Why do lake whitefish move long distances in Lake Huron? Bayesian variable selection of factors explaining fish movement distance

Yang Li [a*], James R. Bence [a], Zhen Zhang [b], and Mark P. Ebener [c]

[a] Quantitative Fisheries Center, Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI 48824, USA
[b] Enabling Capabilities Technology Center, Dow AgroSciences, Indianapolis, IN 46268, USA
[c] Inter-Tribal Fisheries and Assessment Program, Chippewa Ottawa Resource Authority, Sault Ste. Marie, MI 49783, USA

Abstract

Understanding fish movement patterns is vital for stock assessment and fishery management. We used a variable selection procedure in a Bayesian framework to understand what factors most likely affect the net movement distance of individual fish based on a conventional tag-recovery study of lake whitefish populations in Lake Huron during 2003-2011, where fish of this species with spawning site fidelity were tagged during the spawning season and recovered throughout the year. We found that fish with greater total length, and those that were tagged and released from tagging sites near Cheboygan and Alpena, Michigan, moved longer net distances than fish from other tagging sites. Habitat conditions also had a profound effect on net movement distance. We found that shorter movement distances by lake whitefish can be expected if the relative density of the benthic amphipod *Diporeia* spp. was higher near the tagging site during the recovery year. We also found evidence that lake whitefish may start their annual spawning migration runs earlier during warmer years. More generally, our Bayesian framework for analysis of conventional tagging data has potential for wide applicability, and model details and our code are provided to facilitate this.

[*]Corresponding author: E-mail address: liyang11@msu.edu (Y. Li); Tel +1 (517) 355-0126.

Key words: Fish Movement, Variable Selection, Tag Recovery, Lake Whitefish, Bayesian.

1. Introduction

Many fish species move for long distances at various times during their life cycle, and movements made by individuals vary from regular and predictable migration to less-predictable resource driven nomadism (Runge et al., 2014). Most previous research that evaluated changes in fish spatial locations focused on either the triggering factors or distance between initial and final fish location (e.g., Albanese et al., 2004; Radinger and Wolter, 2014), or on estimating net movement/migration rates of populations (Polacheck et al., 2006; Vandergoot and Brenden, 2014).

Fish movement is essential from both conservation and management perspectives. Movement behavior can influence how fish are distributed, whether their populations persist in the face of ecosystem changes, and how stocks are assessed. Fish movement can further influence ecological interactions and evolution (Lidicker and Stenseth, 1992). Management problems such as inaccurate assessment results, or inappropriate catch limits, can occur when actual fish movements do not agree with the spatial assumptions made in stock assessments and management decisions, which can result in local population depletion and population collapse (Fu and Fanning, 2004; Hutchings, 1996; Li et al., 2015; Mitchell and Beauchamp, 1988; Rothschild, 2007).

Despite its ecological and management importance, understanding of fish movement patterns in time and space, and how movements are related to environmental variables, is still limited. Moreover, most previous research that focused on the triggering factors (i.e., factors causing the initiation of movement) and net fish movement distance were limited to stream fish, given the easy calculation of net distance moved from conventional tagging data. Much less is known about movement of fish that live in large water areas. Most of which is known has been derived from electronic tagging data, although there are many long-term conventional tagging programs. While technological advances make the use of acoustic or pop-up tags increasingly useful, conventional tags are still more widely used for estimating population size, mortality, and tracking individual growth, given their lower price. Conventional tagging data can also provide information on the location at tag release and tag recovery, which could be used for the estimation of movement route and intensity (e.g., net fish movement distance) (e.g., Albanese et al., 2004; Gilliam and Fraser, 2001).

The goal of this study was to develop a model framework for analysis of how factors impact the distance fish move from when they are tagged until they are recovered ('net fish movement distance' hereafter) in a larger water body, based on conventional tag-recovery results. We based our research on several lake whitefish (*Coregonus clupeaformis)* spawning stocks in Lake Huron of the Laurentian Great Lakes of North America. As an ecological and economically important fish species in the Great Lakes, lake whitefish have been found to move freely among multiple management units during the non-spawning period, but show a high degree of natal homing, so nearly all mature fish return to spawn at the same location each year (Ebener et al., 2010b). Previous research on lake whitefish movement patterns provides a useful platform for us to derive a priori hypotheses about the potential factors that influence movement. Since the establishment of dreissenid mussels in the early 1990s, the ecosystem of four of the five Great Lakes have changed substantially, including an overall decrease in the density of lake whitefish's preferred food- *Diporeia* spp. (Barbiero et al., 2011; McNickle et al., 2006; Mohr and Nalepa, 2005). In this context, Rennie et al. (2012) evaluated the

relationship between lake whitefish migration distance and growth rate, and found that the least mobile population of lake whitefish was supported by a remnant *Diporeia* spp. population. Ebener et al. (2010b) found that stock identity and season of recapture affected net movement distance most strongly, while the influence of variables such as sex, year, fish total length, and time at large was weaker. Although the role of temperature has not been directly implicated in explaining patterns in the fish movement, the association between lake whitefish harvest and surface water temperature suggested that such a connection may exist (Price et al., 2003).

The pioneering studies of net movement distance used either a regression-tree based approach or ANOVA models to test whether net movement distance varied significantly in association with the factors they evaluated (e.g., Albanese et al., 2004; Ebener et al., 2010b; Radinger and Wolter, 2014). Because some studies estimated the effects of different factors as additive (i.e., causing a given distance change rather than a percentage change in net movement distances), it is hard to generalize the results from studies with different spatial and temporal scales. When jointly considering multiple factors and continuous covariates, the ANOVA approach can provide only a rough picture of the continuous relationship between net movement distance and explanatory factors. Thus, a more thorough regression analysis is needed. The regression-tree based approach seeks to approximate nonlinearity and interactions in the relationships between the net movement distances and multiple factors by recursively partitioning the data points according to the categorization of the factors (Ebener et al., 2010b). Such partitioning may have difficulty in interpreting the effects, if the observations from the same tag or recovery area happen to be separated into different branches of the tree. Some regression-tree applications have partitioned data by site (i.e., different sites on different branches), and this can make it difficult to develop a general understanding of movement (Ebener et al., 2010b). In addition, although it is possible for regression-tree based approaches to rank or select variables based on variable importance measures, they do not provide any further insight of the uncertainty associated with their rankings or selections. Also information criteria, such as Akaike's information criterion and the Bayesian information criterion, commonly used as penalization terms for the number of parameters in model, are not applicable for nonparametric tree-based models (Claeskens and Hjort, 2008).

We therefore considered a global linear regression model that accounts for joint effects of multiple factors and the heterogeneity among sites, to study the relationship between the net movement distance and individual factors. We further conducted a variable selection procedure under a Bayesian framework to explore the plausibility of alternative regression models that include various explanatory variables, and assess the associated uncertainty. Bayesian variable selection treats the regression model itself as random among all possible models with different sets of variables. Thus, it accounts for model uncertainty in the overall assessment of uncertainty by making inferences on how probable alternative models are after consideration of the data. The implementation of Bayesian variable selection via the reversible jump Markov chain Monte Carlo (rjMCMC) (Green, 1995) procedure is substantially more efficient in exploring the model space than the traditional approaches such as all-subsets-regression (Woznicki et al., 2016). While we believe our approach has substantial advantages over regression-tree approaches, it could miss some nonlinear effects that could be identified by regression-

trees. Thus, as a check on robustness we compared our results with those from regression-tree methods.

We considered how net distance moved from tagging to recapture locations changed monthly and over years, and how this net movement pattern depended upon tagging location. In addition, we considered how life history traits, namely total length, and sex, and habitat features, namely *Diporeia* spp. density and water temperature, played a role in these net movement patterns. Thus, the variables we considered as potential explanatory factors in this study were tagging year, recovery year, recovery month, year(s) between tag and recovery, fish total length, sex, tagging (spawning) site, and the habitat variables based on *Diporeia* spp. density and growing degree days.

Our goal was to provide not only insight on how those factors influenced lake whitefish movement in Lake Huron, but also a model framework for analyzing movement mechanism based on conventional tagging data. Although Bayesian variable selection in linear regression is a long-established approach (Mitchell and Beauchamp, 1988), it was rarely used in ecology or more specifically for uncovering explanations for movements (Drouineau et al., 2017; Ethier et al., 2017). Drouineau et al. (2017) used a Bayesian state-space model to analyze the effects of different environmental factors in triggering migration of silver eel in fragmented rivers. Ethier et al. (2017) used Bayesian models and variable selection to evaluate how environmental variables influenced regional variation in population trends of Bobolink. Both studies used a mixture distribution of priors (i.e., normal plus zero-inflation), which were estimated using a Gibbs sampler. However, their variable selection procedure did not introduce a penalty such as BIC for increasing number of selected variables. Also the Gibbs sampler usually involves scanning all variables at each iteration, which could be computational expensive, especially when the number of candidate variables is large.

To the best of our knowledge, this study is the first to apply the Bayesian variable selection approach to compare the effects of various factors on fish net movement distance by introducing an explicit prior penalty on model complexity, and the most comprehensive to date in terms of the range of factors affecting whitefish movement. To avoid sampling all indicators within a Gibbs sampler circle as in Drouineau et al. (2017) and Ethier et al. (2017), we adopt the reversible jump MCMC algorithm for model exploration that mimics stepwise selection and subsets regression technique, which is more computationally efficient. Thus our research introduces an approach to fish movement studies, which has the potential to be much more effectively interrogate a large number of predictor variables. To facilitate usage of our approach, we provide the open-source code for MATLAB program which is online available at to implement the method.

2. Methods

2.1 Data collection, selection, and calculation of net-movement distance

Lake whitefish were tagged and released in a study coordinated by one of us (Mark P. Ebener) at 21 individual tagging sites from nine spawning stocks in Lake Huron from late October through December (i.e., spawning season) of 2003-2006. Total length (mm) of all 35,285 tagged fish were measured before release, spatial coordinates of the tagging and release location, and date of release were recorded for each fish. Lake whitefish were tagged on or very near the spawning grounds and subsequently killed when recovered by the commercial or recreational fishery. The commercial fishing season for

lake whitefish is not closed in Ontario waters during the spawning season, but it is closed in Michigan waters. Thus, fish tagged and released at Detour, Cheboygan, Alpena, and Saginaw Bay (Fig. 1) were extant 1-4 weeks before being subjected to fishing and tag recovery. At Burnt Island, the Fishing Islands, and Sarnia fish were also tagged during the spawning season, but commercial fishing was occurring simultaneously during tagging so they had little time to be extant prior to tag recovery. Recovery happened from December 2003 until December 2012, with the majority being recovered by commercial fishermen, and the rest recovered during fishery surveys. Subsets of the data used here were previously reported by Ebener et al. (2010a, 2010b), and details of the tagging methodology are given by Ebener et al. (2010a).

Our analysis focused on drivers of net movement distance of lake whitefish tagged and recovered in Lake Huron. We thus restricted attention to recoveries for which net distance movement could be calculated and for which explanatory variable data were available. Only recoveries that had location information recorded (either by latitude and longitude or by 10-minute by 10-minute statistical grid, treated as though recovered at the grid center) were considered. In addition, we excluded observations from fish that were recovered within two days of release, as well as those without their recapture date, sex, or total length recorded (i.e., explanatory variables). We also removed fish that were recovered from Lake Michigan because of our focus on movement within Lake Huron and because our explanatory variables were from Lake Huron. We further excluded recoveries from two tagging sites that each produced only two total recoveries, and the two fish recovered during 2012. Thus of the total of 2,098 reported lake whitefish recoveries, 1,368 recoveries were used in this study. Details of data exclusion are described in Supplementary Table S1. These recovered fish had total lengths between 375-667 mm at the time of tagging, and were tagged and released from seven spawning sites (Fig. 1).

We used log-transformed net movement distance as a response variable because net movement distances were highly skewed. We calculated net movement distance based on the shortest water distance between tagging and recovery locations, using a Dijkstra type shortest path algorithm (Vincenty, 1975; online Appendix A). We standardized log-transformed net movement distance by subtracting the mean and dividing by standard deviation prior to analysis.

2.2 Explanatory variables

We hypothesized that net movement distance for lake whitefish in Lake Huron would be influenced by 1) life history traits, which included total length, and sex; 2) temporal factors, which included tagging year (tag_Y), recovery year (rec_Y), recovery month (rec_M), and year(s) between tagging and recovery (year_lag); and 3) habitat condition, which included *Diporeia* spp. density, and growing degree days; and 4) tagging (spawning) sites. These hypotheses, related variables, and the expected sign of the associated coefficients, if hypotheses were supported, are in Table 1. Due to the strong spawning site fidelity of lake whitefish (i.e., nearly all lake whitefish move back to where they born each year during the spawning season), we only considered the habitat conditions during the recovery year as a predictor. That is, the net movement is in actuality the net movement since the prior spawning season. We used relative *Diporeia* spp. density, which was the *Diporeia* spp. density of the release location divided by the mean of all sampled stations in Lake Huron for that year. The U.S. EPA Great Lakes

National Program Office collected *Diporeia* samples every August since 1999 at 12 Lake Huron stations (Barbiero et al., 2011). The release location density was defined as the density at the sampled location closest to the release location. Our hypothesis was that lake whitefish tended to stay near their tagging locations when *Diporeia* density was higher in that vicinity.

We proposed two alternative hypotheses for the relationship between growing degree days (GDDs) (i.e., also known as thermal time, a weather-based indicator about heat acumination for assessing fish growth; e.g., Chezik et al., 2014) and lake whitefish net movement distance, and these led to two distinct sets of GDD variables. These two sets were used in two alternative analyses. We calculated GDDs based on mean daily (daytime) surface temperatures from the Great Lakes Surface Environmental Analysis (GLSEA) remote sensing surface water temperature data (See online Appendix A).

Case 1 (*GDD hypothesis 1*)—Lake whitefish respond to growing conditions they had experienced during the current year. Thus, they would tend to be closer to their tagging (spawning) site when the growing degree days (GDD) at the tagging location was greater than the lake average GDD during that same time period. This led us to define the explanatory variable relative GDD difference ("GDD_Diff"), calculated as: $\text{GDD\_Diff} = (\text{GDD}_{\text{tag}} - \text{GDD}_{\text{lake}})/\text{GDD}_{\text{lake}}$, where GDD_{tag} and GDD_{lake} are the cumulated non negative degree days ($°\text{C}.\,\text{days}$) that exceeded $5°\text{C}$ (Rennie et al., 2009) at the tagging location or for the lake-average, respectively, from the first day of the recovery year to the day of recovery.

Case 2 (GDD hypothesis 2)—The spawning season of lake whitefish would be shifted earlier in the year, in years for which GDDs accumulated faster, because individual fish would reach a physiological status allowing spawning earlier under such conditions. Preliminary model fits without a GDD effect indicated that lake whitefish were generally closer to the spawning location during September through December, than at other times of the year. We therefore assumed that GDD might potentially influence net movement distance (to varying degrees) only during these months. Thus, we added four additional interaction variables (recovery month $\times$ GDD_{lake}) for September through December recoveries. We used GDD_{lake} because fish would be living and feeding away from their spawning/tagging sites until moving to those sites for spawning.

After creating dummy variables and choosing the category with the largest number of observations as the baseline category for each factor, we have a total of 34 (for GDD hypothesis 1 case) and 37 (for GDD hypothesis 2 case) candidate variables including the intercept (Table 1). Note that there was no dummy variable created for the baseline category (i.e., tagging site: Detour, recovery month: June, tagging year: 2004, recovery year: 2006, or sex: Male), because it was defined as zero for all other categories for that factor. All explanatory variables were standardized like net movement distance.

2.3 Model framework

We used Bayesian variable selection to identify the highly probable subsets of predictors for the linear regression and, given a set of predictors, we assessed likely parameter values. Given the Bayesian approach we used, inferences were based on a posterior distribution, which depends jointly on assumed prior distributions and the likelihood of the data. Model components (i.e., regression model, prior distributions, and likelihood) are described in Section 2.31 (Model Description) and how we used Markov

chain Monte Carlo (MCMC) techniques to derive posterior distributions in Section 2.3.2. A separate model selection process was conducted for the two cases (GDD hypotheses).

2.3.1 Model description

Each possible model is of the form:

$$Y = X_{\mathcal{J}}\beta_{\mathcal{J}} + \epsilon, \epsilon \sim N(0, \tau^2 I_N) \qquad \text{(Eq.1)}$$

where Y is the response variable (i.e., log-transformed net movement distance) with N observations, X_j is the $N \times q_j$ design matrix (containing data for the predictors included for that regression), $\beta_{\mathcal{J}}$ is a vector of parameter coefficients (an intercept included in every model plus q_j-1 additional coefficients for the predictor variables included for that model) and ϵ is the residual error. We assumed here homogenous, normal and independent residual errors, with variance τ^2. We assumed independent errors given the relatively large distances between tagging sites (Fig. 1) and because tagging and recovery spatial factors were included as potential explanatory variables. As described in online Appendix B, I_N could be replaced with a selected correlation matrix. A model with a specific subset of selected variables is represented by $\mathcal{J}$, which formally is an index set, that maps the q variables in the selected model to the larger set of p possible variables.

The $\beta_{\mathcal{J}} = (\beta_1, \beta_2, \ldots, \beta_q)'$ had a normal prior $\beta_{\mathcal{J}} \sim N(0, \tau^2\Lambda)$, where $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_q\}$. The λ_js were modeled as arising from a higher level inverse-gamma prior distribution ('hyperprior') with shape parameter a_λ and scale parameter b_λ. We assumed a hyperprior with inverse-gamma density for τ^2 with shape a_τ and scale b_τ. The hyperparameters were set to the values $a_\lambda = a_\tau = 2$ and $b_\lambda = b_\tau = 0.001$, which correspond to a rather dispersed prior distribution.

The normal prior with diagonal variance-covariance matrix for the λ_js represents a decision to use a Bayesian counterpart to Ridge regression. The λ_js represent the signal to noise ratio of the effects in the model, and their magnitude played a role in whether an effect was included and the size of selected models. Modeling them as arising from a hyperprior (rather than specifying their values) allowed for adaptive learning on which variables to include during the model search process.

We included an intercept in all models to account for the grand mean level of Y, as is often done for variable selection. There are a total of 2^{p-1} possible models (i.e., an intercept-only model, all possible models with one additional variable, all possible models with two additional variables, etc.). We specified the prior probability of each model as arising from the product of a prior probability for a model of a given size (i.e., $\pi(q)$), multiplied by the probability of a specific model given its size:

$$\pi(\mathcal{J}) = \pi(\mathcal{J}_q, q) = \pi(\mathcal{J}_q|q)\pi(q) \qquad \text{(Eq.2)}$$

We let $\pi(q) \propto \exp\{-\kappa q\}$ for integer q from $\{1, 2, \ldots, p\}$. Here $\propto$ means "proportional to" up to a constant that is irrelevant in making inferences about the hyper-parameter κ. This placed higher prior probability on models with smaller size, as is consistent with common practice in variable selection, and the rate at which the prior probability falls as model size increases was determined by κ. We set $\kappa = \log(N)/2$, which is analogous to a BIC-type penalty on the number of selected variables (Schwarz,

1978). Conditional on q, each model $\mathcal{J}_q$ had an equal chance of being selected, i.e., $\pi(\mathcal{J}_q|q) = 1/\binom{p-1}{q-1}$ for $q > 1$, and for $q = 1$ no selection is needed.

2.3.2 Characterization of the posterior Distribution using MCMC

We used Markov chain Monte Carlo methods to determine the posterior distribution. We used a hybrid reversible jump technique (rjMCMC), because it performs well when selecting among different sets of variables, which involved trans-dimensional states of Markov chain (Green, 1995; Woznicki et al., 2016). Our procedure involved running multiple chains and combining converged portions of these into one set of "retained samples." The retained samples were summarized to highlight desired properties of the posterior distributions.

Details on the implementation of the hybrid of rjMCMC for model search and Gibbs sampler for parameters given the model, as well as procedures for evaluating MCMC convergence and producing the retained samples are given in online Appendix B. We summarized the posterior distributions for regression model parameters in two ways:

Variable-wise summary— This provided a summary conditional on the j-th variable being selected. This was based on summarizing all samples included in the final MCMC chains for a model that included the j^{th} variable. For the corresponding β_j, the posterior mean and 95% (equal probability tail) credible intervals were constructed from these samples. As a measure of the importance of each variable we also calculated the marginal inclusion probability (Barbieri and Berger, 2004), as the proportion of all retained MCMC samples that included the j^{th} variable in the model.

Model-wise summary— This was conditional on one specific model $\mathcal{J}$ in the posterior samples, and thus was based only on retained MCMC samples for that model. We provide such summaries for the 12 "top" models. Here models are ranked based on the posterior probability, calculated as the proportion of all retained MCMC samples that were model $\mathcal{J}$. For the top models, we summarized the posterior distributions of the β_js for all variables in $\mathcal{J}$, again in terms of the posterior mean and 95% credible interval.

2.3.3 Model diagnosis, simulation study, and comparison with tree-based methods

We used the posterior predictive assessment of model fitness using the χ^2-discrepancy (Gelman et al., 1996), based upon which we calculated the Bayesian p-value for the top models in both GDD hypothesis cases. We also conducted simulations to evaluate how well our Bayesian variable selection procedure can discover the true set of important variables and estimate the corresponding effects, under five different scenarios with varying combinations of true predictor variable effects. We also applied two tree-based methods to our data, and compared the top variables from tree-based methods, the gradient boosting regression tree method (Ethier et al., 2017) and the random forests approach (Breiman, 2001), to the selected variables from our variable-wise summary. Detailed methods for our diagnostic procedures, simulations, and tree-based applications are given in online appendices C, D, and E respectively, and performance statistics resulting from the simulations and tree-based methods are also presented in the appendices.

3. Results

The posterior distributions of the number of selected variables were similar for the two GDD cases and suggested that the most probable model sizes had 6 and 7 variables including an intercept (Fig. 2). However, the selected variables were quite different (see Section 3.1).

3.1 Variable-wise summary

GDD hypothesis 1 Case: There were 10 variables with 95% credible intervals that did not cover 0, which we define as "consistent effects" (Fig. 3). Variables that had consistent effects generally had high marginal inclusion probability, and more generally variables with higher probability of inclusion tend to have more of their posterior distribution on one side of zero (Fig. 3). The six top variables (length, tagging site: Cheboygan and Alpena, *Diporeia*, and recovery months October and November) had marginal inclusion probability above 0.75 (i.e., they are selected by more than 75% of the total posterior samples). The variable recovery month September also had a relatively large marginal inclusion probability (0.40). The other variables that were detected as consistent effects had substantially lower marginal inclusion probability (<0.07) are: years lag, tagging site Fishing islands, and recovery month December. According to the posterior mean of those 10 variables with consistent effects, fish with greater length, longer lag between the tagging and recovery years, released at tagging site Cheboygan, Alpena, and Fishing Islands, and recovered in December had greater net movement distance, while fish released at the tagging site with higher density of *Diporeia*, and recovered during September, October, and November had shorter net movement distance. Our first GDD hypothesis was not supported by the variable selection results because the 95% credible interval of the associated effect covered 0, and had a marginal inclusion probability of only 0.004.

GDD hypothesis 2 Case: As was true for the previous case, consistency of effects and the marginal probability of inclusion were positively associated (Fig. 4). The six top variables in terms of marginal inclusion probability (length, tagging site: Cheboygan and Alpena, *Diporeia*, recovery month November, interaction effect between lake-average GDD and recovery month October) were similar to the top variables for GDD hypothesis 1. The major exceptions were for the recovery month October and the GDD associated variables (Fig. 4). Consistent with the results for GDD hypothesis 1, fish with greater length, longer lag between the tagging and recovery years, and released at tagging site Cheboygan, Alpena, and Fishing islands had greater net movement distance, while fish released at the tagging site with higher density of *Diporeia*, and recovered during September, and November had shorter net movement distance. The effect of recovery month October had a similar negative posterior mean, although the effect was less consistent. The less consistent effect of October is likely associated with the inclusion of the GDD associated variables for Hypothesis 2. Our second hypothesis of GDD was well supported by the variable selection results. The interaction effect between lake-average GDD and recovery month October, and the interaction effect between lake-average GDD and recovery month November were both consistent, with a negative posterior mean and the former was smaller than the latter. That is, fish tended to have smaller net movement distance in October and November if the lake-average GDD was greater, and the effect was larger in October than that in November. On the other hand, the interaction between lake-average GDD and recovery month December was also consistent, but with a positive

posterior mean, which suggested that fish tended to have greater net movement distance in December if the lake-average GDD was greater. An overall interpretation of these effects is a shift in the spawning season in association with GDD, with more fish close to the spawning grounds by October and having moved away by December, when GDD was higher.

3.2 Model-wise summary

Given the variable selection result support our second GDD hypothesis, we only present model-wise summary for GDD hypothesis 2 case. One or more of the four interaction variables (recovery month × GDD for the fish that were recovered from September, October, November, and December) were included in at least one out of the top 12 models. In addition to the three variables with marginal inclusion probability equals 1 in Fig. 4 (*Diporeia*, tagging site: Cheboygan and Alpena), the interaction effect recovery month October × GDD was also included in all 12 top models (Fig. 5). Recovery month November was included in nine of the 12 top models (all but models 5, 7, and 10), while the interaction variable November × $\mathrm{GDD}_{\mathrm{lake}}$ was included in the other three top models. Total length of tagged fish was included in eight out of the 12 top models, recovery month September was included in four out of the top models, the interaction variable December× $\mathrm{GDD}_{\mathrm{lake}}$ was included in three out of the top models, and recovery month December and the interaction variable September × $\mathrm{GDD}_{\mathrm{lake}}$ were included in one out of the top models. The top two models both had a posterior probability greater than 0.14. These models were similar. The best model (i.e., the highest posterior probability model) included six variables and the second best model included all those variables, plus recovery month September. Most estimated β_js are consistent across the top models, suggesting the effect of a variable was relatively uninfluenced by the presence of other variables in the models.

The best model (Model 1 in Fig. 5) for the fit with GDD hypothesis 2 is summarized in Table 1. From the best model, fish that were tagged and released from tagging sites Cheboygan and Alpena had longer net distance than fish released at other tagging sites. Lake whitefish with greater total length also tended to have greater net distance. Fish that were recovered in November consistently had shorter net distance than fish recovered in other months. In addition, shorter movement distance could be expected if the relative *Diporeia* density was higher near the spawning locations during the recovery year. The interaction term of month October and lake-average GDD resulted in shorter net distance when lake-average GDD was high.

3.3 Model diagnosis, simulation study, and comparison with tree-based methods

Model diagnosis— There is no evidence for lack-of-fit of the top models under both GDD hypothesis cases. In particular the scatterplot of predicted and realized χ^2appear consistent with a 1:1 relationship (Fig. S1 in online Appendix C) and the Bayesian p-values are much larger than 0.05, indicating that the null hypothesis that the observed data follow the hypothesized model is not rejected. We also did a residual analysis for the top model of both GDD hypothesis cases, and plotted averaged standard residuals for the MCMC samples associated with those top models versus selected (including both continuous variables and two way combinations of categorical predictors). We did not observe any suspicious patterns from the plot given: 1) all residuals are nearly symmetric

about zero, majority within (-3, 3), according to the 3-sigma rule, 2) there were no obvious trends in variation or mean across different values of the predictors.

Simulation study— In general, our BVS method had consistent performance at identifying important variables, and in identifying an appropriate model under scenarios with varying combinations of candidate variables (see online Appendix D). Effects of interactions, and of continuous and categorical variables were all likely to be selected when they actually had effects, and not to be selected when they did not have effects on the response variable. Across all scenarios, the true model was very likely to be included in the top two models (i.e., probability >=0.9), and most likely to be our top model (i.e., probability >=0.74).

Comparison with tree-based methods— The top variables from both tree-based methods in Figure S3 (online Appendix E) are consistent with Bayesian variable selection (BVS) results, although there were several exceptions. The first exception was for the GDD hypothesis 1 case, where GDD_Diff was not selected as important variable by BVS, but was selected as top variables by both boosted regression tree and random forest approaches. We believe that this is due to several high-leverage GDD_Diff observations (Fig. S4 in online Appendix E), which the regression tree methods see as nonlinear effects. A second exception, also for the GDD hypothesis 1 case, was that fish length had a high inclusion probability (0.78 with BVS) and was also a top variable for boosted regression trees but was not included in the top list for the random forests approach. A third exception was that the rank of the variable September was lower for the tree-based approaches than for BVS, and this was true for both GDD hypotheses, albeit the three approaches rank variable importance in different ways (probability of inclusion for BVS, see X axis of Fig. S3 and Fig. S5 for tree-based methods).

4. Discussion

The goal of this study was to develop a model framework for analysis of how factors impact net fish movement distance in a larger water body, based on conventional tag-recovery results, and apply the framework to lake whitefish spawning stocks in Lake Huron of the Laurentian Great Lakes of North America. Our framework used a data-driven Bayesian variable selection (BVS) method, where the candidate variables represented hypothesis about drivers of net movement distance. The hypotheses we evaluated were that the net movement distance of adult lake whitefish in the main basin of Lake Huron was related to 1) fish total length, 2) sex, 3) tag and release year, 4) recovery year, 5) recovery month, 6) year(s) between tagging and recovery, 7) *Diporeia* spp. density near the spawning locations relative to the lake-wide *Diporeia* spp. density, 8) relative difference between the tagging site and lake-wide growing degree days, and 9) the interaction term between lake-wide growing degree days and recovery month. Some of the above hypotheses were well supported by the results presented.

There was a consistent positive relationship between lake whitefish net movement distance and fish total length at the time of tagging. This is consistent with conclusions from previous studies of stream-dwelling fish, in which longer movement and home range was observed for larger fish (Gatz and Adams, 1994; Gunning and Shoop, 1963). This greater movement may be due to the increasing mass-specific bioenergetic costs of mobility with decreasing body size (Roff, 1991). Minns (1995) also found that the home range is related to body size in freshwater fisheries and is consistently larger in lakes than in rivers.

Because of the spawning site fidelity of lake whitefish, recovery months were expected to have effects on net movement distance. Ebener et al. (2010b), analyzing some of the same data but focused on different spatial and temporal scales with fewer predictor variables, also demonstrated that season of recapture played an important role in the distance moved by lake whitefish. Here, net movement distance was found to be negatively related to recovery months September, October and November, and positively related to December. This suggested that the spawning migration movement for lake whitefish generally occurred within months from September to November, and after that, fish tended to leave their spawning site and were actually further from the spawning location than in the baseline month of June.

Past research has documented that some life history events such as reproduction can be accelerated with warmer water temperature (Forseth et al., 1999). For example, the spawning of walleye has occurred earlier with earlier ice-out related to warmer temperature (Schneider et al., 2010). We found similar patterns in our study. When lake average GDD was higher, lake whitefish tended to move or stay closer to their spawning sites from September to November, and to be further away from their spawning sites in December. This suggests that fish may start their annual spawning migration runs earlier in warmer years after acquiring and processing energy needed for spawning. The underlying mechanism could be that fish have to either achieve a critical condition before the cost of migration/spawning can be offset (Forseth et al., 1999), or to accumulate enough energy to survive a winter starvation period before spawning.

Although the decline of *Diporeia* spp. density in the Laurentian Great Lakes due to the establishment of dreissenid mussels has been argued as the main reason of lake whitefish expanding their movement range (Ebener et al., 2010b; Rennie et al., 2012), we know of no other direct evaluation of an effect of *Diporeia* density on movement. Our study evaluated this hypothesis by including relative *Diporeia* spp. density as a predictor for lake whitefish net movement distance, and we found that when relative *Diporeia* spp. density was high near the spawning grounds, lake whitefish tended to stay closer to their spawning site. This implied that fish might expand their foraging area when *Diporeia* density was low near their preferred habitat. Our analysis also found an effect of the relative density of *Diporeia* within a year, which suggests a pattern related to the density of this prey, not just a general change in movement over time throughout the lake as *Diporeia* declined.

Lake whitefish tagged and released from the tagging sites Cheboygan and Alpena had consistently greater net distance than those released from other areas. The underlying reasons may be relate to the bathymetry and shoreline features of Lake Huron. Deep water (>80 m) near Cheboygan and Alpena may restrict the movement of Cheboygan and Alpena spawning stocks to north-south direction where there is a large area with relative shallow water. In contrast, the spawning stocks in Detour and Burnt Inlands may be constrained from moving south by the deep water in north of the main basin of Lake Huron, so that they tended to move in the east-west direction. Considering the shape of Lake Huron and the locations of those spawning stocks, movement in the north-south direction allows longer movement distance than in the east-west direction.

There was similarity but also some differences in variable selection between our Bayesian variable selection and tree-based methods. One notable difference between tree-based methods and the Bayesian method is in the inclusion of GDD difference in GDD

hypothesis 1 case for the tree-based methods but not by BVS. The overall neutral effect and low importance for the BVS was apparently because a few high-leverage points were treated as noise. By recursively partitioning the data according to different ranges of predictors, the tree-based methods are less sensitive to those points. However, such localized results based on small samples can hardly provide any general predictability. Rätsch et al. (2001) also found that overfitting can occur for regression tree-based methods using a boosting algorithm when there is a lot of noise.

Our BVS method can be used for various different species and any water system meeting our input requirements. For conventional tagging studies done in large lakes (e.g., Lake Huron as in our case) or oceans, shortest water distance can be used as response variable; while for a tagging study done in a river, a river network needs to be built /considered for calculating (net) movement distance. Given that our Bayesian variable selection method penalizes the number of selected variables, it has the potential to perform well for other cases with more candidate explanatory variables than we used in our application. In addition, the approach is adaptable to situations where residuals might be correlated. We assumed no such correlations given the spatial distribution of tagging sites and inclusion of spatial covariates (e.g., tagging sites), but in other situations there could be spatial structure that should be accounted for in random part of the model. In such cases correlations could be made a function of a measured quantity like distance between tagging sites, and our code and detailed description of the model in the supplement outlines how this can be done. In addition, our Bayesian method also allows extra flexibility such as including: (1) random effects to cope with grouping variables with a large number of outcomes, which can greatly improve the prediction by better explaining the variability; (2) prior information for the effects of variables with flexible choices that can be leveraged from a broad catalog in the Bayesian variable selection literature. Thus we believe our work established a framework that could facilitate additional studies of animal movement based on conventional tagging data.

We made several simplifying assumptions and choices in our analysis. Firstly, we assumed 100% spawning site fidelity, so for the environmental factors *Diporeia* spp. density and GDDs, only data for the year of recovery were used. While fidelity is likely not 100%, available data suggest it is quite high for lake whitefish (Ebener et al., 2010b). Secondly, the T_0 used for the calculation of cumulative GDD is 5°C (Rennie et al., 2009), but it is possible that this is not the best threshold or that fish are responding to temperature in a different or more complex fashion than we assumed. We believe that violation of the 100% fidelity assumption and the GDD assumptions would act to obscure effects of *Diporeia* and tagging site rather than cause us to discover artefactual effects. Thirdly, we assumed similar tag reporting rates across all recovery basins, so data were not weighted across different recovery basins. Violation of this assumption could be influencing details of our results. However, we suspect the larger qualitative effects are real rather than artifacts of such a violation. If there were dominating differences in tag reporting rates among basin, we would have expected that to be reflected in consistent tagging site effects for sites within basins, which we did not see in our results.

Acknowledgements

Funding for this project was provided by Quantitative Fisheries Center supporting partners, Michigan DNR through the Partnership for Ecosystem Research and Management program, and grants from the Great Lakes Fishery Commission Fishery

573 Research Program and the Great Lakes Fishery Trust (Project 2012.1250). We thank the
574 commercial fisherman and the staff of the Inter-Tribal Fisheries and Assessment
575 Program, the Hammond Bay Biological Station, the Great Lakes Fishery Commission,
576 and Michigan DNR for their helps in tagging fish in very uncomfortable weather
577 conditions. We thank Dr. Richard P. Barbiero for providing *Diporeia* abundances data,
578 and Dr. Lacey Mason for providing surface water temperature data. We acknowledge the
579 support of the Michigan State University High Performance Computing Center and the
580 Institute for Cyber-Enabled Research. We wish to thank the Modeling Subcommittee of
581 the Technical Fisheries Committee for 1836 Treaty-ceded waters for providing input on
582 our research. This is publication 2017-XX of the Michigan State University Quantitative
583 Fisheries Center.
584

Appendix A. Calculation of shortest water distance and GDD

Calculation of shortest water distance and GDD can be found in the online version, at XXX.

Appendix B. Model implementation

Model implementation can be found in the online version, at XXX.

Appendix C. Model diagnostics

Model diagnostics can be found in the online version, at XXX.

Appendix D. Simulation study

Simulation study methods and results can be found in the online version, at XXX.

Appendix E. Comparison with two tree-based methods

Two tree-based methods and their results can be found in the online version, at XXX.

Appendix F. Codes for Bayesian variable selection, and tree based methods

Code for Bayesian variable selection and tree-based methods can be found online, at https://doi.org/10.6084/m9.figshare.5177206.

References

Albanese, B., Angermeier, P.L., Dorai-Raj, S., 2004. Ecological correlates of fish movement in a network of Virginia streams. Can. J. Fish. Aquat. Sci. 61, 857–869. doi:10.1139/f04-096

Barbieri, M.M., Berger, J.O., 2004. Optimal predictive model selection. Ann. Stat. 32, 870–897. doi:10.1214/009053604000000238

Barbiero, R.P., Schmude, K., Lesht, B.M., Riseng, C.M., Warren, G.J., Tuchman, M.L., 2011. Trends in Diporeia populations across the Laurentian Great Lakes, 1997–2009. J. Great Lakes Res. 37, 9–17. doi:10.1016/j.jglr.2010.11.009

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324

Chezik, K.A., Lester, N.P., Venturelli, P.A., Tierney, K., 2014. Fish growth and degree-days II: selecting a base temperature for an among-population study. Can. J. Fish. Aquat. Sci. 71, 1303–1311. doi:10.1139/cjfas-2013-0615

Claeskens, G., Hjort, N.L., 2008. Model selection and model averaging. Cambridge University Press.

Drouineau, H., Bau, F., Alric, A., Deligne, N., Gomes, P., Sagnes, P., 2017. Silver eel downstream migration in fragmented rivers: use of a Bayesian model to track movements triggering and duration. Aquat. Living Resour. 30. doi:10.1051/alr/2017003

Ebener, M.P., Brenden, T.O., Jones, M.L., 2010a. Estimates of fishing and natural mortality rates for four Lake Whitefish stocks in Northern Lakes Huron and Michigan. J. Great Lakes Res. 36, 110–120. doi:10.1016/j.jglr.2009.06.003

Ebener, M.P., Brenden, T.O., Wright, G.M., Jones, M.L., Faisal, M., 2010b. Spatial and temporal distributions of lake whitefish spawning stocks in Northern lakes Michigan and Huron, 2003–2008. J. Great Lakes Res. 36, 38–51. doi:10.1016/j.jglr.2010.02.002

Ethier, D.M., Koper, N., Nudds, T.D., 2017. Spatiotemporal variation in mechanisms driving regional-scale population dynamics of a Threatened grassland bird. Ecol. Evol. 7, 4152–4162. doi:10.1002/ece3.3004

Forseth, T., Nesje, T.F., Jonsson, B., Harsaker, K., 1999. Juvenile migration in brown trout: a consequence of energetic state. J. Anim. Ecol. 68, 783–793. doi:10.1046/j.1365-2656.1999.00329.x

Fu, C., Fanning, L.P., 2004. Spatial Considerations in the Management of Atlantic Cod off Nova Scotia, Canada. North Am. J. Fish. Manag. 24, 775–784. doi:10.1577/M03-134.1

Gatz, A.J., Adams, S.M., 1994. Patterns of movement of centrarchids in two warmwater streams in eastern Tennessee. Ecol. Freshw. Fish 3, 35–48. doi:10.1111/j.1600-0633.1994.tb00105.x

Gelman, A., Meng, X.-L., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. 6, 733–807. doi:10.1.1.142.9951

Gilliam, J.F., Fraser, D.F., 2001. MOVEMENT IN CORRIDORS: ENHANCEMENT BY PREDATION THREAT, DISTURBANCE, AND HABITAT STRUCTURE. Ecology 82, 258–273. doi:10.1890/0012-9658(2001)082[0258:MICEBP]2.0.CO;2

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732. doi:10.1093/biomet/82.4.711

Gunning, G.E., Shoop, C.R., 1963. Occupancy of home range by longear sunfish, Lepomis m. megalotis (Rafinesque), and bluegill, Lepomis m. macrochirus Rafinesque. Anim. Behav. 11, 325–330. doi:10.1016/S0003-3472(63)80119-0

Hutchings, J.A., 1996. Spatial and temporal variation in the density of northern cod and a review of hypotheses for the stock's collapse. Can. J. Fish. Aquat. Sci. 53, 943–962. doi:10.1139/f96-097

Li, Y., Bence, J.R., Brenden, T.O., 2015. An evaluation of alternative assessment approaches for intermixing fish populations: a case study with Great Lakes lake whitefish. ICES J. Mar. Sci. 72, 70–81. doi:10.1093/icesjms/fsu057

Lidicker, W.Z., Stenseth, N.C., 1992. To disperse or not to disperse: who does it and why?, in: Animal Dispersal. Springer Netherlands, Dordrecht, pp. 21–36. doi:10.1007/978-94-011-2338-9_2

McNickle, G.G., Rennie, M.D., Sprules, W.G., 2006. Changes in Benthic Invertebrate Communities of South Bay, Lake Huron Following Invasion by Zebra Mussels (Dreissena polymorpha), and Potential Effects on Lake Whitefish (Coregonus clupeaformis) Diet and Growth. J. Great Lakes Res. 32, 180–193. doi:10.3394/0380-1330(2006)32[180:CIBICO]2.0.CO;2

Minns, C.K., 1995. Allometry of home range size in lake and river fishes. Can. J. Fish. Aquat. Sci. 52, 1499–1508. doi:10.1139/f95-144

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian Variable Selection in Linear Regression. J. Am. Stat. Assoc. 83, 1023–1032.

Mohr, L.C., Nalepa, T.F., 2005. Proceedings of a workshop on the dynamics of lake whitefish (Coregonus clupeaformis) and the amphipod Diporeia spp. in the Great Lakes, Great Lakes Fish. Comm. Tech. Rep.

Polacheck, T., Eveson, J.P., Laslett, G.M., Pollock, K.H., Hearn, W.S., 2006. Integrating catch-at-age and multiyear tagging data: a combined Brownie and Petersen estimation approach in a fishery context. Can. J. Fish. Aquat. Sci. 63, 534–548. doi:10.1139/f05-232

Price, H., Pothoven, S.A., McCormick, M.J., Jensen, P.C., Fahnenstiel, G.L., 2003. Temperature Influence on Commercial Lake Whitefish Harvest in Eastern Lake Michigan. J. Great Lakes Res. 29, 296–300. doi:10.1016/S0380-1330(03)70434-1

Radinger, J., Wolter, C., 2014. Patterns and predictors of fish dispersal in rivers. Fish Fish. 15, 456–473. doi:10.1111/faf.12028

Rätsch, G., Onoda, T., Müller, K.-R., 2001. Soft Margins for AdaBoost. Mach. Learn. 42, 287–320. doi:10.1023/A:1007618119488

Rennie, M.D., Ebener, M.P., Wagner, T., 2012. Can migration mitigate the effects of ecosystem change? Patterns of dispersal, energy acquisition and allocation in Great Lakes lake whitefish (Coregonus clupeaformis). Adv. Limnol. 63, 455–476. doi:10.1127/advlim/63/2012/455

Rennie, M.D., Sprules, W.G., Johnson, T.B., 2009. Factors affecting the growth and condition of lake whitefish (Coregonus clupeaformis). Can. J. Fish. Aquat. Sci. 66, 2096–2108. doi:10.1139/F09-139

Roff, D.A., 1991. Life History Consequences of Bioenergetic and Biomechanical Constraints on Migration. Am. Zool. 31, 205–216. doi:10.1093/icb/31.1.205

Rothschild, B.J., 2007. Coherence of Atlantic Cod Stock Dynamics in the Northwest Atlantic Ocean. Trans. Am. Fish. Soc. 136, 858–874. doi:10.1577/T06-213.1

691 Runge, C.A., Martin, T.G., Possingham, H.P., Willis, S.G., Fuller, R.A., 2014.
692 Conserving mobile species. Front. Ecol. Environ. 12, 395–402. doi:10.1890/130237
693 Schneider, K.N., Newman, R.M., Card, V., Weisberg, S., Pereira, D.L., 2010. Timing of
694 Walleye Spawning as an Indicator of Climate Change. Trans. Am. Fish. Soc. 139,
695 1198–1210. doi:10.1577/T09-129.1
696 Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Stat. 6, 461–464.
697 doi:10.1214/aos/1176344136
698 Vandergoot, C.S., Brenden, T.O., 2014. Spatially Varying Population Demographics and
699 Fishery Characteristics of Lake Erie Walleyes Inferred from a Long-Term Tag
700 Recovery Study. Trans. Am. Fish. Soc. 143, 188–204.
701 doi:10.1080/00028487.2013.837095
702 Vincenty, T., 1975. DIRECT AND INVERSE SOLUTIONS OF GEODESICS ON THE
703 ELLIPSOID WITH APPLICATION OF NESTED EQUATIONS. Surv. Rev. 23,
704 88–93. doi:10.1179/sre.1975.23.176.88
705 Woznicki, S.A., Nejadhashemi, A.P., Abouali, M., Herman, M.R., Esfahanian, E.,
706 Hamaamin, Y.A., Zhang, Z., 2016. Ecohydrological modeling for large-scale
707 environmental impact assessment. Sci. Total Environ. 543, 274–286.
708 doi:10.1016/j.scitotenv.2015.11.044
709

Table 1. Summary of candidate variables/terms and their interpretation and relationship to hypotheses. A. For continuous variables, "Hypothesis" (first column) states our a priori hypothesis associated with the variable, and the second column indicates sign of associated coefficient that would support that hypothesis. B. Similarly for interaction terms, but here a single hypothesis (our GDD Hypothesis 2) is associated with all interaction terms, and the second column describes the interpretation of coefficients and the pattern in their sign that would support the hypothesis. C. For categorical (dummy) variables we did not have explicit a priori hypothesis for the sign of coefficients but did hypothesize that these factors could influence net distance. For these variables one level of a factor is the baseline with coefficient fixed at zero, and this level (category) is given in the first column and interpretation of the sign of other coefficients in the second column. For A through C, "X" in the "GDD H1" column indicates that the variable was a candidate variable/term in our variable selection process for the GDD Hypothesi1 1 Case, and the GDD H2 column likewise indicates if the variable/term was a candidate variable for the GDD Hypothesis 2 Case. The last row summarizes the total number of candidate variables for each GDD hypothesis.

Table 2. Posterior mean and 95% credible intervals for parameters of the highest posterior probability model.

Fig. 1. Map of the study area (Lake Huron) and seven tag release (spawning) sites. Of total 1368 recoveries, 659 were from Detour, 300 from Cheboygan, 243 from Burnt Island, 42 from Saginaw Bay, 43 from Sarnia, 56 from Alpena, and 25 from Fishing Islands.

Fig. 2. Posterior distributions for the number of selected variables (i.e., $q - 1$). The x-axis starts at 5 because all models selected at least five variables.

Fig. 3. Variable-wise summary results (posterior mean with 95% credible intervals) of the effect of variables (the β_j), with variables named on y-axis for the case with GDD hypothesis 1. Bars are highlighted by red color when the 95% credible interval does not cover 0, which is defined as a consistent effect. The number above each bar is the marginal inclusion probability.

Fig. 4. Variable-wise summary results (posterior mean with 95% credible interval of the effect β_j for the jth variable, as indicated in y-axis) for the case with GDD hypothesis 2. Bars are highlighted by red color when the 95% credible interval does not cover 0, which was defined as a consistent effect. The number above each bar is the marginal inclusion probability.

Fig. 5. Model-wise summary for top 12 models ranked according to their posterior probability mass, for the case of GDD hypothesis 2. Variables that were included in the top 12 models are given on the y-axis. Horizontal bar represents posterior 95% credible intervals and symbols on each bar the posterior mean for each coefficient included in a model, with the associated model given to the left of the bar. Thus when more bars are given for a variable it was included in more models.

Table 1

A. Continuous Variables				
Variable Name	Hypothesis	If support, sign of covariate	GDD H1	GDD H2
Length	Greater total length, fish range further from tagging site.	>0	X	X
years_lag	Longer lag between tagging and recovery year, recoveries tend to be further from tagging site.	>0	X	X
Diporeia	Higher relative *Diporeia* spp. density near the tagging site, fish stay closer to their tagging site.	<0	X	X
GDD_Diff	Greater GDD at the tagging location than the lake average, fish stay closer to their tagging site.	<0	X	
B. Interaction Terms				
Names	Hypothesis	Sign of coefficient	GDD H1	GDD H2
Sep× GDD_{lake} Oct× GDD_{lake} Nov× GDD_{lake} Dec× GDD_{lake}	In years when lake average GDD is higher there is a shift in spawning timing. This is reflected in shorter net distances in one or more adjacent spawning months, and longer net distances in later months.	If <0, fish are closer to tagging site with higher GDD_{lake} during that month, and if >0 further away. Support for hypothesis would be >0 coefficient for one or more adjacent months of Sep – Nov, and <0 coefficient for later months.		X
C. Categorical (Dummy) Variables				
Variable Names	Baseline category (effect was 0)	Interpretation of coefficient	GDD H1	GDD H2
tag_site: Cheboygan tag_site: Burnt_Island tag_site: Alpena tag_site: Sarnia	Fish tagged and released from Detour (Figure 1)	If >0, larger net distance than baseline; if <0, shorter net distance than baseline	X	X

tag_site: Saginaw_Bay tag_site: Fish_Islands sex: Female rec_Y: 2003 rec_Y: 2004 rec_Y: 2005 rec_Y: 2007 rec_Y: 2008 rec_Y: 2009 rec_Y: 2010 rec_Y: 2011 rec_M:7 rec_M:8 rec_M:9 rec_M:10 rec_M:11 rec_M:12 rec_M:1 rec_M:2 rec_M:3 rec_M:4 rec_M:5 tag_Y: 2003 tag_Y: 2005 tag_Y: 2006	Male tagged fish Tagged fish recovered from 2006 Tagged fish recovered in June of each year Fish tagged and released from 2004	Same as above Same as above Same as above Same as above	X X X X	X X X X
Total number of candidate variables for each case (without intercept)			33	36

Table 2

Variable	Mean	Lower	Upper
Rec_M:11	-0.49	-0.63	-0.35
Oct × GDD_{lake}	-0.45	-0.58	-0.32
Diporeia	-0.17	-0.22	-0.12
length	0.09	0.05	0.14
tag_site: Cheboygan	0.69	0.57	0.80
tag_site: Alpena	1.04	0.78	1.30

Figure 1


Detour
Burnt Island
Killarney
Mackinac Island
Manitoulin Island
Cheboygan
Petoskey
Alpena
Fishing Islands
MICHIGAN
Saginaw Bay
N
0
40
80 km
Flint
Sarnia
London
46
45
44
43
−85
−84
−83
−82
−81
Latitude
Longitude

Figure 2

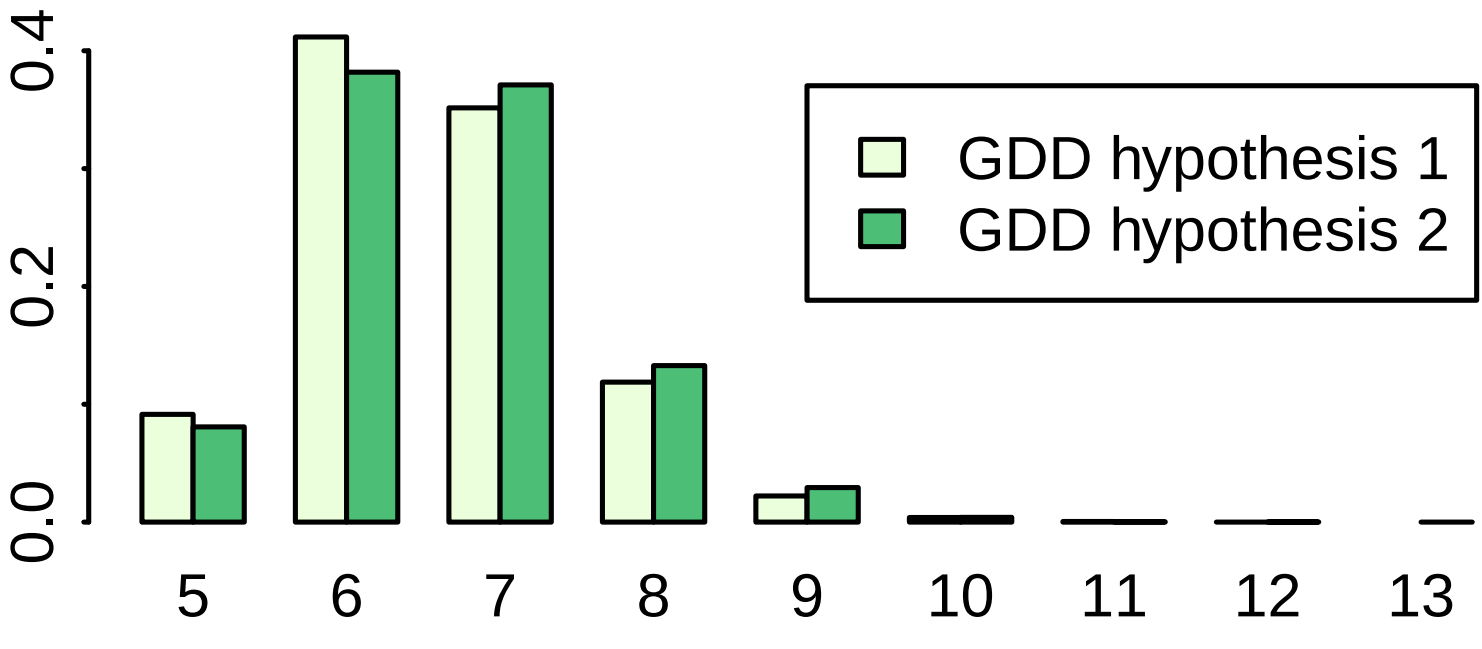


Figure 3

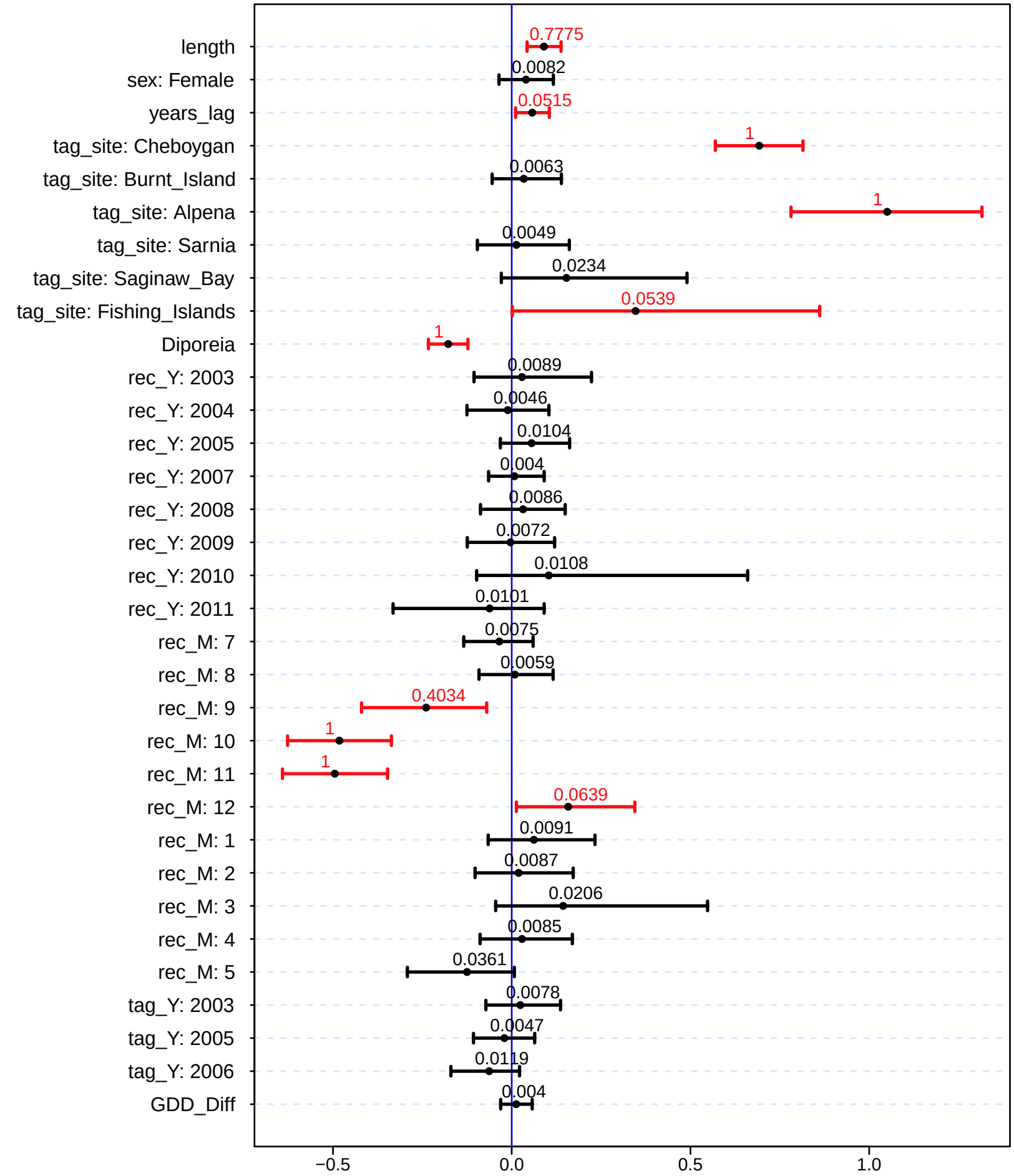

length
sex: Female
years_lag
tag_site: Cheboygan
tag_site: Burnt_Island
tag_site: Alpena
tag_site: Sarnia
tag_site: Saginaw_Bay
tag_site: Fishing_Islands
Diporeia
rec_Y: 2003
rec_Y: 2004
rec_Y: 2005
rec_Y: 2007
rec_Y: 2008
rec_Y: 2009
rec_Y: 2010
rec_Y: 2011
rec_M: 7
rec_M: 8
rec_M: 9
rec_M: 10
rec_M: 11
rec_M: 12
rec_M: 1
rec_M: 2
rec_M: 3
rec_M: 4
rec_M: 5
tag_Y: 2003
tag_Y: 2005
tag_Y: 2006
GDD_Diff
0.7775
0.0082
0.0515
1
0.0063
1
0.0049
0.0234
0.0539
1
0.0089
0.0046
0.0104
0.004
0.0086
0.0072
0.0108
0.0101
0.0075
0.0059
0.4034
1
1
0.0639
0.0091
0.0087
0.0206
0.0085
0.0361
0.0078
0.0047
0.0119
0.004
−0.5
0.0
0.5
1.0

Figure 4

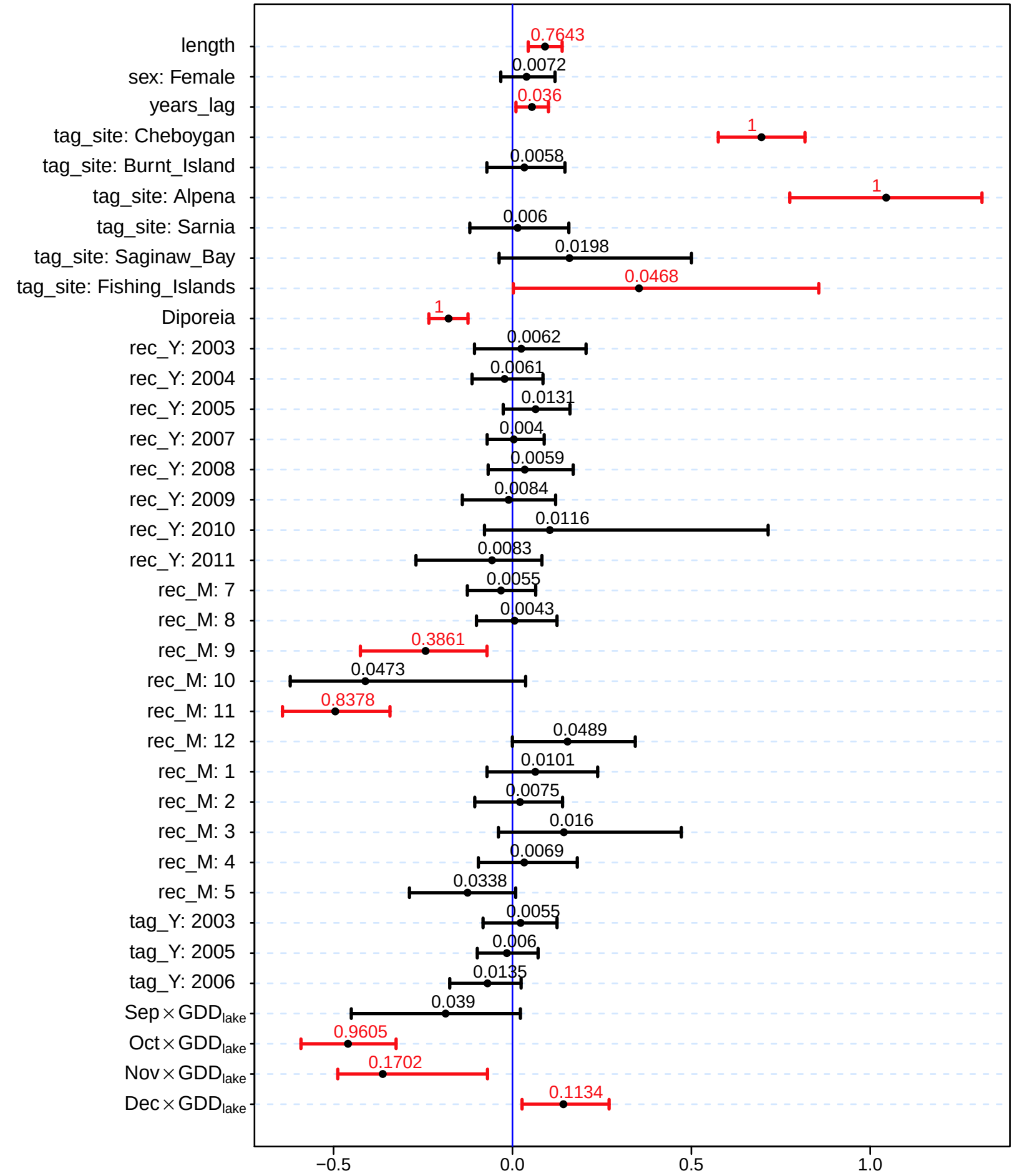
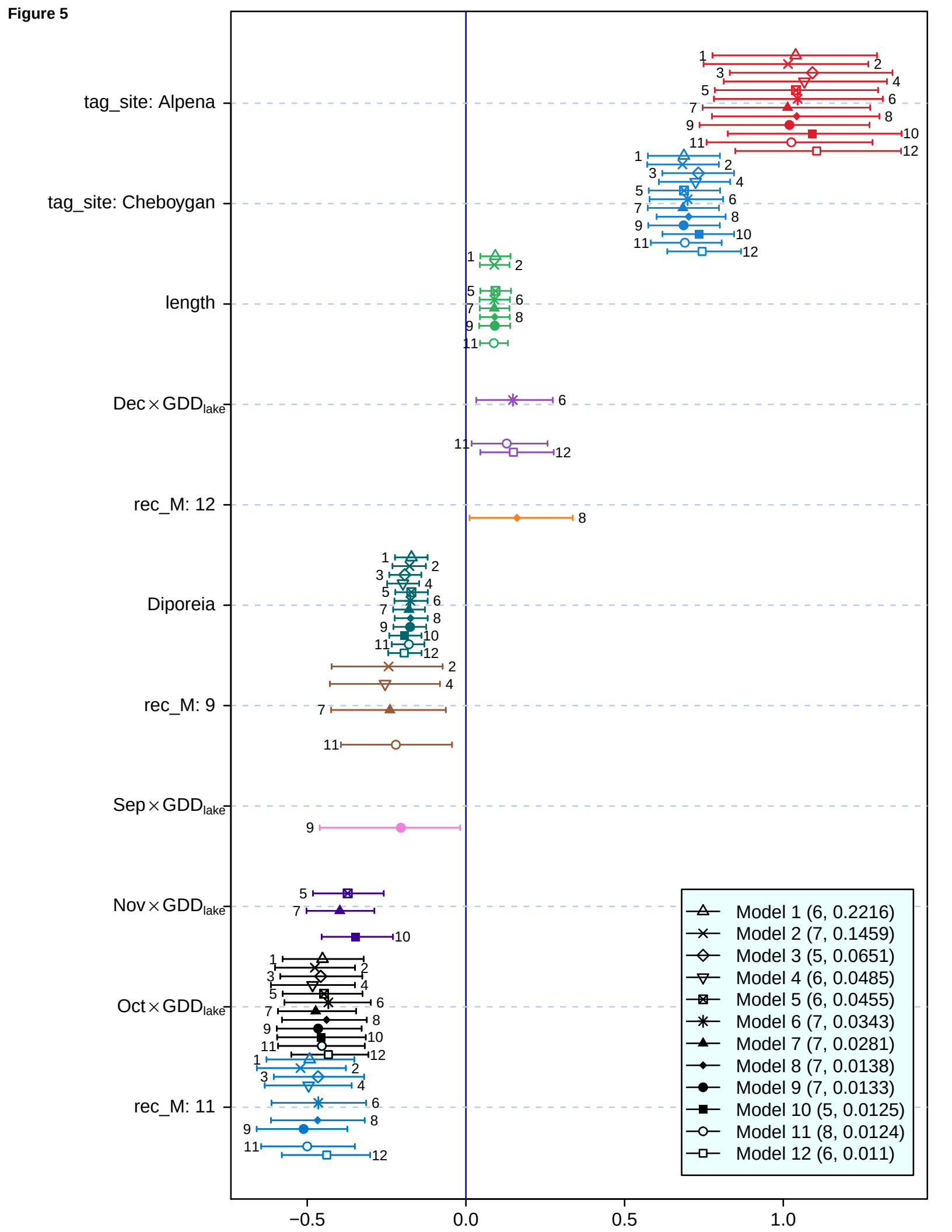

length
sex: Female
years_lag
tag_site: Cheboygan
tag_site: Burnt_Island
tag_site: Alpena
tag_site: Sarnia
tag_site: Saginaw_Bay
tag_site: Fishing_Islands
Diporeia
rec_Y: 2003
rec_Y: 2004
rec_Y: 2005
rec_Y: 2007
rec_Y: 2008
rec_Y: 2009
rec_Y: 2010
rec_Y: 2011
rec_M: 7
rec_M: 8
rec_M: 9
rec_M: 10
rec_M: 11
rec_M: 12
rec_M: 1
rec_M: 2
rec_M: 3
rec_M: 4
rec_M: 5
tag_Y: 2003
tag_Y: 2005
tag_Y: 2006
Sep × GDD lake
Oct × GDD lake
Nov × GDD lake
Dec × GDD lake
0.7643
0.0072
0.036
1
0.0058
1
0.006
0.0198
0.0468
1
0.0062
0.0061
0.0131
0.004
0.0059
0.0084
0.0116
0.0083
0.0055
0.0043
0.3861
0.0473
0.8378
0.0489
0.0101
0.0075
0.016
0.0069
0.0338
0.0055
0.006
0.0135
0.039
0.9605
0.1702
0.1134
−0.5
0.0
0.5
1.0

Figure 5


tag_site: Alpena
tag_site: Cheboygan
length
Dec × GDD_lake
rec_M: 12
Diporeia
rec_M: 9
Sep × GDD_lake
Nov × GDD_lake
Oct × GDD_lake
rec_M: 11
−0.5
0.0
0.5
1.0
Model 1 (6, 0.2216)
Model 2 (7, 0.1459)
Model 3 (5, 0.0651)
Model 4 (6, 0.0485)
Model 5 (6, 0.0455)
Model 6 (7, 0.0343)
Model 7 (7, 0.0281)
Model 8 (7, 0.0138)
Model 9 (7, 0.0133)
Model 10 (5, 0.0125)
Model 11 (8, 0.0124)
Model 12 (6, 0.011)

Supplementary material for on-line publication only

Click here to download Supplementary material for on-line publication only: Second Revision_Supplementary Materials.docx