1 - 2,664	52
SLIM	0.14	<0.01	0.17	<0.01	1 - 11,595	155
SMLT	<0.01	<0.01	0.11	0.41**	1 - 181,082	895
SPOT	0.37*	0.09	0.41**	0.45**	1 - 12,055	227
STK3	--	--	--	--	1 - 16,701	144
TRPR	0.23*	0.02	0.32*	0.41**	1 - 23,917	287

The standardized residuals versus the fitted values for the Lake Trout GWR showed that there was a cone like pattern indicating that there is not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the homogeneity of the variance assumption (Figure 40). When the Lake Trout standardized residuals were mapped both the highest (≥ 1.5) and lowest (≤ -1.5) deviations from the mean were dispersed throughout the study area (Figure 41).

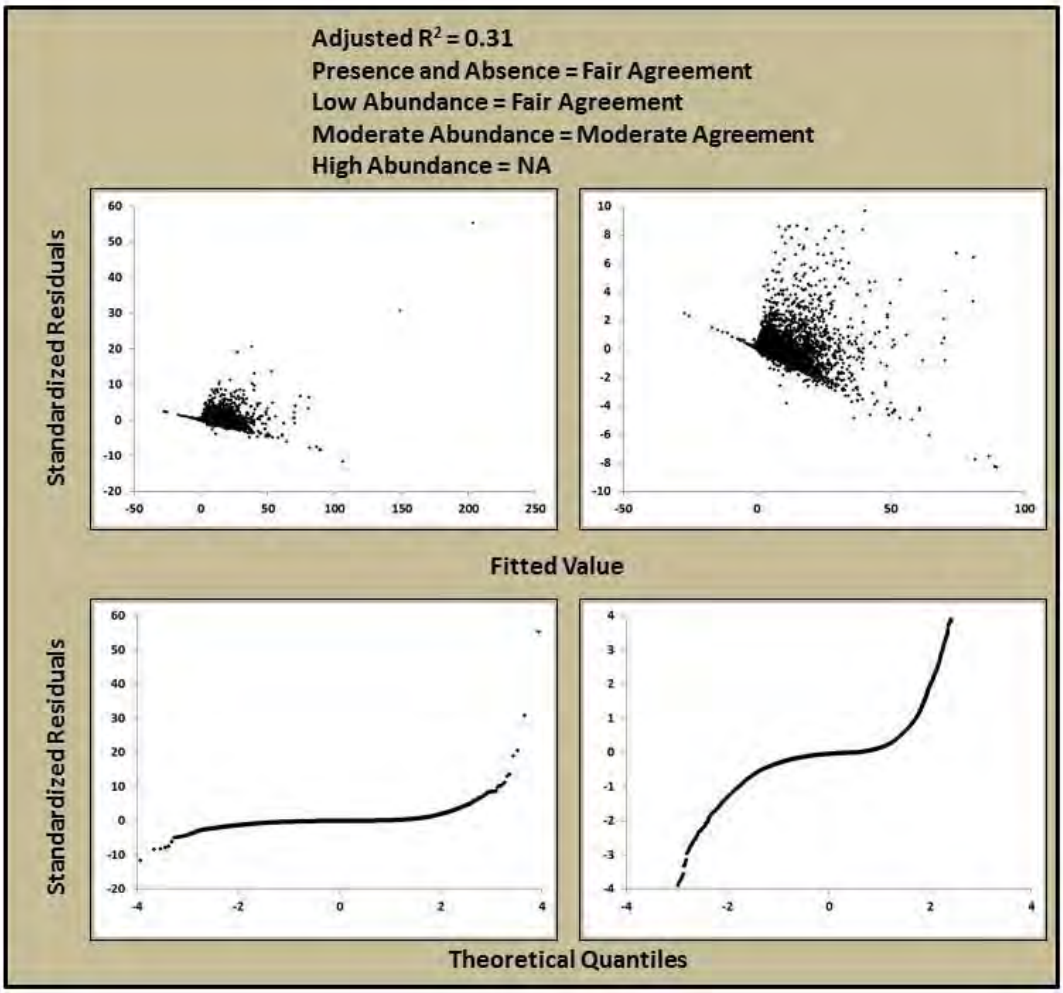

Figure 40 Standardized Residual versus Fitted Value and QQ Plot for Lake Trout GWR (1990-2014)

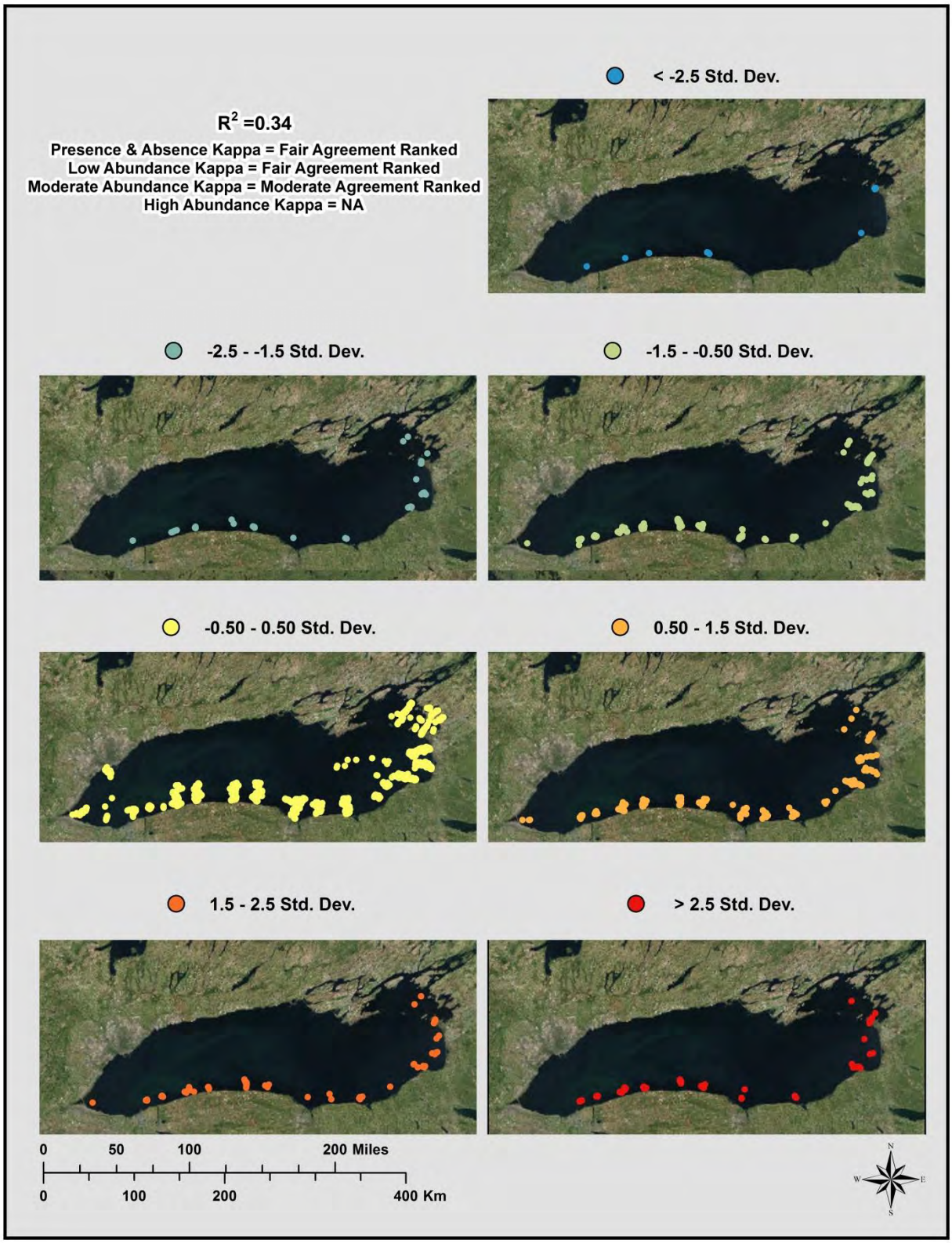

Figure 41 Spatial Distribution of Standardized Residuals for Lake Trout GWR (1990-2014)

The local R^2 values for Lake Trout ranged from less than 0.01 to 0.92. The higher values were dispersed around the lake (Figure 42). The trawling event located in deeper waters tended to have the lower local R^2 values.

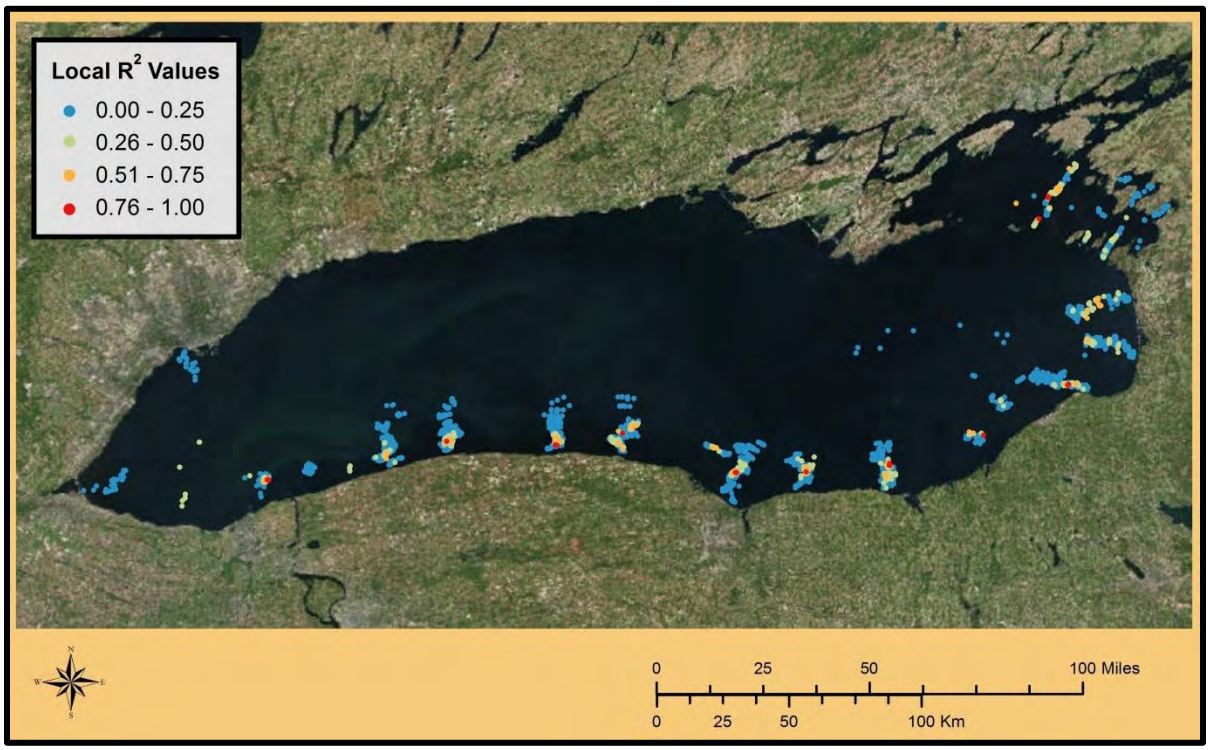

Figure 42 GWR Local R^2 Values for Lake Trout (1990 - 2014)

The coefficients for Lake Trout are shown in Figure 43. The coefficient for the square root of Rainbow Smelt abundance had a small range of values from -9.24 to 0.82. The highest values were mainly located in a number of areas in the eastern portion of the lake (Figure 43). The majority of the study area had lower values across the lake. The coefficient for the intercept had a large range of values from -10.7 to 285.9. The highest values were isolated in small areas in the eastern portion of the lake and also in areas in the western portion of the lake (Figure 43-A). The majority of areas had low values throughout the lake, with the lowest values primarily in the eastern portion of the lake (Figure 43-B).

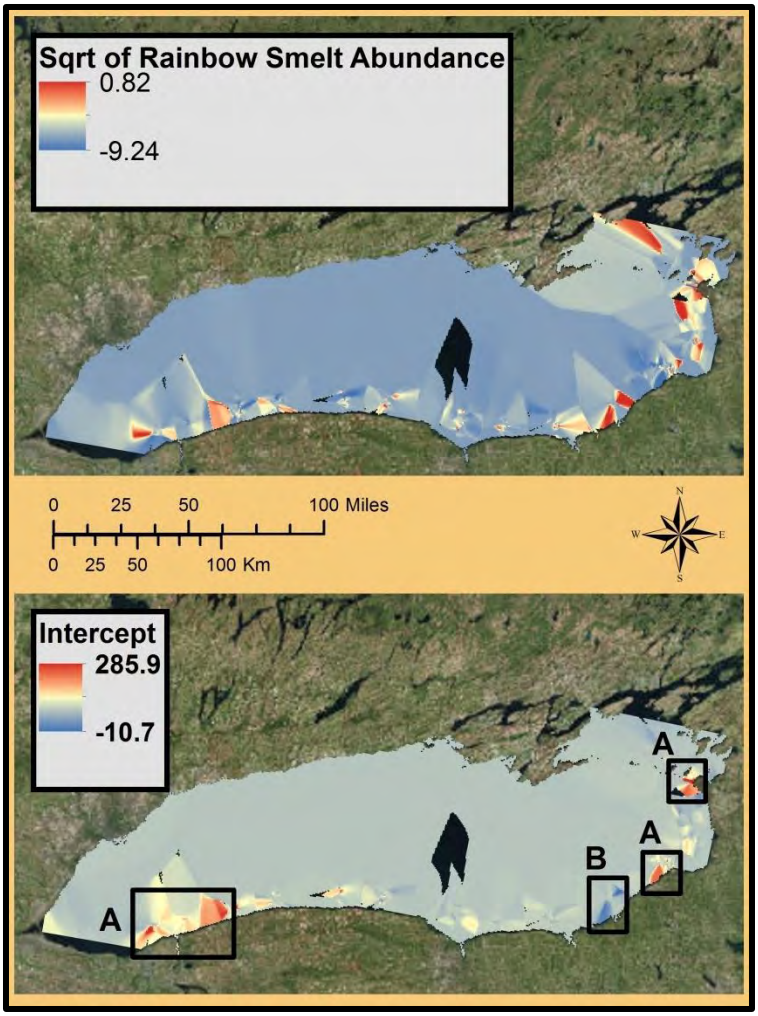

Figure 43 Lake Trout (1990-2014) GWR Coefficient Rasters

5.3.4. 2004 - 2014 dataset

The 2004 - 2014 dataset was used to model only for Round Goby and produced an adjusted R^2 value of 0.48 using 47 neighbors and only a single variable, temperature at fishing depth (Table 28). The results from the Cohen's Kappa showed that the GWR did better at predicting the moderate and high abundance classes, getting a fair and moderate agreement ranking respectively. The Round Goby GWR did not perform as well for the low abundance or the presence and absence which only got a slight agreement ranking between the observed and predicted values (Table 29).

Table 28 Adjusted R^2 value of the GWR for Round Goby (2004 - 2014)

Species	Adjusted R^2	Local R^2 Range	Neighbors	# of Variables
GOBY	0.48	<0.01 - 0.89	47	1

Table 29 Cohen's Kappa values of the GWR for Round Goby (2004 – 2014)
(*) denotes fair agreement, (**) denotes moderate agreement.

Species	Presence	Low Abundance	Moderate Abundance	High Abundance	When Species Present	
					Range of abundances	Average abundance
GOBY	0.14	0.01	0.21*	0.44**	1 - 13,076	215

The standardized residuals versus the fitted values for the Round Goby GWR showed that there was a non-random pattern indicating that there was not homogeneity of the variance. The QQ - Plot also indicates that the model did not meet the homogeneity of the variance assumption (Figure 44). When the Round Goby standardized residuals were mapped, the lowest (≤ -1.5) deviations below the mean were distributed in the central portion of the lake and the highest (≥ 1.5) deviations above the mean were dispersed in both the eastern and central portion of the lake (Figure 45).

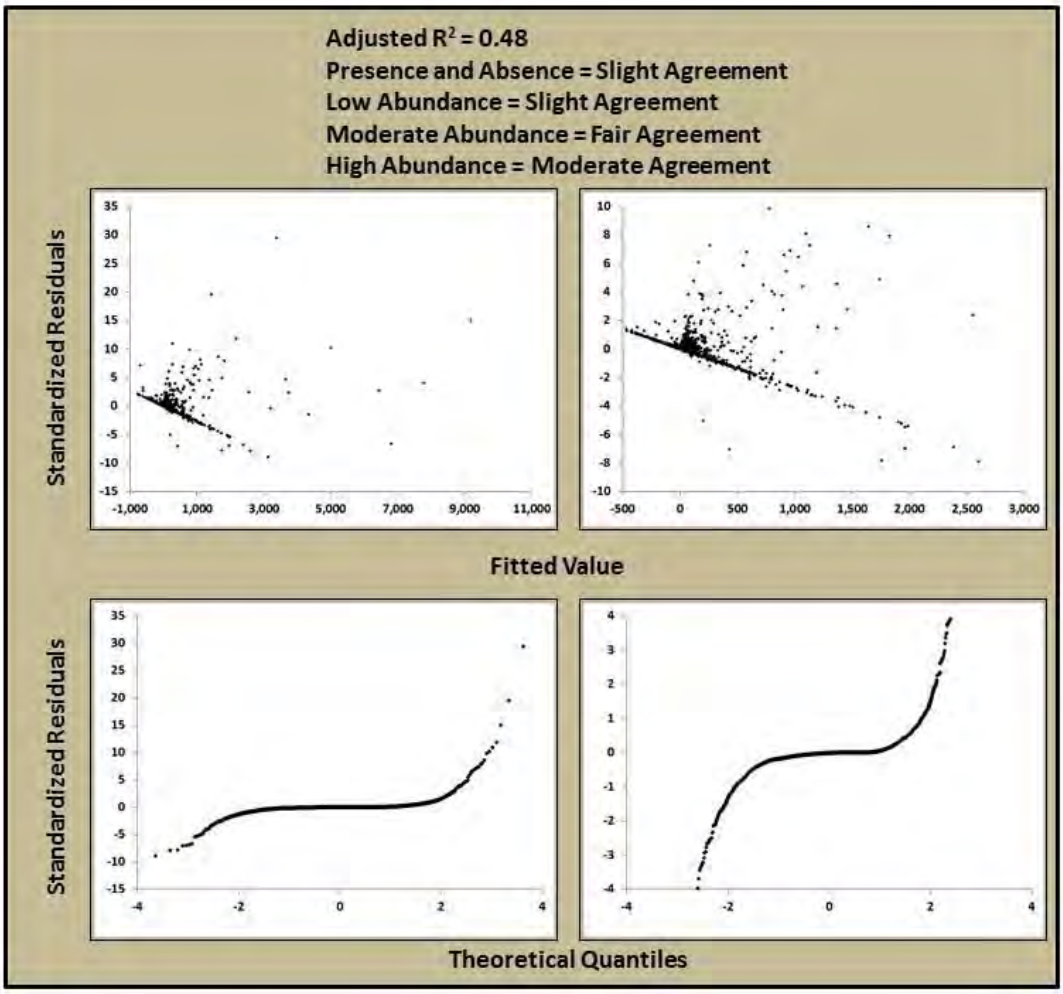

Figure 44 Standardized Residual versus Fitted Value and QQ Plot for Round Goby GWR (2004-2014)

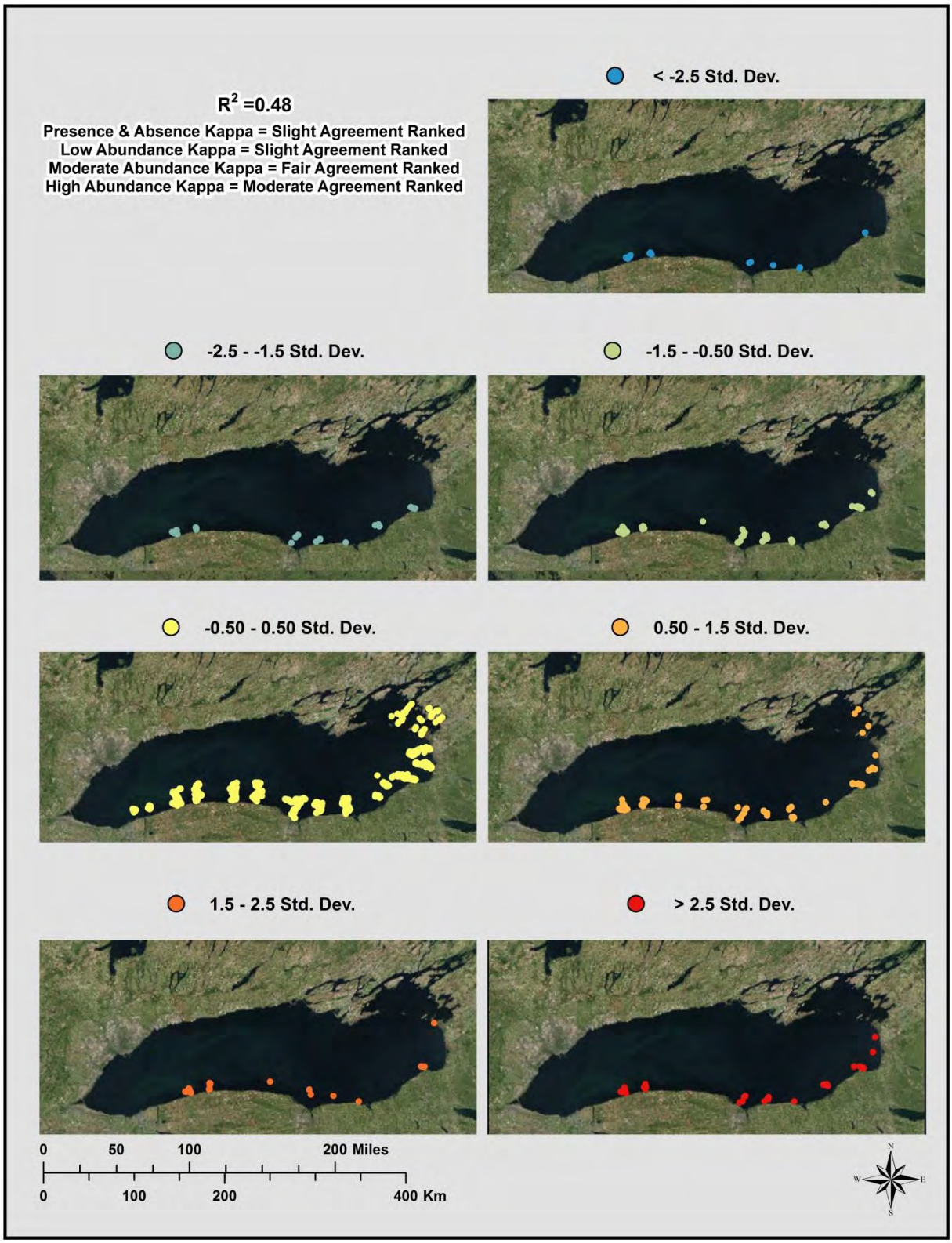

Figure 45 Spatial Distribution of Standardized Residuals for Round Goby GWR (2004-2014)

The local R^2 values for Round Goby ranged from less than 0.01 to 0.89. The higher values were dispersed around the lake (Figure 46). The trawling event located in deeper waters tended to have the lower local R^2 values.

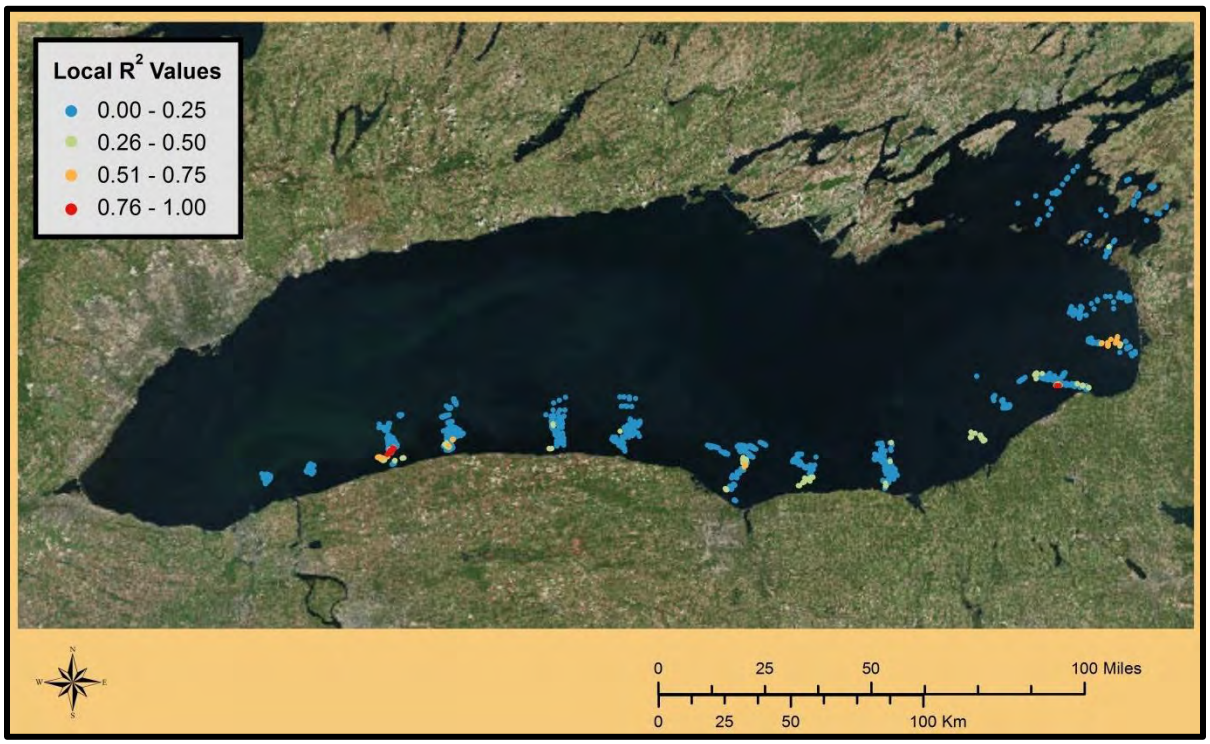

Figure 46 GWR Local R^2 Values for Round Goby (2004 - 2014)

The coefficients for Round Goby are shown in Figure 47. The coefficient for fishing depth temperature had a large range of values from -317 to 914.7. The highest values were mainly isolated to three small areas throughout the lake (Figure 47-A). The majority of the areas have lower values across the lake. The coefficient for the intercept also had a large range of values from -4172.1 to 1463.8. The lowest values were isolated in the western and eastern portion of the lake (Figure 47-B). The rest of the lake had relatively high values.

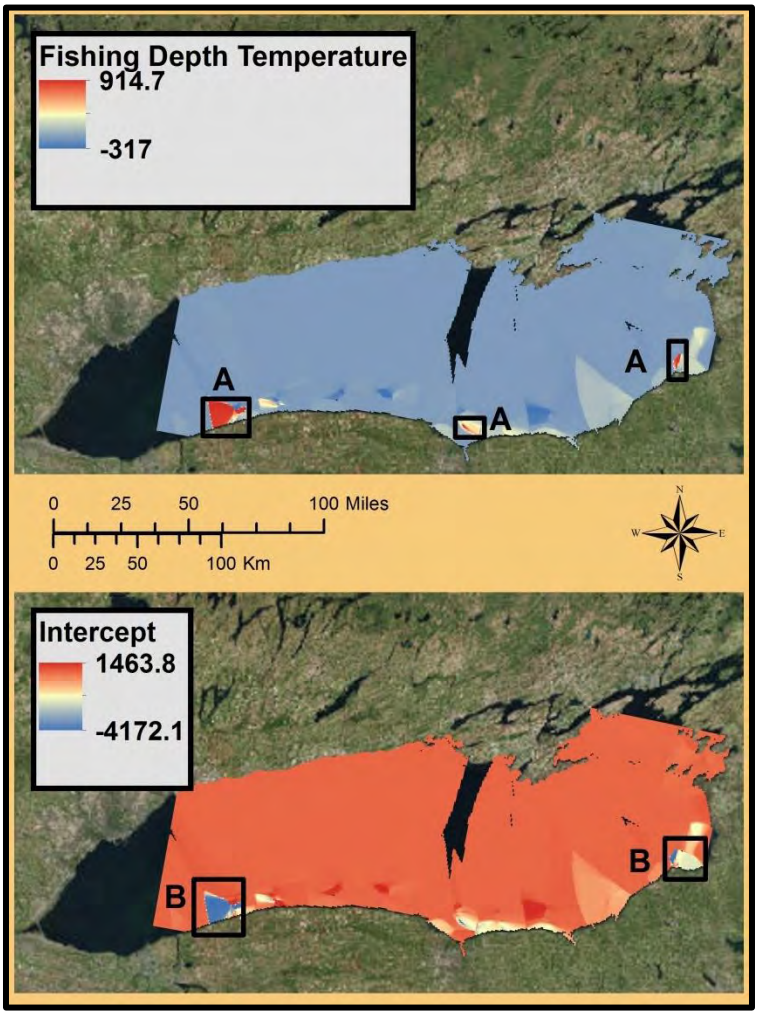

Figure 47 Round Goby (2004-2014) GWR Coefficient Rasters

5.4 Summary

The result of this analysis of successful models across the datasets and methods shows that the GAM method often achieved the highest adjusted R^2 values, most commonly with the Poisson distribution. The AIC values were often lowest with the GWR for the Gaussian models and GAM for the Poisson models. The GWR method was the only method to obtain Cohen's Kappa values that surpassed slight agreement in all categories. When the low abundance category was excluded the GAM and GWR methods often had the highest values. The GLM method was the least successful in both adjusted R^2 and Cohen's Kappa values. The next chapter focuses on comparative analysis of model results collectively.

Chapter 6 Model Comparison

The ultimate objective of this thesis was to compare the success of models generated by three different modeling methods. This chapter develops these comparisons in depth. Before comparing the models produced by GLM, GAM, and GWR, the different distribution types were compared between the GLM and GAM methods. Distribution types were compared using the adjusted R^2 and Cohen's Kappa values. Then a comparison was done between modeling methods using the adjusted R^2 and Cohen's Kappa values. A comparison of AIC values was done between the modeling methods within each distribution type to see which method achieved the lowest value for each distribution type. Model variables were also reviewed to determine if model relationships agreed with what was to be expected in reality.

6.1 Gaussian Distribution Versus Poisson Distribution

A comparison of Gaussian and Poisson distribution was only done with GLM and GAM, due to the fact that the Esri Spatial Analyst tool Geographically Weighted Regression did not have the option of using any other distribution family besides Gaussian. Sixty percent of the GLMs saw improvement in adjusted R^2 value when using a Poisson distribution with a change in value that ranged from a 0.02 to 0.35. Ninety percent of the GAMs saw improvement in adjusted R^2 value when using a Poisson distribution with a change in value that ranged from 0.01 to 0.57. The Cohen's Kappa values for the presence and absence category were commonly higher for the Gaussian distribution whereas the Poisson distribution had more success predicting the moderate and high abundance classes. Neither distribution had much success in predicting the low abundance class.

6.2 GLM, GAM & GWR Adjusted R^2 and AIC Comparison

The highest adjusted R^2 value for the 1978 - 2014 dataset was 0.48 using the GAM method with a Poisson distribution for Round Goby (Figure 48 & Table 30). The GAM with Poisson distribution received higher adjusted R^2 values for seven of the ten species. The GWR method achieved higher adjusted R^2 values for two of the ten species, and could not develop a model for Threespine Stickleback. Both GAM methods and the GWR received similar results for Trout Perch. The GWR method also commonly received the second highest adjusted R^2 value for the species being modeled. The GLM modeling methods performed the poorest. The GLM with Poisson only managed to outperform another method, the GAM with Gaussian distribution, with three species. While the GLM with Poisson distribution did outperform the GAM with Gaussian distribution it failed to be the top performing model.

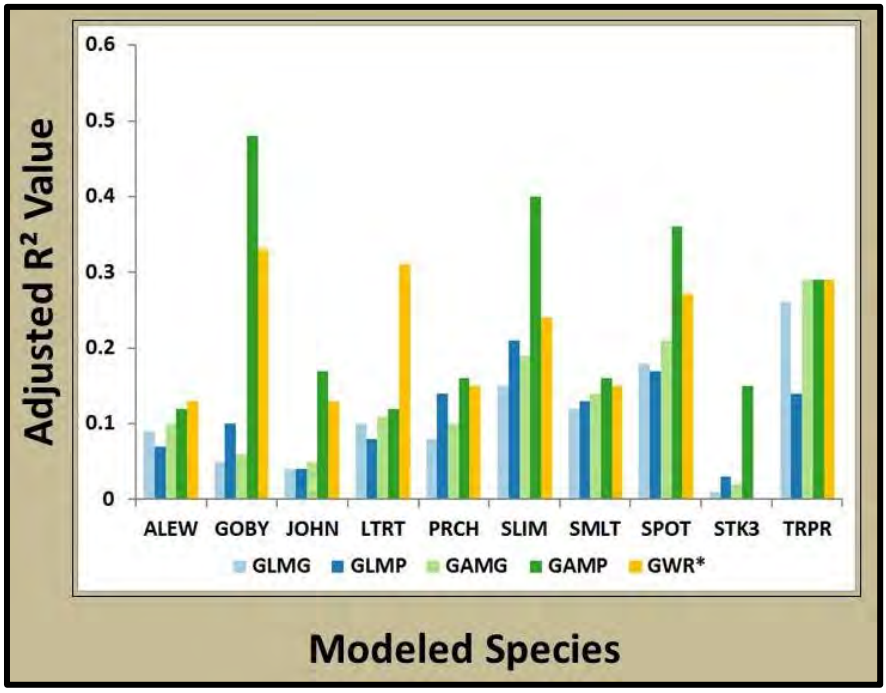

Figure 48 Comparison of Adjusted R^2 Values for GLM, GAM, and GWR (1978 - 2014). Alewife (ALEW), Round Goby (GOBY), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR). * signifies that the GWR could not develop a model for Threespine Stickleback

Table 30 Comparison of adjusted R^2 values for GLM, GAM, and GWR (1978 - 2014). Alewife (ALEW), Round Goby (GOBY), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR).

Species	**Adjusted R^2 Highest value ←→Lowest value**				
ALEW	0.13^{GWR}	0.12^{GAMP}	0.10^{GAMG}	0.09^{GLMG}	0.07^{GLMP}
GOBY	0.48^{GAMP}	0.33^{GWR}	0.10^{GLMP}	0.06^{GAMG}	0.05^{GLMG}
JOHN	0.17^{GAMP}	0.13^{GWR}	0.05^{GAMG}	0.04^{GLMG}	0.04^{GLMP}
LTRT	0.31^{GWR}	0.12^{GAMP}	0.11^{GAMG}	0.10^{GLMG}	0.08^{GLMP}
PRCH	0.16^{GAMP}	0.15^{GWR}	0.14^{GLMP}	0.10^{GAMG}	0.08^{GLMG}
SLIM	0.40^{GAMP}	0.24^{GWR}	0.21^{GLMP}	0.19^{GAMG}	0.15^{GLMG}
SMLT	0.16^{GAMP}	0.15^{GWR}	0.14^{GAMG}	0.13^{GLMP}	0.12^{GLMG}
SPOT	0.36^{GAMP}	0.27^{GWR}	0.21^{GAMG}	0.18^{GLMG}	0.17^{GLMP}
STK3	0.15^{GAMP}	0.03^{GLMP}	0.02^{GAMG}	0.01^{GLMG}	NA^{GWR}
TRPR	0.29^{GAMP}	0.29^{GWR}	0.29^{GAMG}	0.26^{GLMG}	0.14^{GLMP}

The AIC values from the 1978 - 2014 dataset showed that between the Gaussian models, GWR had the lowest values, indicating the better model (Table 31). Threespine Stickleback could not develop a model using GWR so GAM was shown to have the lowest AIC value. Trout Perch which could get a GWR model to develop also had GAM produce the lowest AIC value. Between the GLM and GAM with Poisson distribution the GAM had the lowest AIC value for all species besides the Slimy Sculpin model.

Table 31 ΔAIC values for GLM, GAM, & GWR models based on distribution type (1978 - 2014)

Species	ΔAIC				
	GLMG	GAMG	GWR	GLMP	GAMP
ALEW	468	354	0	2,546,802	0
GOBY	3,826	3,735	0	367,204	0
JOHN	1,082	925	0	231,527	0
LTRT	2,108	1,911	0	9,968	0
PRCH	959	585	0	10,420	0
SLIM	1,282	718	0	0	952,690
SMLT	411	76	0	3,110,170	0
SPOT	1,406	953	0	152,350	0
STK3	129	0	NA	412,754	0
TRPR	441	0	176	690,503	0

The 1978 - 1989 dataset developed the highest adjusted R^2 value of all the datasets with a value of 0.74 for Spottail Shiner (Figure 49 & Table 32). The GAM method obtained the highest adjusted R^2 values for seven of the nine species, with the GAM using Poisson distribution being the method that often got the highest values of the GAMs. The GWR method obtained a higher adjusted R^2 values for only two of the nine species. The GWR failed to develop a model for Threespine Stickleback as was the case with the 1978 - 2014 dataset. The GLM with Poisson distribution saw an increase in performance with this dataset by receiving the second highest adjusted R^2 values for four out of the nine species.

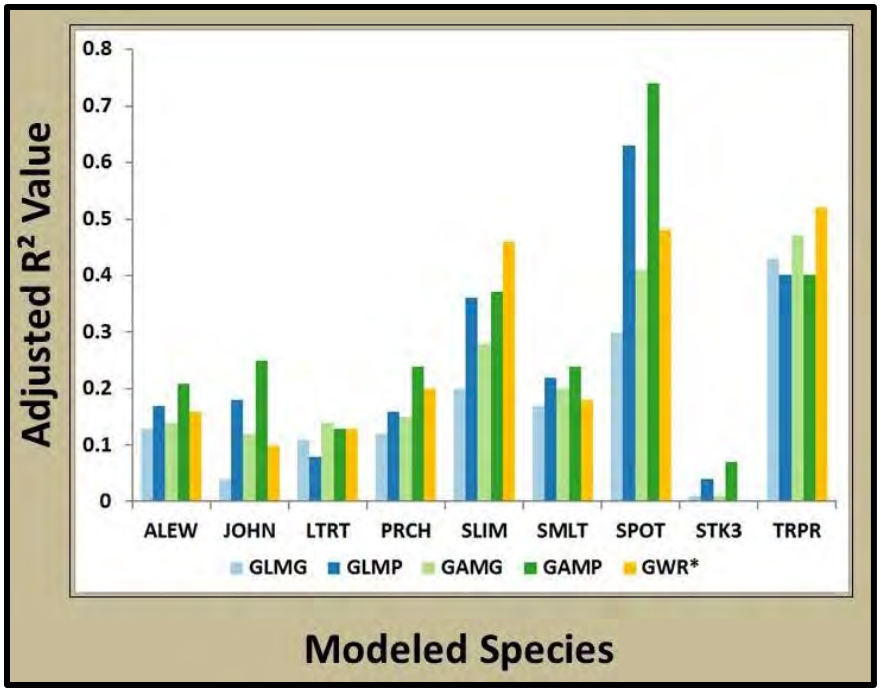

Figure 49 Comparison of Adjusted R^2 Values for GLM, GAM, and GWR (1978 - 1989). Alewife (ALEW), Round Goby (GOBY), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR). * signifies that the GWR could not develop a model for Threespine Stickleback

Table 32 Comparison of adjusted R^2 values for GLM, GAM, and GWR (1978 - 1989). Alewife (ALEW), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR).

Species	**Adjusted R^2 Highest value ←→Lowest value**				
ALEW	0.21^{GAMP}	0.17^{GlMP}	0.16^{GWR}	0.14^{GAMG}	0.13^{GLMG}
JOHN	0.25^{GAMP}	0.18^{GLMP}	0.12^{GAMG}	0.10^{GWR}	0.04^{GLMG}
LTRT	0.14^{GAMG}	0.13^{GAMP}	0.13^{GWR}	0.11^{GLMG}	0.08^{GLMP}
PRCH	0.24^{GAMP}	0.20^{GWR}	0.16^{GLMP}	0.15^{GAMG}	0.12^{GLMG}
SLIM	0.46^{GWR}	0.37^{GAMP}	0.36^{GLMP}	0.28^{GAMG}	0.20^{GLMG}
SMLT	0.24^{GAMP}	0.22^{GLMP}	0.20^{GAMG}	0.18^{GWR}	0.17^{GLMG}
SPOT	0.74^{GAMP}	0.63^{GLMP}	0.48^{GWR}	0.41^{GAMG}	0.30^{GLMG}
STK3	0.07^{GAMP}	0.04^{GLMP}	0.01^{GAMG}	0.01^{GLMG}	NA^{GWR}
TRPR	0.52^{GWR}	0.47^{GAMG}	0.43^{GLMG}	0.40^{GAMP}	0.40^{GLMP}

The 1978 - 1989 dataset AIC values showed that between the Gaussian models GWR and GAM had the lowest values depending on the species (Table 33). Threespine Stickleback could not develop a model using GWR and had only a value of three differences making it difficult to identify which modeling method was better. Between the GLM and GAM with Poisson distribution the GAM had the lowest AIC value for all species models.

Table 33 ΔAIC values for GLM, GAM, & GWR models based on distribution type (1978 - 1989)

Species	ΔAIC				
	GLMG	**GAMG**	**GWR**	**GLMP**	**GAMP**
ALEW	113	113	0	1,073,431	0
JOHN	326	0	111	49,894	0
LTRT	111	0	27	7,756	0
PRCH	313	177	0	5,298	0
SLIM	1,464	1,087	0	46,174	0
SMLT	91	0	105	433,941	0
SPOT	1,064	409	0	31,323	0
STK3	3	0	NA	208	0
TRPR	679	369	0	81,163	0

The 1990 - 2014 dataset achieved another of the highest adjusted R^2 values, 0.71 for Spottail Shiner (Figure 50 & Table 34). The GAM with Poisson distribution, like the previous dataset results, received the highest adjusted R^2 values for eight of ten species. The GWR method was able to get the highest adjusted R^2 values for two of the ten species. The Lake Trout GWR saw a considerable increase compared to the other methods. GWR also developed the second highest value for four of the ten models The GWR failed to develop a model for Threespine Stickleback as was the class with both of the previous datasets. The GLM with Poisson distribution once again saw an increase in performance with this dataset than was seen with the 1978 - 2014 dataset.

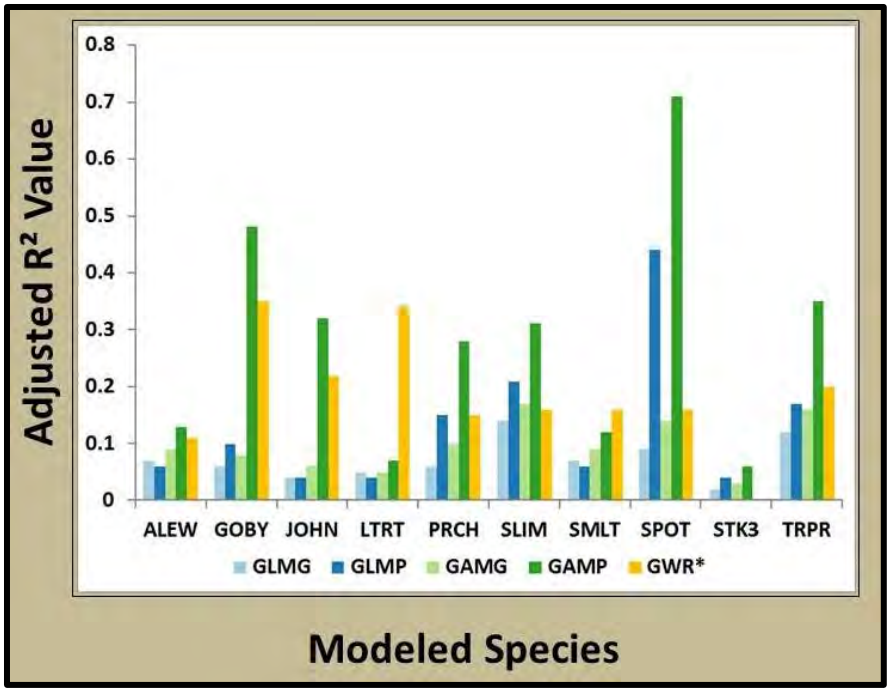

Figure 50 Comparison of Adjusted R^2 Values for GLM, GAM, and GWR (1990 - 2014). Alewife (ALEW), Round Goby (GOBY), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR). * signifies that the GWR could not develop a model for Threespine Stickleback

Table 34 Comparison of adjusted R^2 values for GLM, GAM, and GWR (1990 - 2014). Alewife (ALEW), Round Goby (GOBY), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR).

Species	Adjusted R^2 Highest value ←→Lowest value				
ALEW	0.13^{GAMP}	0.11^{GWR}	0.09^{GAMG}	0.07^{GLMG}	0.06^{GLMP}
GOBY	0.48^{GAMP}	0.35^{GWR}	0.10^{GLMP}	0.08^{GAMG}	0.06^{GLMG}
JOHN	0.32^{GAMP}	0.22^{GWR}	0.06^{GAMG}	0.04^{GLMG}	0.04^{GLMP}
LTRT	0.34^{GWR}	0.07^{GAMP}	0.05^{GAMG}	0.05^{GLMG}	0.04^{GLMP}
PRCH	0.28^{GAMP}	0.15^{GWR}	0.15^{GLMP}	0.10^{GAMG}	0.06^{GLMG}
SLIM	0.31^{GAMP}	0.21^{GLMP}	0.17^{GAMG}	0.16^{GWR}	0.14^{GLMG}
SMLT	0.16^{GWR}	0.12^{GAMP}	0.09^{GAMG}	0.07^{GLMG}	0.06^{GLMP}
SPOT	0.71^{GAMP}	0.44^{GLMP}	0.16^{GWR}	0.14^{GAMG}	0.09^{GLMG}
STK3	0.06^{GAMP}	0.04^{GLMP}	0.03^{GAMG}	0.02^{GLMG}	NA^{GWR}
TRPR	0.35^{GAMP}	0.20^{GWR}	0.17^{GLMP}	0.16^{GAMG}	0.12^{GLMG}

The 1990 - 2014 dataset AIC values showed that between the Gaussian models GWR had the lowest values for all species besides Slimy Sculpin and Threespine Stickleback (Table 35). Threespine Stickleback could not develop a model using GWR and had only a value of three differences making it difficult to identify which modeling method was better. Between the GLM and GAM with Poisson distribution the GAM had the lowest AIC value for all species models.

Table 35 ΔAIC values for GLM, GAM, & GWR models based on distribution type (1990 - 2014)

Species	ΔAIC				
	GLMG	GAMG	GWR	GLMP	GAMP
ALEW	287	122	0	3,483,420	0
GOBY	2,758	2,672	0	359,398	0
JOHN	1,536	1,412	0	167,300	0
LTRT	2,109	2,071	0	4,907	0
PRCH	761	393	0	12,664	0
SLIM	214	0	141	229,169	0
SMLT	717	564	0	1,999,770	0
SPOT	608	209	0	110,352	0
STK3	64	0	NA	53,431	0
TRPR	788	458	0	216,431	0

The 2004 - 2014 dataset, which only modeled Round Goby achieved similar adjusted R^2 values for GAMP (0.49) and GWR (0.48) (Figure 51 & Table 36). The Gaussian distribution for both GLM and GAM achieved the lowest values. The GLM with a Poisson (0.26) did better than the GLM and GAM with Gaussian distribution.

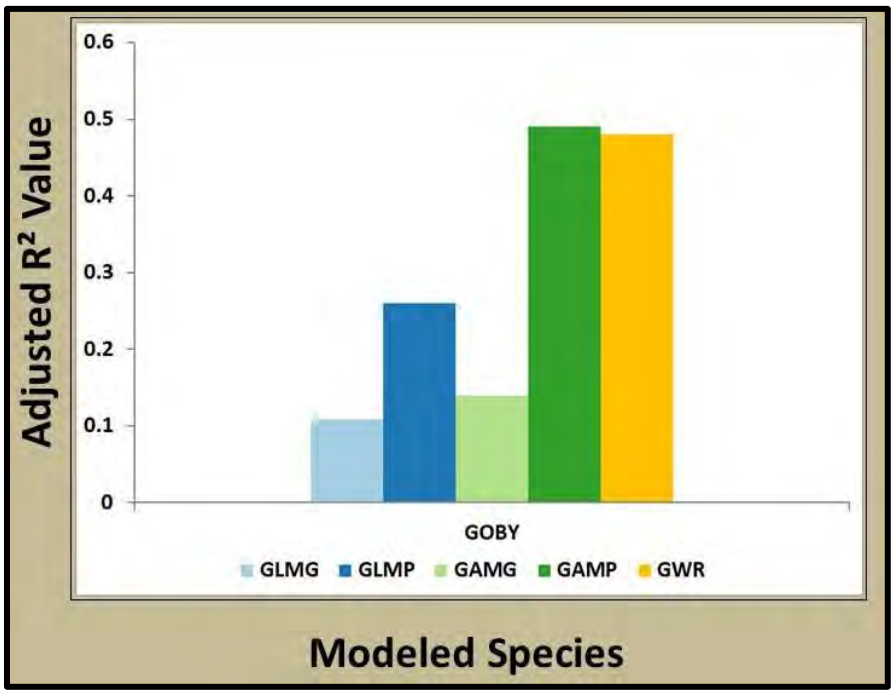

Figure 51 Comparison of Adjusted R^2 Values for GLM, GAM, and GWR (2004 - 2014). Round Goby (GOBY)

Table 36 Comparison of adjusted R^2 values for GLM, GAM, and GWR (2004 - 2014). Round Goby (GOBY).

Species	Adjusted R^2 Highest value ←→Lowest value				
GOBY	0.49^{GAMP}	0.48^{GWR}	0.26^{GLMP}	0.14^{GAMG}	0.11^{GLMG}

The 2004 - 2014 dataset AIC values showed that between the Gaussian models the GWR method had the lowest values for Round Goby (Table 37). Between the GLM and GAM with Poisson distribution the GAM had the lowest AIC value of the two methods.

Table 37 ΔAIC values for GLM, GAM, & GWR models based on distribution type (2004 – 2014)

Species	ΔAIC				
	GLMG	**GAMG**	**GWR**	**GLMP**	**GAMP**
GOBY	1,672	1,548	0	184,450	0

6.3 GLM, GAM, & GWR Cohen's Kappa Comparison

The Cohen's Kappa values for the 1978 - 2014 dataset showed that the GWR method was the only method to get a fair or better agreement ranking between observed and predicted values in each abundance category for a single species, Lake Trout (Figure 52). It should be noted that Lake Trout did not have abundances that fell within the high abundance category. All modeling methods had the most difficulty with predicting the low abundance category. None of the methods were able to successfully predict the high abundance category of Yellow Perch. Besides the low abundance category the GWR method was able to get a fair or higher agreement ranking in the other categories for Spottail Shiner and Trout Perch. The GAM with Poisson distribution was able to get a fair or better agreement ranking for every category besides low abundance for Spottail Shiner and Threespine Stickleback.

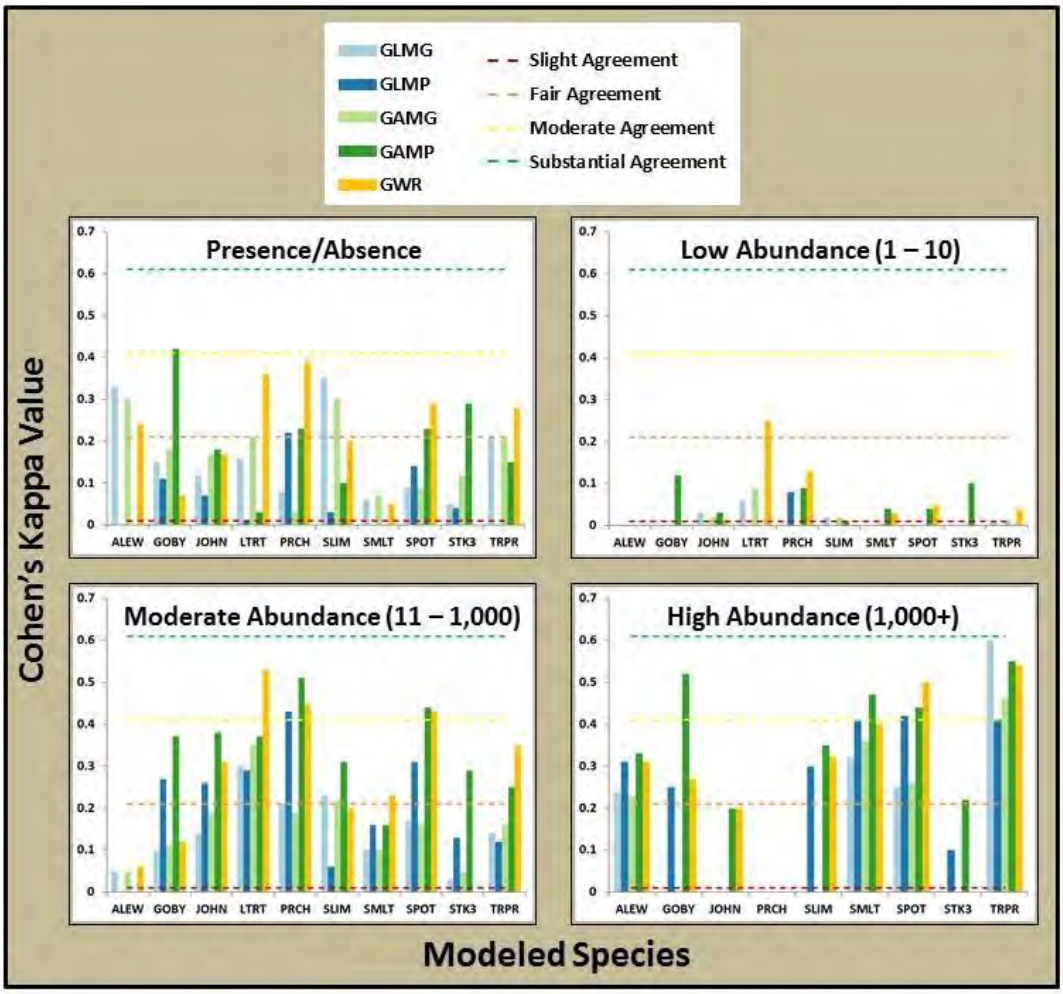

Figure 52 Cohen's Kappa Values for the 1978 - 2014 Dataset. Alewife (ALEW), Round Goby (GOBY), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Rainbow Smelt (SMLT), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR)

The Cohen's Kappa values for the 1978 - 1989 dataset also showed that all methods had difficulty predicting the low abundance category for all species and the high abundance category for Yellow Perch (Figure 53). Slimy Sculpin was able to get a fair or better agreement ranking between observed and predicted values in all categories, other than the low abundance category, using the GAM with Gaussian distribution and GWR methods. Spottail Shiner also achieved a fair or better ranking in all categories but low abundance with the GLM with Gaussian

distribution, GAM with Poisson distribution, and GWR methods. For the Spottail Shiner models the GWR and GAM outperformed the GLM by having higher values.

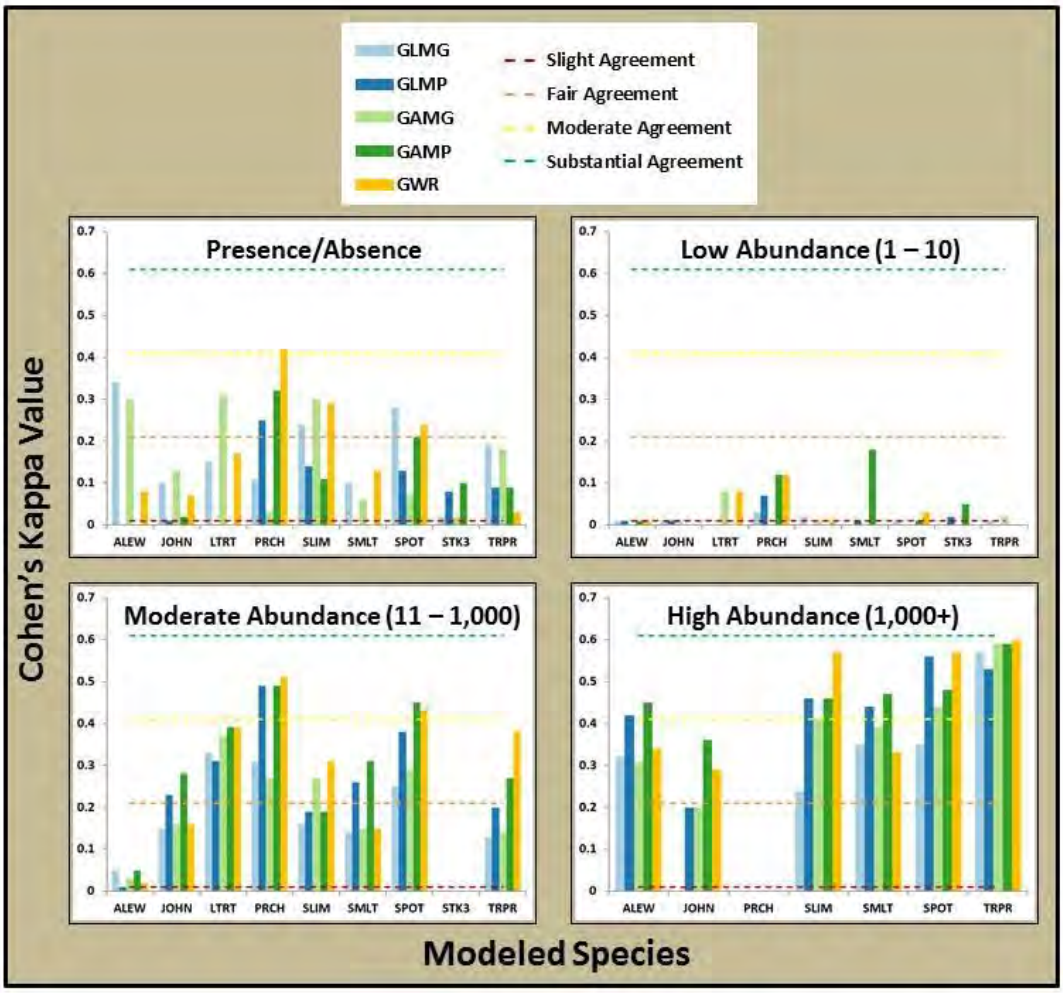

Figure 53 Cohen's Kappa Values for the 1978 - 1989 Dataset. Alewife (ALEW), Round Goby (GOBY), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Rainbow Smelt (SMLT), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR)

The 1990 - 2014 dataset Cohen Kappa values continued to show that the low abundance category was the most difficult to predict (Figure 54). Of all the models only the Lake Trout model using the GWR method achieved fair or better agreement rankings between observed and predicted values for all categories. When the low abundance category was ignored Round Goby (GAM with Poisson distribution), Johnny Darter (GWR & GAM with Poisson distribution),

Yellow Perch (GAM with Poisson distribution), Spottail Shiner (GWR & GAM with Poisson distribution), and Trout Perch (GWR) obtained a fair or better agreement ranking in all other categories.

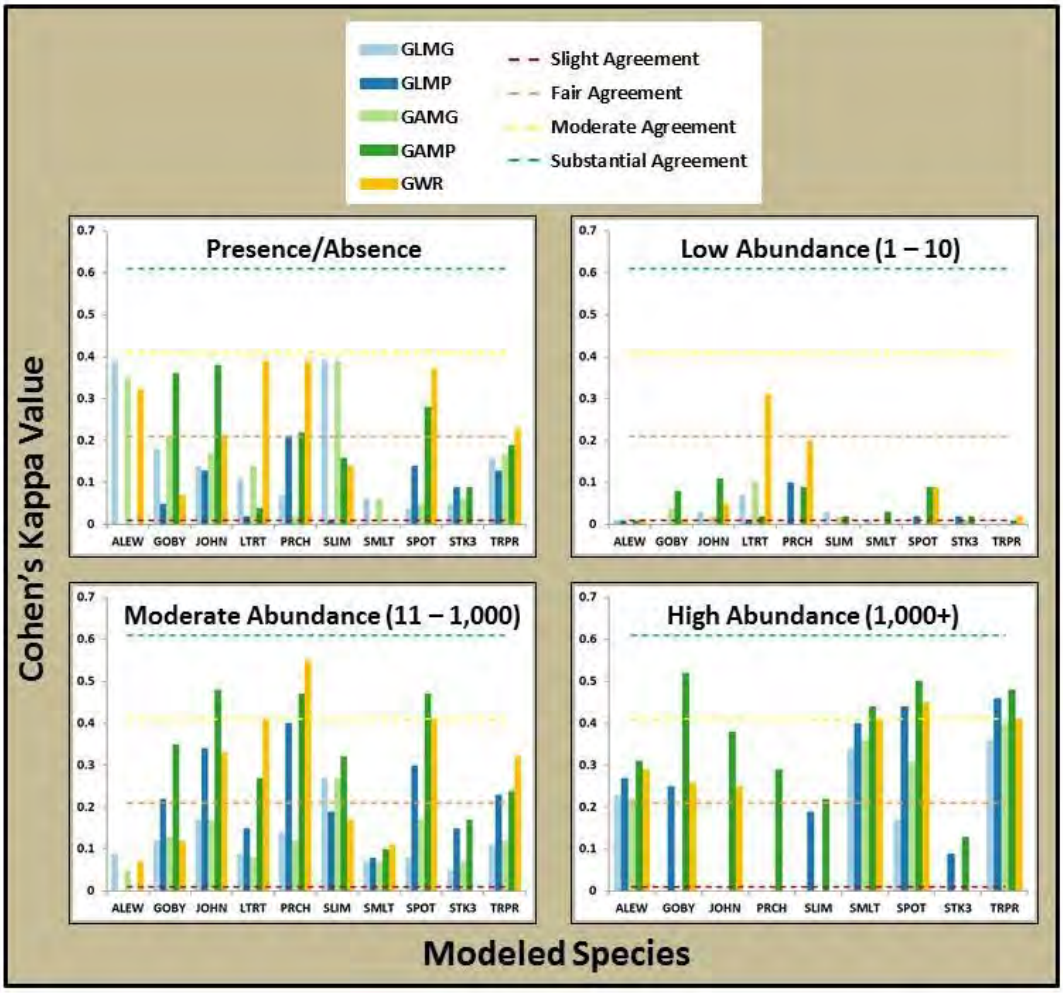

Figure 54 Cohen's Kappa Values for the 1990 - 2014 Dataset. Alewife (ALEW), Round Goby (GOBY), Johnny Darter (JOHN), Lake Trout (LTRT), Yellow Perch (PRCH), Slimy Sculpin (SLIM), Rainbow Smelt (SMLT), Spottail Shiner (SPOT), Threespine Stickleback (STK3), Trout Perch (TRPR)

The 2004 - 2014 dataset saw poor overall results for all abundance categories (Figure 55). As was the case with all other datasets the low abundance category was the most difficult to predict. Of all the methods used the GWR was the only method to produce a slight or better agreement ranking between observed and predicted values in all categories. The GAM with

either distribution method was the second best performing method producing a slight or better agreement ranking for all categories besides the low abundance.

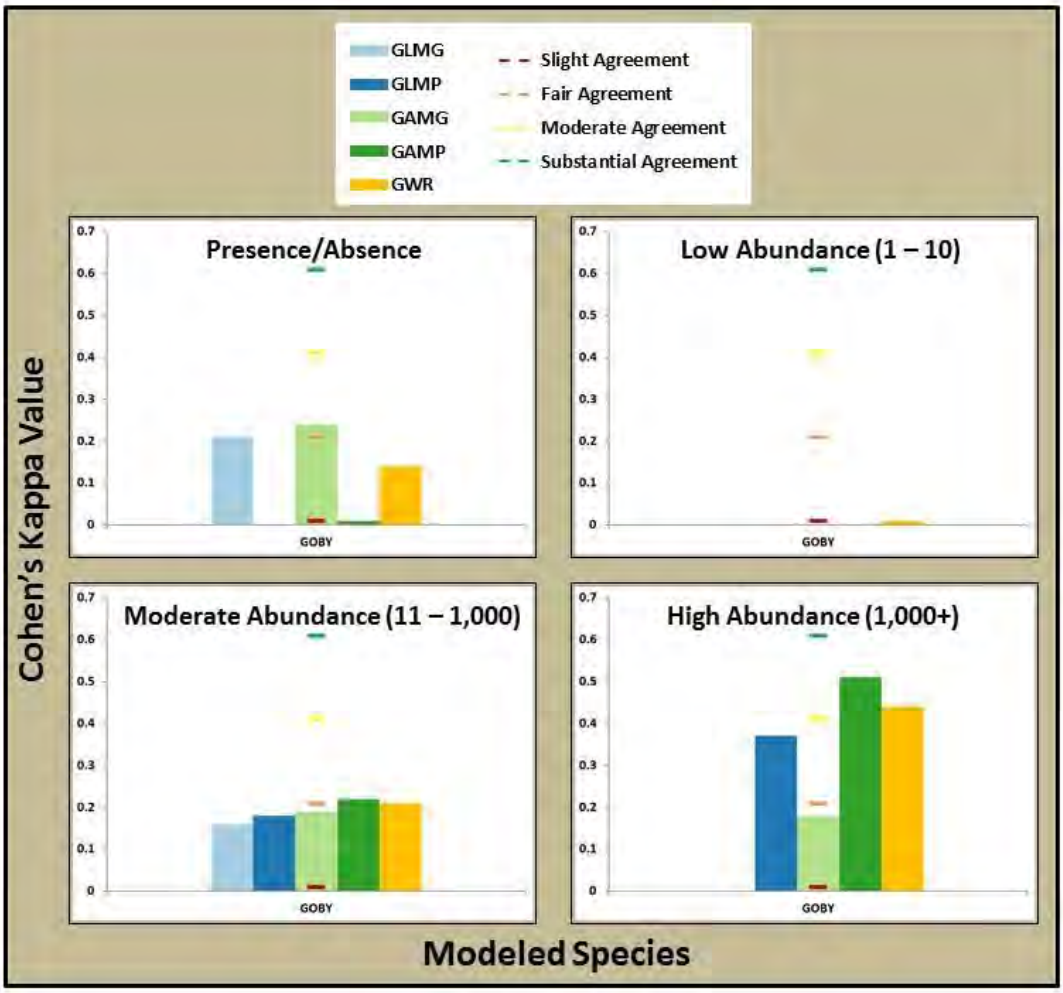

Figure 55 Cohen's Kappa Values for the 2004 - 2014 Dataset. Round Goby (GOBY)

The Cohen's Kappa values for all datasets showed that the GWR method was the only method to achieve a rank above slight agreement between observed and predicted values for the low abundance class. The GAM method also produced some of the highest values for all categories except for low abundance. The GLM method rarely produced agreement rankings between the observed and predicted values above slight agreement.

6.4 Assessment of Model Function and Structure

As discussed above, all GLMs and GAMs were fitted with five variables. The GWR models include five or fewer predictor variables depending on the species. The reduction of variables below the parsimonious initial goal of five was necessary due to local multicollinearity. As was stated in the methods, only variables that had a p-value of 0.05 or less were used, indicating that they were significant. All predictor variables within each model set had a VIF value below 7.5, which is an indication that collinearity among predictor variables is low.

Since the variables included in each best model changed with the dataset and/or the method used, any variable that repeatedly appeared can shed light on which variables have the most importance. As a further means of assessing model success, it is useful to assess whether the relationships shown by the different sets of variables show the same overall pattern and make sense.

Because of the nonlinear nature of the GAM and the local regression method of the GWR, variables used in the GAM and GWR models could have both positive and negative relationships. The GAMs in this study allowed for smoothing functions with three degrees of freedom, two curves. Relationships between response variables and predictor variables in a GWR are heavily influenced on geographic location. Due to the geographic change of coefficients with a GWR, it can be difficult describing the relationships between predictors and response variables. However, the GWR tool used in this study produces coefficient surface rasters as seen earlier in this chapter. Because of the difficulty discussing GWR coefficients this section focuses only on GLM and GAM coefficient relationships.

Month was a recurring variable in the GLM and GAMs, showing up in the majority of the successful models; however the direction of the relationship of individual months changes

depending on the species, the method used and the dataset. The likely reason for the reoccurrence of Month as a predictor variable is that it can represent the time of the year when some fish species assemble together for some activity, therefore increasing their numbers for easier capture. These activities could include spawning or feeding on booms in primary production in select areas.

For example the month of May consistently had a positive relationship with the abundance of Alewife. This relationship is reasonable due to its connection to Alewife spawning which occurs in and around the month of May (Durbin et al. 1979). Another example of the month variable highlighting spawning time is with Slimy Sculpin. Slimy Sculpin spawning is triggered by water temperature and at greater depth the water temperature is not often met until later in the year (Owens and Noguchi 1998). The model for Slimy Sculpin showed that the month of October, the time deeper waters would start becoming warmer, was consistently positive for all models. Johnny Darter was another species that showed a primarily positive relationship between the month of October and the species' abundance for the 1978 - 2014 and 1990 - 2014 datasets. The 1978 - 1989 dataset saw a negative relationship for the month of October. Past research showed that in October Johnny Darter was more present in Lake Trout diets (Elrod and O'Gorman 1991). The decrease in Lake Trout population could explain why the relationship is negative for the 1978 - 1989, when Lake Trout were more present, than the other datasets. The reason for the positive relationship is unclear but could be related to seasonal migration or spawning.

Exploring all the relationships between month and individual species would be extensive and require in-depth knowledge of the life cycles of each species. The Alewife, Slimy Sculpin,

and Johnny Darter examples do suggest that month is an important variable and could explain the increased abundance of some species at a location during one month and not another.

6.4.1. Alewife model structure

For Alewife, the common variables are the square roots of Slimy Sculpin and Rainbow Smelt abundances (Table 38). The square root of abundance for Slimy Sculpin was only used in nine of the fifteen models for Alewife. The relationship between Slimy Sculpin and Alewife was negative for the GLMs and primarily negative in the GAMs (Table 39). The negative relationship primarily seen with the Slimy Sculpin is reasonable due to the potential predation that could occur from Slimy Sculpin. The Rainbow Smelt relationship with Alewife was primarily positive for both the GLMs and GAMs. The primarily positive relationship between Alewife and Rainbow Smelt is reasonable as they have similar dietary habits (Lantry and Stewart 1993).

Table 38 Variables included in Alewife models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	ProtDist_Log, SqrtSLIM, SqrtLTRT, SqrtSMLT	Month
		Poisson	SqrtSMLT, SqrtSLIM, SqrtLTRT, OpenDist_Log	Month
	1978-1989	Gaussian	Year, ProtDist_Log, OpenDist, Depth	Month
		Poisson	SqrtSLIM, Year, ProtDist_Log, SqrtLTRT	Month
	1990-2014	Gaussian	Temp, Depth, SqrtSLIM, SqrtSMLT	Month
		Poisson	SqrtSMLT, SqrtSLIM, SqrtLTRT, Depth	Month
GWR	1978-2014	Gaussian	SqrtSMLT	Month(Apr), Month(May), Month(Jun)
	1978-1989	Gaussian	Depth	Month(Apr), Month(May), Month(Jun)
	1990-2014	Gaussian	SqrtSMLT	Month(Apr), Month(May), Month(Jun), Month(Oct)

Table 39 Relationships of reoccurring model variables for Alewife

SqrtSLIM	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	-	-	- / + / -	- / + / -	**NA**
1978-1989	**NA**	-	**NA**	- / + / -	**NA**
1990-2014	-	-	- / + / -	- / + / -	**NA**
SqrtSMLT	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+ / - / +	+ / - / +	- / +
1978-1989	**NA**	**NA**	**NA**	**NA**	**NA**
1990-2014	+	+	+ / - / +	- / +	- / +

6.4.2. Round Goby model structure

For Round Goby, the most common variable is temperature at fishing depth (Table 40). Temperature at fishing depth occurred in every model developed. As Table 41 shows, for the GLM models the relationship was positive, indicating that Round Goby were higher in abundance as water temperature increased. The relationship between temperature and Round

Goby for the GAM show that the relationship is positive until the water gets too warm and then abundances start to decrease. The trend of abundance increasing until a certain temperature then dropping off is more reasonable than the strict positive relationship seen in the GLM. The reasoning is that while Round Goby enjoy warmer waters, they do not enjoy when it gets too hot.

Table 40 Variables included in Round Goby models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	Temp, Year, ProtDist, SqrtJOHN	Month
		Poisson	Year, Depth, Temp, ProtDist	Month
	1978-1989	Gaussian	Temp, ProtDist, Year, SqrtJOHN	Month
		Poisson	Year, Depth, Temp, ProtDist	Month
	1990-2014	Gaussian	Temp, ProtDist, SqrtSMLT, SqrtTRPR	Month
		Poisson	Depth, Temp, ProtDist, SqrtAlew	Month
GWR	1978-2014	Gaussian	Temp	
	1978-1989	Gaussian	Temp	
	1990-2014	Gaussian	Temp	

Table 41 Relationships of reoccurring model variables for Round Goby

Round Goby					
Temp	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+ / -	+ / - / +	- / +
1990-2014	+	+	+ / -	+ / -	- / +
2004-2014	+	+	+ / -	+ / -	- / +

6.4.3. Johnny Darter model structure

For Johnny Darter the most common variables are temperature at fishing depth and the square root of Slimy Sculpin (Table 42). The square root of Slimy Sculpin variable was present in all models except for a single GWR, 1978 - 1989 dataset, and showed a primarily positive relationship with the abundance of Johnny Darter for the GLMs and GAMs (Table 43). The temperature at fishing depth variable was used in each model and showed that Johnny Darter had

a positive relationship to a point, when it would become negative. The positive relationship between Slimy Sculpin and Johnny Darter is not fully understood, but could indicate similar diets or habitat preferences.

Table 42 Variables included in Johnny Darter models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	Temp, SqrtSLIM, SqrtSPOT, SqrtPRCH	Month
		Poisson	Depth, Temp, SqrtGOBY, SqrtSLIM	Month
	1978-1989	Gaussian	Temp, SqrtSPOT, SqrtPRCH, SqrtSLIM	Month
		Poisson	Temp, SqrtSLIM, SqrtLTRT, SqrtSPOT	Month
	1990-2014	Gaussian	Temp, SqrtSLIM, SqrtSMLT, SqrtSTK3	Month
		Poisson	Depth, Year, SqrtSLIM, Temp	Month
GWR	1978-2014	Gaussian	Temp, SqrtSLIM	
	1978-1989	Gaussian	SqrtSPOT, Temp	
	1990-2014	Gaussian	Temp, SqrtSLIM	

Table 43 Relationships of reoccurring model variables for Johnny Darter

Johnny Darter					
SqrtSLIM	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+ / - / +	+ / - / +	+ / -
1978-1989	+	+	+ / -	+ / - / +	**NA**
1990-2014	+	+	+ / - / +	+ / - / +	+ / -
Temp	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+ / -	+ / -	+ / -
1978-1989	+	+	+ / - / +	+	+ / -
1990-2014	+	+	+ / -	+ / -	+ / -

6.4.4. Lake Trout model structure

For Lake Trout, the most common variable is the square root of Rainbow Smelt abundance which was appears in every model (Table 44). The relationship between Rainbow Smelt and Lake Trout was primarily positive in all models for GLM and GAM (Table 45). This

is an expected relationship with Lake Trout, as Rainbow Smelt is a key prey fish for them (Elrod and O'Gorman 1991).

Table 44 Variables included in Lake Trout models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	Fetch_Log, DeltaDist, Year, SqrtSMLT	Month
		Poisson	Year, SqrtSMLT, DeltaDist, SqrtALEW	Month
	1978-1989	Gaussian	SqrtSMLT, ProtDist, SqrtALEW, OpenDist	Month
		Poisson	SqrtSMLT, SqrtALEW, ProtDist, OpenDist	Month
	1990-2014	Gaussian	SqrtSMLT, Fetch, DeltaDist, ProtDist	Month
		Poisson	Year, SqrtSMLT, Fetch, DeltaDist	Month
GWR	1978-2014	Gaussian	SqrtSMLT	
	1978-1989	Gaussian	ProtDist, Temp, SqrtSMLT, SqrtSLIM	
	1990-2014	Gaussian	SqrtSMLT	

Table 45 Relationships of reoccurring model variables for Lake Trout

Lake Trout					
SqrtSMLT	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+ / -	+ / - / +	+ / -
1978-1989	+	+	+ / - / +	+ / - / +	+ / -
1990-2014	+	+	+ / -	+ / - / +	+ / -

6.4.5. Yellow Perch model structure

For Yellow Perch the most common variables are $\log_{10}$ transformed distance to open type wetland and square root of Trout Perch abundance (Table 46). The relationship between Yellow Perch and distance to open type wetland was negative for all GLM and GAM models (Table 47). This relationship suggests that Yellow Perch abundances decrease the further from open type wetlands one searches. The square root of Trout Perch abundance was used in all models from the 1978 - 2014 and 1978 - 1989 datasets. The relationship was positive for all GLM and GAM models. It is likely that the positive relationship is related to their shared diet of Diporeia,

amphipods (Wells 1980). Diporeia largely disappeared from Lake Ontario in the early 1990s (Dermott 2001). This disappearance of amphipods may explain why Trout Perch is no longer a significant predictor in the 1990-2014 dataset.

Table 46 Variables included in Yellow Perch models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	SqrtSPOT, OpenDist_Log, SqrtTRPR, DeltaDist_Log, SqrtJOHN	
		Poisson	OpenDist_Log, Depth, RivDist, Fetch_Log, SqrtTRPR	
	1978-1989	Gaussian	SqrtTRPR, SqrtSPOT, SqrtJOHN, OpenDist_Log, ProtDist_Log	
		Poisson	OpenDist_Log, Depth, SqrtTRPR, RivDist, Fetch	
	1990-2014	Gaussian	SqrtSPOT, OpenDist_Log, DeltaDist_Log, Temp, ProtDist_Log	
		Poisson	OpenDist_Log, Depth, RivDist, SqrtSPOT, Fetch_Log	
GWR	1978-2014	Gaussian	SqrtTRPR	
	1978-1989	Gaussian	SqrtTRPR	
	1990-2014	Gaussian	Temp	

Table 47 Relationships of reoccurring model variables for Yellow Perch

Yellow Perch					
OpenDist_Log	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	-	-	- / + / -	-	**NA**
1978-1989	-	-	- / + / -	- / + / -	**NA**
1990-2014	-	-	- / + / -	-	**NA**
SqrtTRPR	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+	+	+ / -
1978-1989	+	+	+	+	+ / -
1990-2014	**NA**	**NA**	**NA**	**NA**	**NA**

6.4.6. Slimy Sculpin model structure

For Slimy Sculpin, the most common variables was depth (Table 48). The relationship between Slimy Sculpin and depth showed a positive relationship for the GLMs and a mixed

positive and negative relationship for the GAMs (Table 49). This relationship suggests that Slimy Sculpin prefer deeper water when looking at the GLMs, however the GAMs suggest that at certain depth abundances may decrease. This decrease could however be due to a reduction of gear efficiency at the deepest depths.

Table 48 Variables included in Slimy Sculpin models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	Year, Depth, OpenDist, SqrtSMLT	Month
		Poisson	Depth, SqrtTRPR, OpenDist, Year	Month
	1978-1989	Gaussian	Depth, OpenDist, RivDist, SqrtTRPR	Month
		Poisson	OpenDist, SqrtTRPR, SqrtSMLT, SqrtALEW	Month
	1990-2014	Gaussian	Year, Depth, SqrtTRPR, SqrtSMLT	Month
		Poisson	Year, SqrtTRPR, Depth, Fetch_Log	Month
GWR	1978-2014	Gaussian	Depth, SqrtSMLT	Month(Oct)
	1978-1989	Gaussian	Depth	Month(Oct)
	1990-2014	Gaussian	Depth	Month(Oct)

Table 49 Relationships of reoccurring model variables for Slimy Sculpin

Slimy Sculpin					
Depth	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+ / -	+ / -	+ / -
1978-1989	+	**NA**	+ / -	**NA**	+ / -
1990-2014	+	+	+ / -	+ / -	+ / -

6.4.7. Rainbow Smelt

For Rainbow Smelt, the most common variables were the event year, the square root of Lake Trout abundance, and the distance to open type wetland (Table 50). The year variable had a negative relationship with the GLMs for the 1978 - 2014 and 1990 - 2014 dataset, but a positive relationship for the 1978 - 1989 dataset (Table 51). The GAM models had a positive relationship that becomes negative for the 1978 - 2014 and 1990 - 2014 datasets and a positive relationship

for the 1978 - 1989 dataset. These relationships show that as time moved forward the abundance of Rainbow Smelt started to decrease. Because the variable is simply time, it is difficult to say the exact reason for the decline but determining when the decline takes place according to the model could assist in figuring out what could be the cause.

Table 50 Variables included in Rainbow Smelt models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	SqrtLTRT, OpenDist, Temp, Year, SqrtTRPR	
		Poisson	Depth, OpenDist, Year, SqrtLTRT, SqrtTRPR	
	1978-1989	Gaussian	OpenDist, Depth, Temp, Year, SqrtLTRT	
		Poisson	Depth, OpenDist, Year, SqrtLTRT	Month
	1990-2014	Gaussian	Year, SqrtLTRT, OpenDist, SqrtTRPR, Temp	
		Poisson	Year, OpenDist, Depth, SqrtLTRT, SqrtTRPR	
GWR	1978-2014	Gaussian	SqrtLTRT, Depth	
	1978-1989	Gaussian	ProtDist, Depth	
	1990-2014	Gaussian	SqrtLTRT, SqrtTRPR	

Table 51 Relationships of reoccurring model variables for Rainbow Smelt

Rainbow Smelt					
Year	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	-	-	+ / -	+ / -	**NA**
1978-1989	+	+	+	+	**NA**
1990-2014	-	-	+ / - / +	+ / - / +	**NA**
SqrtLTRT	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+	+ / - / +	+ / -
1978-1989	+	+	+	+	**NA**
1990-2014	+	+	+ / -	+ / -	+ / -
OpenDist	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+	+	**NA**
1978-1989	+	+	+	+	**NA**
1990-2014	+	+	+	+	**NA**

The Rainbow Smelt abundances were also shown to increase in all GLMs as the abundance of Lake Trout increases. The GAMs also had this positive relationship with Lake

Trout but at a certain point the relationship turns negative and Rainbow Smelt abundances begin to decrease as Lake Trout increase. This is a reasonable relationship since Lake Trout are a known predator of Rainbow Smelt, so a large number of Lake Trout could indicate a large number of Rainbow Smelt. But if too many predators are present the number of Rainbow Smelt will be too diminished by the Lake Trout.

The GLMs and GAMs saw positive relationships with distance to open type wetlands. This suggests that Rainbow Smelt are not commonly found near open type wetlands.

6.4.8. Spottail Shiner model structure

For Spottail Shiner, the most common variables are the square roots of Trout Perch and Yellow Perch abundances (Table 52). The GLMs showed that Trout Perch abundance had a positive relationship with all datasets (Table 53). The GAMs showed that Trout Perch abundance had a positive relationship for the 1978 - 2014 and 1978 - 1989 dataset. The 1990 – 2014 dataset model using the Gaussian GAM method showed that Trout Perch had a positive relationship that turns negative with Spottail Shiner. This could possibly suggest that Spottail Shiner share similar diets with the Trout Perch but at certain abundances the Trout Perch could possibly outcompete the Spottail Shiner for resources. The GLMs also showed a positive relationship with Yellow Perch abundances. The GAMs showed that Yellow Perch abundance was positive at first but then the relationship turns negative. This change was seen in all datasets and suggests that while the two species might have some overlap in diet; Yellow Perch may also be feeding on the Spottail Shiner.

Table 52 Variables included in Spottail Shiner models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	SqrtTRPR, SqrtPRCH, DeltaDist_Log, Depth, RivDist_Log	
		Poisson	SqrtTRPR, DeltaDist_Log, SqrtJOHN, Depth	Month
	1978-1989	Gaussian	SqrtTRPR, SqrtPRCH, Fetch_Log, SqrtJOHN, Temp	
		Poisson	SqrtTRPR, Depth, SqrtJOHN, OpenDist_Log	Month
	1990-2014	Gaussian	SqrtTRPR, SqrtPRCH, DeltaDist_Log, Temp, Depth	
		Poisson	DeltaDist_Log, ProtDist, SqrtPRCH, SqrtTRPR	Month
GWR	1978-2014	Gaussian	SqrtTRPR, SqrtPRCH	
	1978-1989	Gaussian	SqrtTRPR	
	1990-2014	Gaussian	SqrtTRPR, SqrtPRCH	

Table 53 Relationships of reoccurring model variables for Spottail Shiner

Spottail Shiner					
SqrtTRPR	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+	+	+ / -
1978-1989	+	+	+	+	+ / -
1990-2014	+	+	+ / -	+ / - / +	+ / -
SqrtPRCH	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	**NA**	+ / -	**NA**	+ / -
1978-1989	+	**NA**	+ / -	**NA**	**NA**
1990-2014	+	+	+ / -	+ / -	+ / -

6.4.9. Threespine Stickleback model structure

For Threespine Stickleback, the most common variable was the $\log_{10}$ transformation of fetch (Table 54). The majority of models for Threespine Stickleback had a negative relationship for GLMs and GAMs for fetch. Only the GAM with Poisson distribution showed a positive relationship with some values of fetch (Table 55). Due to the poor overall results of the Threespine Stickleback models the true relevance of fetch and abundance is questionable.

Table 54 Variables included in Threespine Stickleback models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	Year, Depth, Fetch_Log, SqrtJOHN, SqrtSMLT	
		Poisson	Depth, Year, Fetch_Log, SqrtJOHN	Month
	1978-1989	Gaussian	Fetch_Log, Temp, SqrtTRPR, SqrtPRCH, SqrtLTRT	
		Poisson	SqrtLTRT, Fetch_Log, SqrtTRPR, SqrtPRCH	Month
	1990-2014	Gaussian	Fetch_Log, SqrtJOHN, Temp, RivDist_Log	Month
		Poisson	Depth, SqrtJOHN, Fetch, SqrtGOBY	Month
GWR	1978-2014	Gaussian	NA	NA
	1978-1989	Gaussian	NA	NA
	1990-2014	Gaussian	NA	NA

Table 55 Relationships of reoccurring model variables for Threespine Stickleback

Threespine Stickleback					
Fetch_Log	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	**-**	**-**	**-**	**-**	**NA**
1978-1989	**-**	**-**	**-**	**+ / - / +**	**NA**
1990-2014	**-**	**NA**	**-**	**NA**	**NA**

6.4.10. Trout Perch model structure

For Trout Perch, the most common variables was the $\log_{10}$ transformation of distance to open type wetland and the square root of Spottail Shiner abundance (Table 56). The relationship between distance to open type wetland and Trout Perch was negative for all GLM and GAM models; this suggested that Trout Perch were more abundant the closer they were to open wetlands (Table 57). The relationship between Trout Perch and Spottail Shiner was positive for all GLM models. The GAM relationships with Spottail Shiner were more diverse. The Gaussian GAM in the 1978 - 2014 dataset had a positive relationship that turned negative relationship at some point. The 1990 - 2014 datasets GAMs both showed a shifting pattern from positive to negative at different points.

Table 56 Variables included in Trout Perch models for each dataset

Method	Dataset Name	Response Variable Distribution	Continuous Predictor Variables	Categorical Predictor Variable
GLM and GAM	1978-2014	Gaussian	OpenDist_Log, Fetch, SqrtSMLT, SqrtPRCH, SqrtSPOT	
		Poisson	OpenDist_Log, SqrtSPOT, Temp, Year	Month
	1978-1989	Gaussian	SqrtSPOT, SqrtPRCH, Fetch_Log, Year, DeltaDist_Log	
		Poisson	SqrtSPOT, OpenDist_Log, Temp, SqrtSLIM	Month
	1990-2014	Gaussian	SqrtSPOT, SqrtSMLT, OpenDist_Log, SqrtPRCH, Temp	
		Poisson	OpenDist_Log, Year, Temp, SqrtSPOT	Month
GWR	1978-2014	Gaussian	SqrtSMLT, SqrtPRCH, Temp	
	1978-1989	Gaussian	SqrtSPOT, SqrtPRCH	
	1990-2014	Gaussian	SqrtSMLT, SqrtSPOT	

Table 57 Relationships of reoccurring model variables for Trout Perch

Trout Perch					
OpenDist_Log	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	-	-	-	-	**NA**
1978-1989	**NA**	-	**NA**	-	**NA**
1990-2014	-	-	-	-	**NA**
SqrtSPOT	**GLMG**	**GLMP**	**GAMG**	**GAMP**	**GWR**
1978-2014	+	+	+ / -	+	**NA**
1978-1989	+	+	+	+	+ / -
1990-2014	+	+	+ / - / +	+ / - / +	+ / -

6.5 Standardized Residuals

The standardized residual versus fitted value plots for the best performing models showed that none of the models had randomly distributed residuals. This was a strong indicator that one or more key components were absent from the models. All QQ plots also did not have a linear pattern also suggesting that key components were absent.

When standardized residuals were mapped, the highest and lowest deviations of the mean were often located in smaller isolated areas. This isolation of standard residuals in small areas

rather than dispersed throughout the study area suggests that those areas have an additional strong influence that is causing the over and under estimations. The majority of the standardized residuals were isolated to the eastern portion of the lake. This could suggest that the islands located in the eastern portion or the St. Lawrence River could be having a major impact as there are no similar comparisons to these features elsewhere in the lake.

6.6 Variations Between Results From Different Time Period Datasets

Of the four datasets, each consisting of different stretches of time, the 1978 - 1989 and 1990 - 2014 datasets were the only ones to achieved an adjusted R^2 greater than 0.70 and had the highest Cohen's Kappa values over the two remaining datasets. While the 1978 - 2014 dataset could have offered more insight in long lasting influences for fish species the adjusted R^2 values were often lower than the 1978 - 1989 and 1990 - 2014 dataset. As well as having the lower Cohen's Kappa values. The 2004 - 2014 dataset that was used to only model for Round Goby was considered the poorest performing dataset. While some methods produced relatively high adjusted R^2 the Cohen's Kappa values showed that they models were often biased towards certain abundance classes. This dataset could have been useful in the modeling of an invasive species but the biased predictions would have made them unreliable. The 1978-1989 and 1990-2014 datasets are likely to have been the best datasets because of how they helped to reflected the state of the lake before and after one of the most influential events to occur within Lake Ontario, the invasion of dreissenid mussels.

6.7 Summary

The comparison of methods showed that GAM and GWR had similar adjusted R^2 and Cohen's Kappa values. The GWR had the lower AIC values among the other Gaussian models. The GWRs also used fewer variables in model development than the GLMs and GAMs.

Assessment of model function and structure showed that there was an overlap in the variables selected for a method and dataset to be used for model development. Variables that appear in a majority of a species' models could indicate the most influential variables that were available. Month was included in a number of species, with positive relationships being associated most likely to spawning practices. Rainbow Smelt, Trout Perch, and Slimy Sculpin abundances were reoccurring variables for a number of species with positive relationships associated with other species that had similar dietary needs as the target species. The negative relationships associated with these abundances could be due to predation on the target species. The GAMs allowed the temperature variable to better represent reality by including a point where a positive relationship with temperature can turn negative when it gets too warm.

Chapter 7 Discussion and Conclusion

While Chapters 5 and 6 provide rich detail regarding the modeling results and comparisons, this chapter provides the overall conclusions achieved in this study. The best modeling method was determined by the method that had higher overall adjusted R^2, Cohen's Kappa values, and simpler model complexity. The review of reoccurring predictor variables showed that they followed reasonable and expected relationships. These model indictors along with standardized residuals versus fitted value plots were used to help determine if the models were better at making predictions. This chapter also addresses the issues encountered and possible improvements for future research.

7.1 General Conclusions

There are several general conclusions that can be made from the comparative results discussed in the previous chapter. These are outlined in the following sections.

7.1.1. GWR is the best modeling method

The result of this study suggests that the best modeling method was GWR. While the GWR did not commonly achieve the highest adjusted R^2 values, it did have the most success at predicting all abundance categories. Both GAM and GWR generally outperformed the GLM in adjusted R^2 and Cohen's Kappa values. Between GAM and GWR methods, the GAM method had the higher adjusted R^2 value, but GWR often had the second highest while using fewer predictor variables. The GAM and GWR methods also had similar Cohen's Kappa values for the moderate and high abundances categories, with GWR also getting better results for the low abundance category. The similarity in the results for adjusted R^2 values and Cohen's Kappa values between GAM and GWR with the added weight of the GWR having a less complex

model with fewer predictor variables suggest that GWR is the best overall modeling method to use.

While AIC was not used as an indicator for choosing the best model because of the uncertainty in use comparing different distribution families, it has identified which model is best within distribution families. For the Gaussian families, the GWR had the lowest AIC values for most of the species modeled, approximately 73 percent. The trend for Gaussian models was that the GLM had the highest AIC values and GWR mostly received the lowest values, indicating that GWR is the better model. When the GWR did not have the lowest AIC value, it was the GAM that did. For the Poisson models only, the GLM and GAM could be compared because no GWR with Poisson distribution was used. The GAM method managed to have the lowest AIC between both methods for every species beside Slimy Sculpin in the 1978 - 2014 dataset.

The AIC values showed that among the Gaussian models the GWR method was the better of the three. The GWR AIC values could not be compared to the Poisson models due to differences in distribution families. If a GWR that allowed Poisson distribution was used it could have a similar trend as the Gaussian models and see further improvement over the GAM method.

The use of the Esri Spatial Analyst extension tool Geographically Weighted Regression also offered an easy and user friendly interface to create predictor coefficient surfaces. The predictor coefficient surfaces can be used to see how the influence of the predictor changes throughout the study area. The added pressure of accounting for local multicollinearity could also help in model creation by choosing the most influential variables.

7.1.2. Local regression is better than global

Since this study concludes that GWR is the best overall modeling method, this also suggests that a local regression performs better than a global regression. The local R^2 values

produced by the GWR also increase model usefulness. By studying the local R^2 values, smaller areas within the study area can be considered for future research based on how well or poorly that area was modeled. The reason that the local regression performed better than the global regression could be due to habitat and biological values changing dynamically, in both time and space, in an area as large as Lake Ontario. These variations in variables could cause global regression to be over generalized.

7.1.3. Good models cannot be produced

No good models, those with adjusted $R^2 \geq 0.7$ and Cohen's Kappa rankings of moderate or better, could be produced with any of the modeling methods. While some models for some species, like Spottail Shiner, were able to achieve the required adjusted R^2 value, no model was able to get moderate agreement or better in all the abundance categories. While moderate agreement rankings were obtained for the moderate and high abundances categories of many species the same could not be said of the presence and absence category or the low abundance category.

This inability to develop good models was most likely due to missing predictor variables. This was suggested when reviewing the standardized residual versus fitted values and QQ - Plots, which strongly indicate that at least one key variable is missing in all models. The use of multiple species and several different temporal range datasets with differing abundances and number of observations has helped determine if a particular method works better with a specific species or dataset. For example, Trout Perch often did best with the use of a Gaussian distribution when most species did better with a Poisson distribution. The inclusion of numerous zero observations in the models could also have added to the difficulty in developing a significant model. All models consistently over-predicted for zeros.

While the models weren't considered to be good with the broad requirements set by this study, the requirements could be relaxed with more information on the nature of the individual species. The models, such as Spottail Shiner, that were close to the requirements could become classified as a good model with the inclusion of an additional predictor variable or the refinement of an existing model. To refine the existing model results that do not meet the stated requirements as good would require further knowledge of the specific species that includes their diet, life cycle, and behavior.

7.2 Issues Encountered

The biggest issue in developing models for Lake Ontario fish species was the lack of dynamic, both temporal and spatial, environmental and biological data to develop better models. While datasets for plankton levels, macroinvertebrate community, and water quality exist, they are often have a broad temporal and spatial scale that does not match the more limited temporal scale of the benthic trawling surveys. These dynamic variables are often sparsely sampled within the study area and values extrapolated to the rest of the lake. Because of this sparse sampling the dispersed sampling points would likely not be within the coverage or receive questionable extrapolated values. The inclusion of these dynamic variables could be what the model needs to develop models with higher adjusted R^2 values and better Cohen's Kappa value agreement rankings.

7.3 Future Research

This study was a successful starting point from which to determine which regression method would be best suited for a study area the size of Lake Ontario. However, in order to create good models, which this study was unable to accomplish, future research could focus solely on improving the GWR models. Improvements on the GWR could be as simple as

reducing the study area to areas that received a decent local R^2 or as much as adding new predictor variables.

One future approach would determine the best spatial scale at which to aggregate the observational data. Determining the proper spatial scales for modeling could possibly be determined by deeply investigating how the trawling vessel performs the survey. This would require detailed tracking of trawling vessels over a large number of sampling events.

Another obstacle to overcome in future research is developing a way that temporal variables--including temperature, season, and other factors that change with time--are not lost or over generalized to spatial aggregation. This study focused on individual observational events so temporal factors, such as catch numbers and fishing depth temperature, would not be averaged together. By overcoming this obstacle larger time period datasets could be better utilized without losing seasonal effects from the model development.

Another improvement that could be implemented in future research is the inclusion of better biological environmental data such as macroinvertebrate density, plankton density, light levels, concentration of oxygen, pH, and a number of other characteristics. This data does exist but does not match the temporal scale of the data. Macroinvertebrate and plankton density surveys are often done annually but not with the dispersal and frequency that the fisheries data is collected. While this biological and environmental data is currently available, it would likely have to be an annual average and values extrapolated to cover the study area so that all observations could have values. Future research would be better focused on collecting these variables at the time and location of the trawling survey. While this would not be practical for every variable, the inclusion of recording instruments onto the trawling gear could collect a good number of them.

Yet another improvement would be to investigate the effectiveness of each individual trawling event. Recently, cameras have been attached to trawls to determine their effectiveness; these videos can be used to create a variable that would act as a weight for event efficiency that would prevent poor preformed trawling events due to gear error from impacting results from more successful trawls. However, the use of these cameras would limit the models to more recent years.

7.4 Conclusion

The results of this study showed that GWR was the best modeling method compared to GLM and GAM. While GAM did outperform with some of the model indicators than the GWR models, the GWR did similarly while reducing the complexity of the models. This study has contributed more evidence that a local regression method like GWR is an improvement in model development. Because of the relatively new use of GWR in fisheries research, more studies like this conducted with GWR can inform other fisheries studies on the issues and possible solutions to those issues.

References

Barry, Simon C., and Alan H. Welsh. 2002. "Generalized additive modeling and zero inflated count data." *Ecological Modeling* 157(2):179-188.

Botts, R., B. Krushelnicki. 1987. "Introduction: The Great Lakes" In *The Great Lakes: An Environmental Atlas and Resource Book*, edited by Kent Fuller, Harvey Shear, and Jennifer Wittig, 1-8. Government of Canada and US Environmental Protection Agency.

Burnham, Kenneth P., and David R. Anderson. 2004. "Multimodel inference understanding AIC and BIC in model selection." *Sociological methods & research* 33(2): 261-304.

Byrne, Rowan, John Fish, Thomas K. Doyle, and Jonathan Houghton. 2009. "Tracking leatherback turtles (*Dermochelys coriacea*) during consecutive inter-nesting intervals: Further support for direct transmitter attachment." *Journal of Experimental Marine Biology and Ecology* 377(2): 68-75.

Cook, David G., and Murray G. Johnson. 1974. "Benthic macroinvertebrates of the St. Lawrence Great Lakes." *Journal of the Fisheries Board of Canada* 31(5): 763-782.

Craney, Trevor A., and James G. Surles. 2002. "Model-dependent variance inflation factor cutoff values." *Quality Engineering* 14(3): 391-403

Dermott Ronald. 2001. "Sudden disappearance of the amphipod Diporeia from eastern Lake Ontario, 1993-1995." *Journal of Great Lakes Research* 27(4): 423-433.

Dormann, Carsten F. 2007. "Assessing the validity of autologistic regression." *Ecological Modeling* 207(2): 234-242.

Durbin, Ann Gall, Scott W. Nixon, and Candace A. Oviatt. 1979. "Effects of the spawning migration of the alewife, Alosa pseudoharengus, on freshwater ecosystems." *Ecology* 60(1): 8-17

Elrod, Joseph H., and Robert O'Gorman. 1991. "Diet of juvenile lake trout in southern Lake Ontario in relation to abundance and size of prey fishes, 1979-1987." *Transactions of the American Fisheries Society* 120(3): 290-302.

Fukuba, Tatsuhiro, Tetsuya Miwa, Shun Watanabe, Noritaka Mochioka, Yoshiaki Yamada, Michael J. Miller, Makoto Okazaki et al. 2014. "A new drifting underwater camera system for observing spawning Japanese eels in the epipelagic zone along the West Mariana Ridge." *Fisheries Science* 81(2): 235-246.

Graham, Catherine H., Simon Ferrier, Falk Huettman, Craig Moritz, A. Townsend Peterson. 2004. "New developments in museum-based informatics and applications in biodiversity analysis." *Trends in Ecology and Evolution* 19: 497-503.

Great Lakes Commission (GLC). 2004. Great Lakes Coastal Wetland Inventory. Available from: the Great Lakes Commission Wetland Consortium. http://glc.org/files/projects/cwc/CWC-GLWetlandsInventory-ClassificationScheme.pdf

Griffiths, Ronald W., Donald W. Schloesser, Joseph H. Leach, and William P. Kovalak. 1991. "Distribution and dispersal of the zebra mussel (Dreissena polymorpha) in the Great Lakes region." *Canadian Journal of Fisheries and Aquatic* Sciences 48(8): 1381-1388.

Guisan, Antoine, Reid Tingley, John B. Baumgartner, Ilona Naujokaitis-Lewis, Patricia R. Sutcliffe, Ayesha IT Tulloch, Tracey J. Regan et al. 2013. "Predicting species distribution for conservation decisions." *Ecology Letters* 16(12):1424-1435.

Heikkinen, Risto K., Miska Luoto, Miguel B. Araujo, Raimo Virkkala, Wilfried Thuiller, and Martin T. Sykes. 2006. "Methods and uncertainties in bioclimatic envelope modeling under climate change." *Progress in Physical Geography* 30(6): 751-777.

Hernandez, Pilar A., Catherine H. Graham, Lawrence L. Master, and Deborah L. Albert. 2006. "The effect of sample size and species characteristics on performance of different species distribution modeling methods." *Ecography* 29(5): 773-785.

Kilgo, Jamie M. 2012. "Spatial patterns and habitat associations of targeted reef fish in and around a marine protected area in St. Croix, US Virgin Islands." MS Thesis, University of Washington.

ICES. 2004. "Report of the workshop on survey design and data analysis (WKSAD)." Fisheries Techonlogy Committee ICES CM 2004/B:07, Ref. D, G. 65p. www.ices.dk/reports/ftc/2004/wksad04.pdf

Landis, J.R., and Gary G. Koch. 1977. "The measurement of observer agreement for categorical data." *Biometrics* 33(1): 159-174

Lantry, Brian F., and Donald J. Stewart. 1993. "Ecological energetics of rainbow smelt in the Laurentian Great Lakes: an interlake comparison." *Transactions of the American Fisheries Society* 122(5): 951-976.

Leathwick, J.R., and M.P. Austin. 2001. "Competitive interactions between tree species in New Zealand's old-growth indigenous forests." *Ecology* 82(9): 2560-2573.

McKenna, James E. and Chris Castiglione. 2014. "Model Distribution of Silver Chub (Macrhybopsis storeriana) in Western Lake Erie." *The American Midland Naturalist* 171(2): 301-310.

McKenna, James E. and Chris Castiglione. 2010. "Hierarchical multi-scale classification of nearshore aquatic habitats of the Great Lakes: Western Lake Erie." *Journal of Great Lakes Research* 36:757-771.

Mills, Edward L., Ron M. Dermott, Edward F. Roseman, Donna Dustin, Eric Mellina, David Bruce Conn, and Adrian P. Spidle. 1993. "Colonization, Ecology, and Population Structures of the Quagga Mussel (Bivalvia Dreissenidae) in the Lower Great Lakes." *Canadian Journal of Fisheries and Aquatic Science* 50(11): 2305-2314.

Murphy, Christina A., Gael Grenouillet, and Emili Garcia-Berthou. 2015. "Natural abiotic factors more than anthropogenic perturbation shape the invasion of Eastern Mosquitofish (*Gambusia holbrooki*)." *Freshwater Science* 34(3): 965-974.

Myers, Donna N., James McKenna, Dora Passino, and Jana S. Stewart. 2002 "Great Lakes aquatic GAP project." Gap Analysis Bulletin 11(2): 59-64.

National Geophysical Data Center. 1999. "Bathymetry of Lake Ontario." National Geophysical Data Center, NOAA. doi: 10.7289/V56H4FBH [January 18th, 2016].

Nishida, Tom, and Ding-Geng Chen. 2004. "Incorporating spatial autocorrelation into the general linear model with an application to the yellowfin tuna (*Thunnus albacares*) longline CPUE data." *Fisheries Research* 70(2): 265-274.

O'Hara, Robert B., and D. Johan Kotze. 2010. "Do not log-transform count data." *Methods in Ecology and Evolution* 1(2): 118-122.

Owens, Randall W., and George E. Noguchi. 1998. "Intra-lake variation in maturity, fecundity, and spawning of slimy sculpin (cottus cognatus) in southern Lake Ontario." *Journal of Great Lakes Research* 24(2): 383-391.

Owens, Randall W., and Dawn E. Dittman. 2003. "Shifts in the diets of slimy sculpin (Cottus cognatus) and lake whitefish (Coregonus clupeaformis) in Lake Ontario following the collapse of the burrowing amphipod Diporeia." *Aquatic Ecosystem Health & Management* 6(3): 311-323.

Pearson, R.G., T.P. Dawson, P.M. Berry, P.A. Harrison. 2002. "SPECIES: A Spatial Evaluation of Climate Impact on the Envelope of Speceis." *Ecological Modeling* 154(3): 289-300.

Pierce, Graham J., Jianjun Wang, Xiaohong Zheng, Jose M. Bellido, Peter R. Boyle, Vencent Denis, and Jean-Paul Robin. "2001. "A cephalopod fishery GIS for the Northeast Atlantic: development and application." *International Journal of Geographical Information Science* 15(8): 763-784.

Schlieper, C. 1972. *Research methods in marine biology*. University of Washington Press, Seattle.

Sims, David W., Nuno Queiroz, Nicolas E. Humphries, Fernando P. Lima, and Graeme C. Hays. 2009. "Long-term GPS tracking of ocean sunfish Mola mola offers a new direction in fish monitoring." *PLoS ONE* 4(10): http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007351.

Sippel, Tim, J. Paige Eveson, Benjamin Galuardi, Chi Lam, Simon Hoyle, Mark Maunder, Pierre Kleiber et al. 2015. "Using movement data from electronic tags in fisheries stock assessment: a review of models, technology and experimental design." *Fisheries Research* 163: 152-160

Sly, Peter G. 1991. "The effects of land use and cultural development on the Lake Ontario ecosystem since 1750." *Hydrobiologia* 213(1): 1-75.

Steenbergen, Marco R. 2012. "Maximum Likelihood Programming in R" *Academia* Accessed March 11, 2016 http://www.academia.edu/3466463/Maximum_Likelihood_Programming_in_R

Stoermer, Eugene F., J.A. Wolin, Claire L. Schelske, and D.J. Conley. 1985 "An Assessment of Ecological Changes During the Recent History of Lake Ontario Based on Siliceous Algal Microfossils Preserved in the Sediments." *Journal of Phycology* 21(2): 257-276.

Stumpf, Richard P., Kristine Holderied, and Mark Sinclair. 2003. "Determination of water depth with high resolution satellite imagery over variable bottom types." *Limnology and Oceanography* 48(1): 547-556.

Sullivan, T.J. 2015. *Air pollutant deposition and its effects on natural resources in New York State.* Ithaca, New York: Cornell University Press

U.S. Army Corps of Engineers (USACE). 1984. *Shore Protection Manual.* Coastal Engineering Research Center. Fort Belvoir, Virginia.

USGS. 2012. "Status of Important Prey Fishes in the U.S. Waters of Lake Ontario, 2011." Report No. 52092479 Department of the Interior, U.S. Geological Survey, Biological Resources Division, Great Lakes Science Center, Lake Ontario Biological Station. glsc.usgs.gov/products/reports/52092479

Usseglio, Paolo. 2015. "Quantify reef fishes: bias in observational approaches." In *Ecology of Fishes on Coral Reefs*, edited by Camilo Mora, 270-273. Cambridge: Cambridge University Press.

Watkins, James M., Ronald Dermott, Stephen J. Lozano, Edward L. Mills, Lars G. Rudstam, and Jill V. Scharold. 2007. "Evidence for remote effects of dreissenid mussels on the amphipod Diporeia: analysis of Lake Ontario benthic surveys, 1972-2003". *Journal of Great Lakes Research* 33(3): 642-657.

Wells, LaRue. 1980. *Food of Alewives, Yellow Perch, Spottail Shiners, Trout-Perch, and Slimy Sculpin and Fourhorn Sculpins in Southeastern Lake Michigan.* United States Fish and Wildlife Service Technical Report 98, Ann Arbor, MI.

Windle, Matthew J.S., George A. Rose, Rodolphe Devillers, and Marie-Josee Fortin. "Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic." *ICES Journal of Marine Science* 67: 145-154.

Zuur, Alain, Elena N. Leno, and Graham M Smith. 2007. *Analysing Ecological Data.* New York: Springer Science & Business Media.