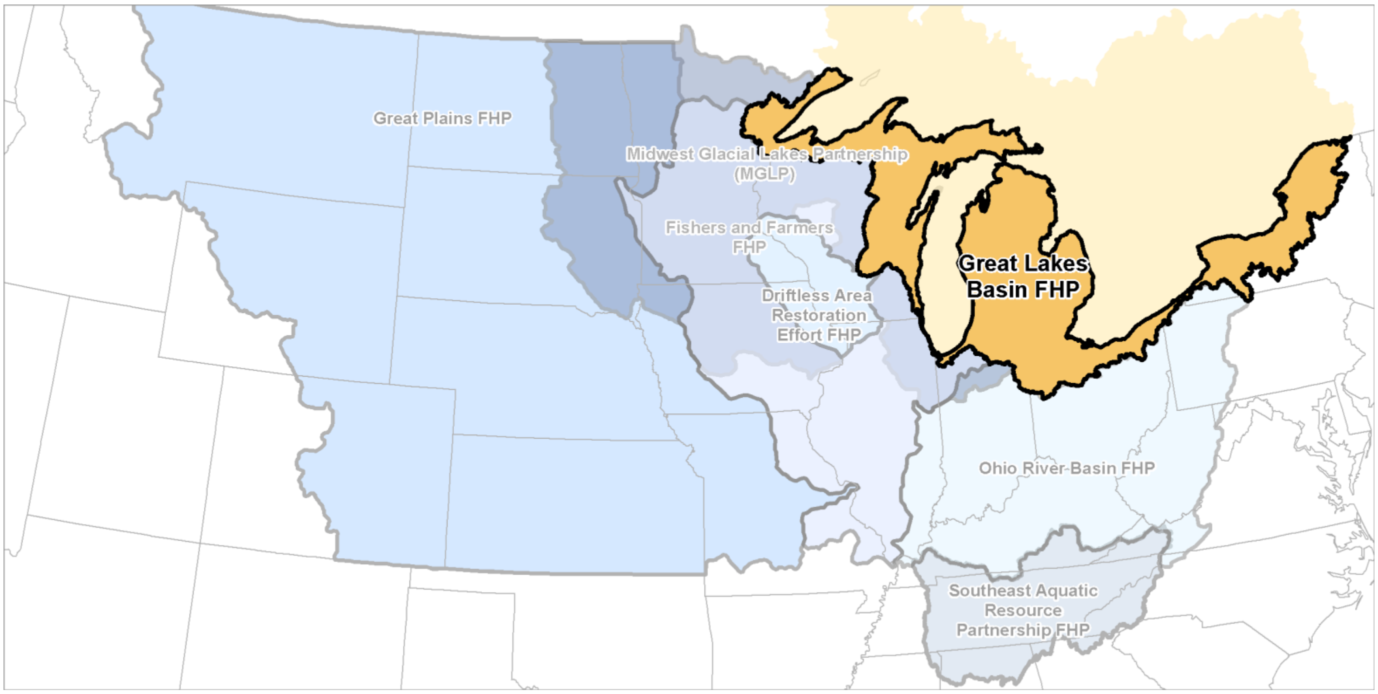
Midwest Fish Habitat Partnership Fish Habitat Modeling Results

Great Lakes Basin Fish Habitat Partnership

Model Summaries
1/30/2015

Brook Trout (*Salvelinus fontinalis*): Probability of Presence
Coldwater Species: Probability of Presence
Lithophilic Species Richness: Probability of Presence
Walleye (*Sander vitreus*): Probability of Presence
Large River Species: Probability of Presence



Authors
Jason Clingerman, Aquatic Ecologist/Modeler – Downstream Strategies
Todd Petty, Aquatic Ecologist – West Virginia University
Fritz Boettner, Project Manager/GIS – Downstream Strategies
Sally Letsinger, GIS – Geodata Basics
Jackie Strager, GIS – West Virginia University
Evan Hansen, Chief Scientist – Downstream Strategies

Downstream Strategies
295 High Street, Suite 3
Morgantown, WV 26505
www.downstreamstrategies.com

Report date:
January 30, 2015

ACKNOWLEDGEMENTS

Coordinator: Mark Brouder, United States Fish and Wildlife Service

GIS and Data Support: Jana Stewart, United States Geological Survey

Science Team members: Randy Claramunt, Michigan Department of Natural Resources; Lucinda Johnson, University of Minnesota-Duluth; Jim Mckenna, United States Geological Survey; Paul Seelbach, United States Geological Survey; Jana Stewart, United States Geological Survey; and Li Wang, University of Michigan.

This project was originally conceived, guided, and supported by Maureen Gallagher.

This project would not have been possible without the coordination and facilitation provided by Pam Dryer to the Great Lakes Basin Fish Habitat Partnership. We would like to thank the United States Fish and Wildlife Service, United States Geological Survey, Michigan Department of Natural Resources, University of Minnesota-Duluth, and University of Michigan for their resources, time, and expertise. We would also like to acknowledge all of the individuals who collected and provided data, feedback, and local expertise to the Great Lakes Aquatic GAP Project, as these data served as the basis for our model development.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

ABBREVIATIONS

BRT	boosted regression tree
ASI	anthropogenic stress index
HQI	natural quality index
CV	cross-validation
DS	Downstream Strategies
FHP	Fish Habitat Partnership
GIS	geographic information systems
GLB	Great Lakes Basin
NHD	National Hydrography Dataset
NPDES	National Pollutant Discharge Elimination System
ROC	receiver operating characteristic
USFWS	United States Fish and Wildlife Service

1. INTRODUCTION

1.1 Document outline

This report provides a summary of the key outcomes resulting from models developed by Downstream Strategies (DS) for use in assessing aquatic habitats for the Great Lakes Basin Fish Habitat Partnership (GLBFHP). The appendices provide additional maps, charts, and metadata useful for evaluating the results of the models.

The following are included for each model's results summary.

- Subsection one, *Modeling inputs*, discusses the predictor and response variables used in the analyses.
- Subsection two, *Modeling process*, reviews the outcomes of the statistical modeling process using BRTs, including information on model certainty. Additional, variable influence and functional relationships between predictor and response variables are included under corresponding headings.
- Subsection three, *Post-modeling*, contains information resulting from the post-modeling process, including information on the top stressors and natural habitat variables and their role in the calculation of the final indices.
- Subsection four, *Mapped results*, contains maps for visualizing conditions at the 1:100k catchment scale and includes maps of expected current probability of presence, stress, and natural quality; it also provides examples of how the two post-modeling indices—habitat quality and anthropogenic stress—can be combined to inform restoration priorities and how those priorities can be visualized in a spatially explicit manner.

1.2 Project background

Fishery and aquatic scientists often assess habitats to understand the distribution, status, stressors, and relative abundance of aquatic resources. Due to the spatial nature of aquatic habitats and the increasing scope of management needs, traditional analytical assessment methods are often limited in their ability to address complex and dynamic aquatic systems. Advancements in the geographic information systems (GIS) field and related technologies have enabled scientists and managers to more effectively collate, archive, display, analyze, and model spatial and temporal data. For example, spatially explicit habitat assessment models allow for a more robust interpretation of many terrestrial and aquatic datasets, including physical and biological monitoring data, habitat diversity, watershed characteristics, and socioeconomic parameters.

Initially, DS was contracted by the United States Fish and Wildlife Service (USFWS) to create a spatially explicit data analysis and modeling system for assessing fish habitat conditions for the GLBFHP based on a range of metrics. These analyses provided data and tools for specific aquatic species and were constructed at the scale of the GLBFHP. These results were useful, but improvements to the post modeling process and the incorporation of additional data formed the motivation to update the models previously created for the GLBFHP. This project built upon the knowledge gained and the framework designed during the initial GLBFHP-scale modeling, but utilized a new methodology for assigning stress and determining natural quality of aquatic habitats. This improved methodology was developed for the Midwest Regional Fish Habitat Assessment and will allow for increased functionality for the resulting web-based decision support tool. Additional response data was also added to the brook trout model along with an updated measure of predicted stream temperature, which was used as a predictor variable in all five model updates.

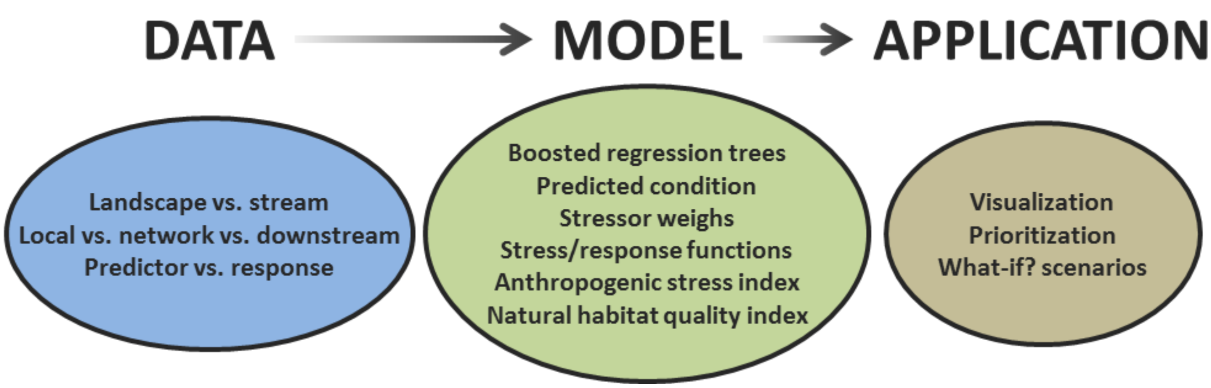
Generally, the models, analyses, and data produced as a result of this project are intended to enable a unique, broad, and spatially explicit understanding of the links between natural habitat conditions, human influences on aquatic habitats, and aquatic health. Specifically, the outcomes can be utilized to conduct fish habitat condition assessments based on a range of stakeholder-specified metrics and modeling endpoints to help determine natural drivers of aquatic conditions, as well as major stressors within the GLBFHP. The ultimate goal is to improve understanding of how local and regional processes influence stream conditions in the region and to provide additional knowledge, data, and tools to help prioritize and drive conservation and restoration projects throughout the GLBFHP.

1.3 Overview of the assessment process

1.3.1 *Modeling*

A diagram of the general assessment process is outlined in Figure 1. DS received landscape and aquatic data from the GLBP to develop models and tools for visualizing expected current and potential future conditions and prioritizing management actions.

Figure 1: Diagram of the habitat assessment process

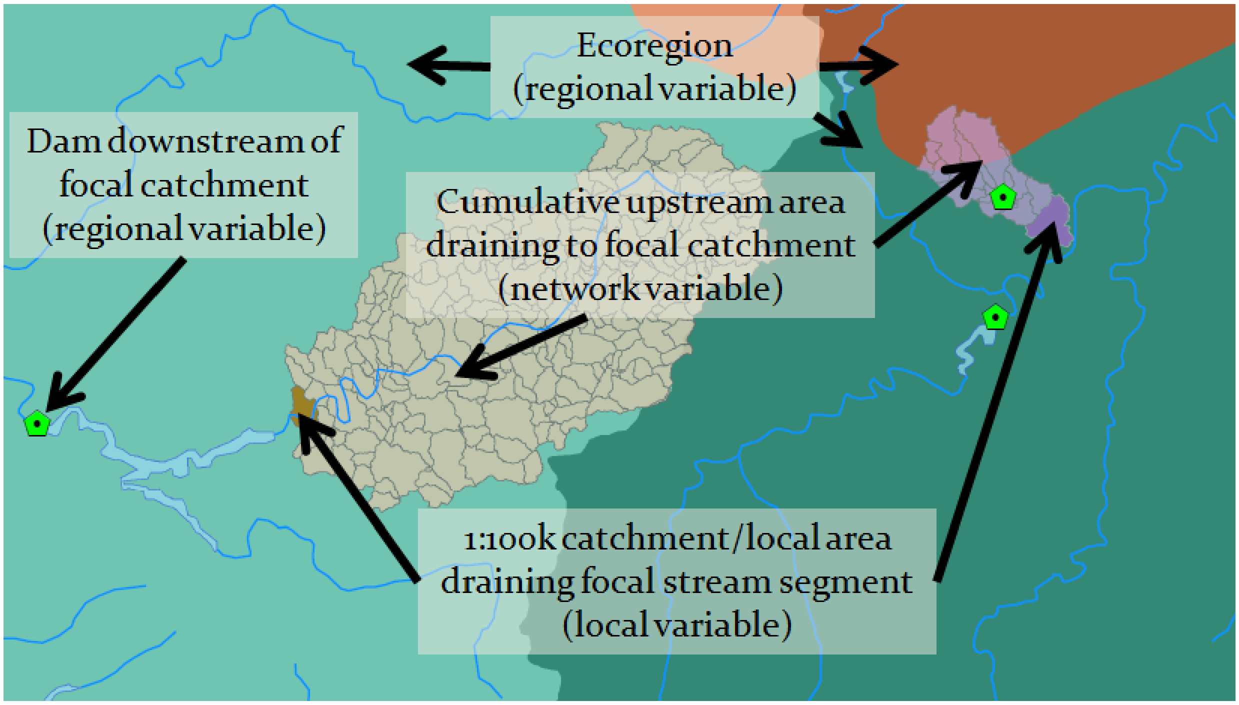


The data provided by the GLBFHP for use in the modeling process can be broken down into two categories: response and predictor variables.

The response variables for this project are presence-absence datasets of freshwater stream fish or fish guilds. For this assessment, a fish guild is defined as a group of fish that have similar habitat requirements and are relatively intolerant to habitat degradation. There were five response variables used in this assessment: (1) brook trout, (2) coldwater species, (3) walleye, (4) large river species, and (5) lithopillic spawners. A separate model was created for each of the five response variables.

The predictor variables are typically measures of land use or land cover derived from GIS, such as percent impervious surface area or road crossing density. Although the response variable is always measured at the same local scale (e.g., individual sample site on a stream), the predictor variables are compiled at multiple scales (Figure 2), including the local scale (e.g., single 1:100k National Hydrography Dataset (NHD) stream catchment), the network scale (e.g., all upstream catchments and the local catchment), or the regional scale (e.g., ecoregion).

Figure 2: Diagram and examples of different scales of data used for predictor variables



For this assessment nearly all of the predictor and response data necessary was already held by DS from the prior GLBFHP assessment. Updated fish sample data—brook trout— were provided and incorporated into the Lake Superior basin. This data was only incorporated into the brook trout model. An updated stream temperature model was also included as a new predictor variable for the entire GLB. The full list of potential predictor variables is shown in Appendix A.

A statistical modeling approach—boosted regression trees (BRT)—is employed to relate the instream response variable to the landscape-based predictor variables. BRT models combine decision trees (i.e. classification and regression trees [CART]) and boosting methodologies, which result in better cross-validated models than other methods (Elith et al., 2006[i]), including CART. Decision trees are advantageous because (1) they can incorporate any type of predictor data (binary, numeric, categorical); (2) model outcomes are unaffected by differing scales of predictors; (3) irrelevant predictors are rarely selected; (4) they are insensitive to outliers and non-normalized data; (5) they can accommodate missing predictor data; and (6) they can automatically handle interactions between predictors (Elith et al., 2008). The boosting algorithm used by BRT improves upon the accuracy of a basic CART approach by following the idea that averaging many rough models offers efficiency over finding a single prediction rule that is highly accurate (Elith et al., 2008[ii]).

The modeling process results in a series of quantitative outcomes, including predictions of expected current conditions to all catchments in the FHP, measurement of prediction accuracy, a quantification of each predictor's relative influence on the predictions (i.e., variable importance), and a series of plots illustrating the modeled functional relationship between each predictor and the response (e.g., plot of impervious area vs. presence-absence). The predictions of current conditions are created by extrapolating the BRT model to each catchment within the modeling area. The units of the predicted current condition for this assessment are probability of presence for the fish guild. These current conditions are useful for assessing habitats and mapping the expected range of species.

Predictive accuracy is quantified using an internal cross-validation (CV) method (Elith et al., 2008). The method consists of randomly splitting the input dataset into ten equally-sized subsets, developing a BRT model on a single subset and testing its performance on the remaining nine, and then repeating that process for the remaining nine subsets. Thus, the accuracy measures, such as the CV receiver operating characteristic (ROC) score (for presence-absence responses) or the CV correlation coefficient, are actually averages of ten separate ROC or correlation measurements. A standard error for the ten estimates is also given. CV measures are designed to estimate how well the model will perform using independent data.

1.3.2 *Post modeling*

Characterizing anthropogenic stress and natural habitat quality of aquatic habitats is a useful and necessary process for helping land and fisheries managers identify place-based conservation and restoration strategies. For each of the five models, a post-modeling process was used to characterize anthropogenic stress and natural habitat quality for all catchments within the study area. Stress and natural habitat quality indices and metrics were identified and calculated based on BRT model outputs, and details of those calculations are below.

Once developed, these indices of stress and habitat quality can be used to generate and visualize restoration and protection priorities by analyzing how stress reduction or habitat improvement can increase the probability of presence. For example, areas of high natural quality and low stress could represent protection priorities, whereas areas of high natural quality and high stress may represent restoration priorities.

Anthropogenic stress

Stress indices are critical for evaluating anthropogenic landscape drivers that structure aquatic responses. Managers can use stress indices and metrics to assess how anthropogenic processes are impacting aquatic responses and can utilize this information to cite restoration projects in order to maximize efficiency. Individual stressors were identified by examining BRT model outputs, both the variable influence table and the functional relationship between predictor variables and response variables. Any predictor variable significantly affected by anthropogenic disturbance was included as a potential stressor. Stressors were not utilized for calculation of stress in the model when the functional relationship between a potential stressor and the response variable was not indicative of a mechanistic relationship (e.g. regional trends were overwhelming mechanistic relationships).

Individual stress metrics were calculated by determining the increase in probability of presence for each catchment when stress for that predictor variable was hypothetically removed. A new predictor variable dataset was produced to calculate each individual stressor metric. The new predictor dataset contained the same values as the original predictor dataset except for a single anthropogenic variable for which a stress metric was calculated. For this variable, the values were all hypothetically set to reflect “no stress.” This provided a hypothetical baseline that represented the removal of all stress from that predictor variable. The existing BRT model was then applied to the new hypothetical landscape data to provide an extrapolation of the current model assuming zero stress for that stressor. The difference between the current predicted probability of presence and the probability of presence under this “no stress” situation indicated the change that could be attributable to stress. This process was repeated for each stressor to generate individual metrics of stress on a potential scale of 0-1. Higher stress values indicated a larger change in predicted probability of presence after removing stress, and lower stress values indicated that the catchment was relatively unaffected by removing stress (Table 1).

For each catchment, the individual stress metrics (e.g. agriculture stress, impervious surface stress, etc.) are then summed to produce an overall stress metric, the anthropogenic stress index (ASI). The generalized formula for calculating individual stress metrics and ASI is as follows:

individual stress metric = probability of presence$_{no\ stress}$ *– probability of presence*$_{current}$

anthropogenic stress index (ASI) = individual stress metric 1 + individual stress metric 2 + ….

Table 1: Example of stress calculations

Comid	Current Condition Predictions	Stressor 1 Predictions	Stressor 1 Metric	Stressor 2 Predictions	Stressor 2 Metric	Anthro. Stress Index (ASI)
Catchment ID	**Predictions using current landscape data**	**Predictions when stressor 1 removed**	**(Stressor 1 pred – Current Pred)**	**Predictions when stressor 2 removed**	**(Stressor 2 pred – Current Pred)**	**Stressor 1 Metric + Stressor 2 Metric**
1234567	0.80	0.90	0.10	0.80	0	0.10
1234568	0.25	0.50	0.25	0.35	.10	0.35
1234569	0.5	0.7	0.2	0.55	.05	0.25

Natural habitat quality

Natural habitat quality metrics provide critical baseline information on the optimal potential condition of a catchment. We defined natural quality as the maximum probability of presence under a zero-stress situation; essentially, the highest attainable condition in the catchment. These metrics allow managers to further classify each catchment and target specific land-based conservation or restoration actions.

The natural habitat quality index (HQI) was calculated directly from the BRT output. Metrics for 'natural' predictor variables were calculated using a different approach than the stressor calculations detailed above. Individual natural quality metrics were not seen as useful by the Midwest and Great Plains FHP Science Team since individual habitat variables were not considered practical management targets (e.g., elevation is a relatively fixed value) and therefore were not used in the calculation of HQI. A single hypothetical 'no stress' dataset was created where all stressors were removed. The existing BRT model was then applied to this hypothetical predictor dataset, and the resulting probability of presence indicated the maximum condition attainable by removing all stress. The probability of presence calculated by the BRT model for this hypothetical 'no stress' dataset is the HQI and this value indicates the maximum condition expected in each catchment.

natural habitat quality index (HQI) = probability of presence$_{all\ stressors\ removed}$

1.3.3 *Assessment summary*

These methods provide current predictions of probability of presence, ASI scores, HQI scores, and potential future probability of presence for each of the three models. Metrics and indices were generated at the 1:100k NHD catchment scale and then mapped in GIS.

2. BROOK TROUT

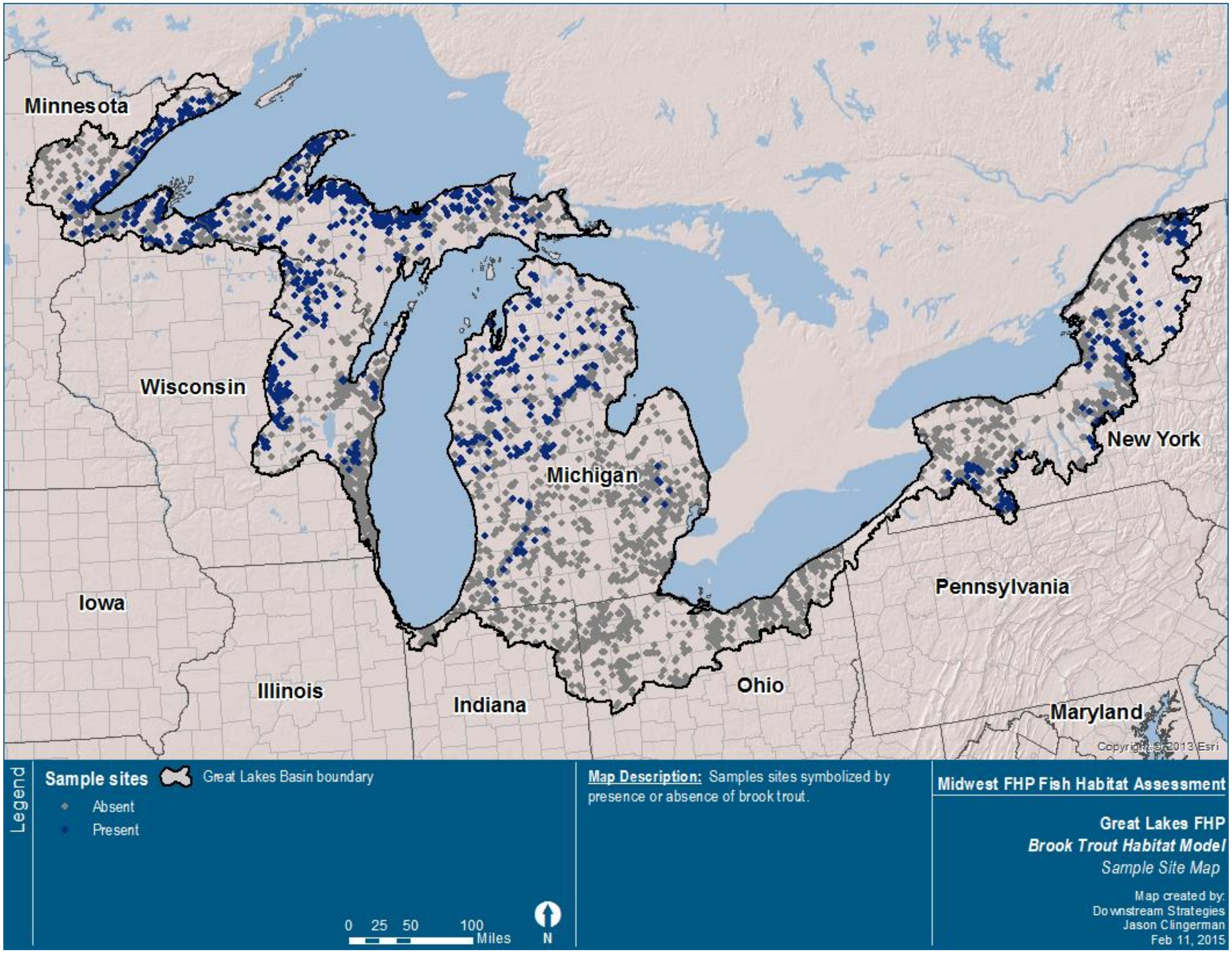
2.1 Modeling inputs

DS used a list of predictor variables selected by GLBFHP to develop a ten-fold CV BRT model for brook trout at the 1:100k catchment scale. The model was used to produce maps of expected brook trout distribution and maps of expected natural habitat quality and anthropogenic stress at the 1:100k scale throughout the extent of the GLBFHP.

DS cooperated with GLBFHP to arrive at a list of landscape-based habitat variables used to predict brook trout throughout the region; those variables were also used to characterizing habitat quality and anthropogenic stress. From an initial suite of 516 catchment attributes, DS and the GLBFHP compiled a list of 72 predictors for evaluation. From that list, 57 variables were removed due to statistical redundancy ($r > 0.6$), logical redundancy, or because of lack of model influence, resulting in a final list of 15 predictor variables for the BRT model and assessment. See Appendix A for a full data dictionary and the metadata document for variable processing notes.

GLBFHP provided DS with fish data collected in streams from 1995 to 2013. This includes data provided to DS during the initial GLB modeling as well as new data from the Lake Superior Basin. DS then processed that data to create a presence-absence dataset for brook trout, which is comprised of 3,696 observations. Figure 3 maps all of the sampling sites that were used to construct the model and the 1:100k catchments used in the model.

Figure 3: Brook trout modeling area and sampling sites



2.2 Modeling process

2.2.1 *Predictive performance*

The final selected model was comprised of 4,550 trees. The model had a CV correlation statistic of 0.640±0.009 and a CV ROC score of 0.899±0.004.

2.2.2 *Variable influence*

The BRT output includes a list of the predictor variables used in the model ordered and scored by their relative importance. The relative importance values are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman and Meulman, 2003). The relative influence score is scaled so that the sum of the scores for all variables is 100, where higher numbers indicate greater influence (Table 2).

Modeled stream temperature was the single most important predictor variable in the model with a relative influence of 32.58%. The next most important predictor was network agricultural land cover with a relative influence of 15.39%. Network agriculture was also the most important anthropogenic stressor.

Table 2: Relative influence of all variables in the final brook trout model

Variable code	Variable description	Relative influence
Merge_temp	Modeled stream temperature	32.58
Lu_agpc	Network agricultural land cover	15.39
Slope	Slope of catchment flowline	8.78
Soil1pc	Network soil hydrologic group A	8.05
Minelevraw	Minimum catchment elevation	6.42
Lu_devpc	Network developed land cover	6.27
Soil3pc	Network soil hydrologic group C	4.03
Precip	Mean annual precipitation	3.46
Cumdrainag	Network drainage area	3.08
Lu_wetpc	Network wetland land cover	3.02
Soil2pc	Network soil hydrologic group B	2.16
Soil4pc	Network soil hydrologic group D	2.05
Geol_maj	Dominant surficial geology texture	1.89
Water_swc	Network surface water consumption	1.50
Cattlec	Network cattle density	1.32

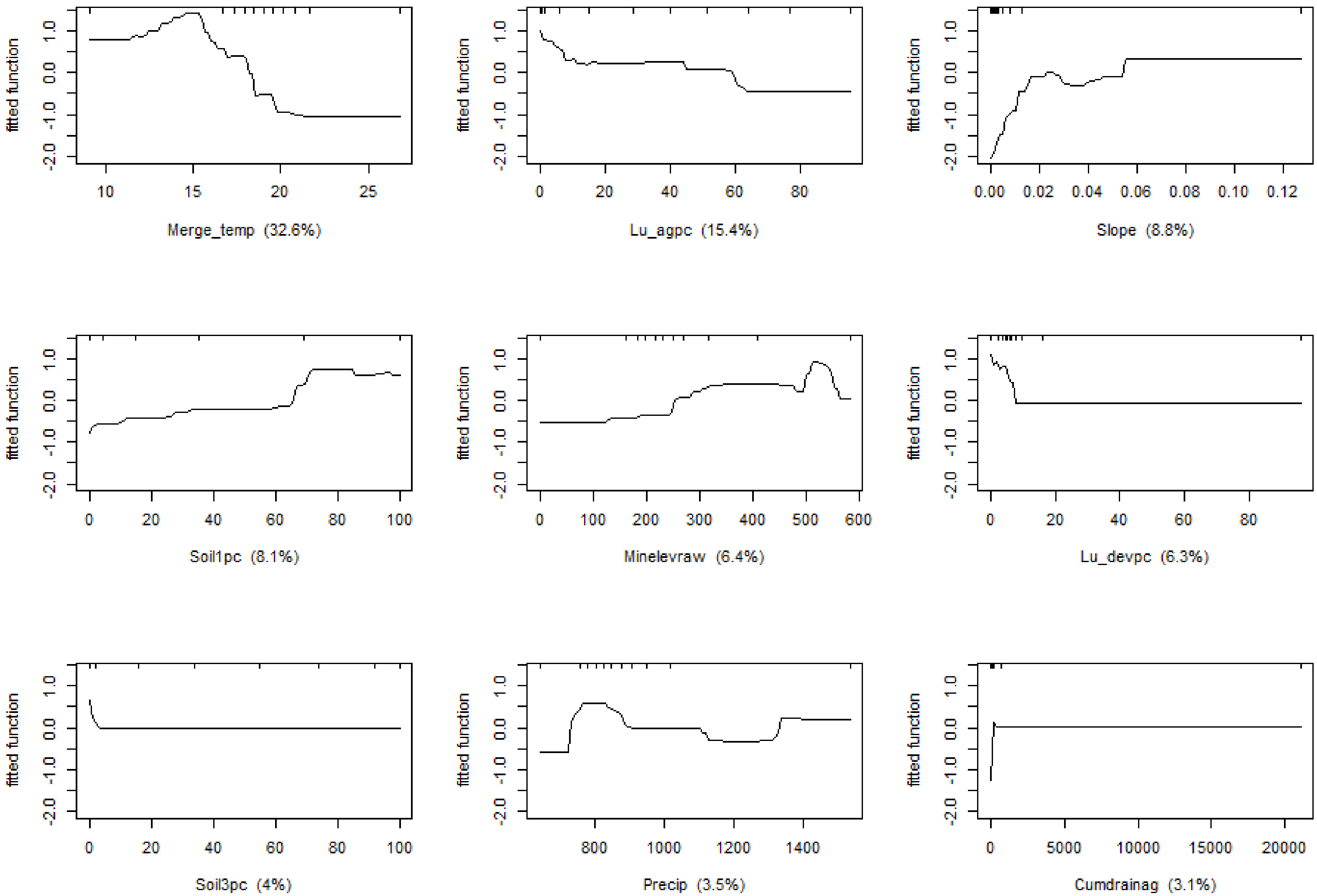
Note: Individual variables are highlighted according to whether they were determined to be anthropogenic (gray shading) or natural (no shading).

2.2.3 *Variable functions*

The BRT output also contains quantitative information on partial dependence functions that can be plotted to visualize the effect of each individual predictor variable on the response after accounting for all other variables in the model. Similar to the interpretation of traditional regression coefficients, the function plots are not always a perfect representation of the relationship for each variable, particularly if interactions are strong or predictors are strongly correlated. However, they do provide a useful and objective basis for interpretation (Friedman, 2001; Friedman and Meulman, 2003).

These plots show the trend of the response variable (y-axis) as the predictor variable (x-axis) changes. The response variable is transformed (usually to the logit scale) so that the magnitude of trends for each predictor variable's function plot can be accurately compared. The dash marks at the top of each function represent the deciles of the data used to build the model. The function plots for the nine most influential variables in the brook trout model (Table 2) are illustrated in Figure 4. The plots for all 15 variables are shown in Appendix B.

Figure 4: Functional responses of the dependent variable to individual predictors of brook trout



Note: Only the top nine predictors, based on relative influence (shown in parentheses; see Appendix A for descriptions of variable codes), are shown here. See Appendix B for plots of remaining predictor variables.

2.3 Post-modeling

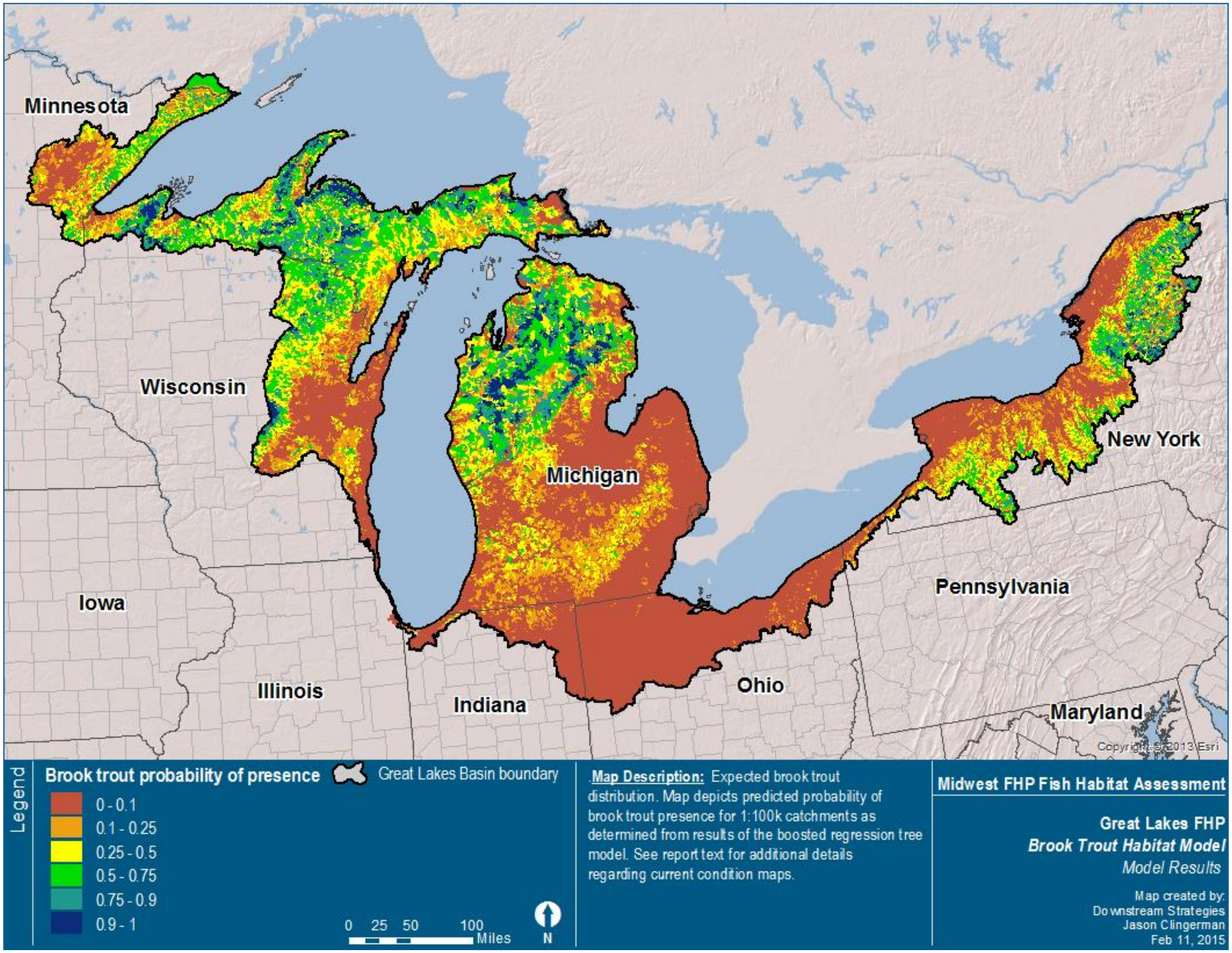
The variable importance table and partial dependence functions of the final BRT model were used to assess the potential stressors for the brook trout model. Within the model, there were five variables considered anthropogenic in nature (Table 2). After reviewing the functional relationships of these four potential stressors, three of the four stressors were removed from ASI calculations. These variables ('Lu_wetpc', 'Water_swc' and 'Cattlec') had function plots that were unintuitive: their relationships to the response likely captured some sort of regional variation in the model rather than a mechanistic relationship with the response. The two remaining stressors, network agriculture land cover (Lu_agpc) and network developed land cover (Lu_devpc), were used to calculate ASI for the brook trout model. Section 1.3.2 details how ASI and HQI were calculated for each model.

2.4 Mapped results

2.4.1 *Expected current conditions*

Brook trout probability of presence was calculated for all 1:100k stream catchments in the study area using the BRT model. The predicted probability values ranged from 0 to 1, where 0 = absent and 1 = 100% probability of presence. The mean predicted probability was 0.209. Of the total 104,343 catchments, there were 7,101 catchments with a predicted probability of presence greater than 0.75 and 9,910 catchments where the probability of presence was between 0.5 and 0.75. These results are mapped in Figure 5.

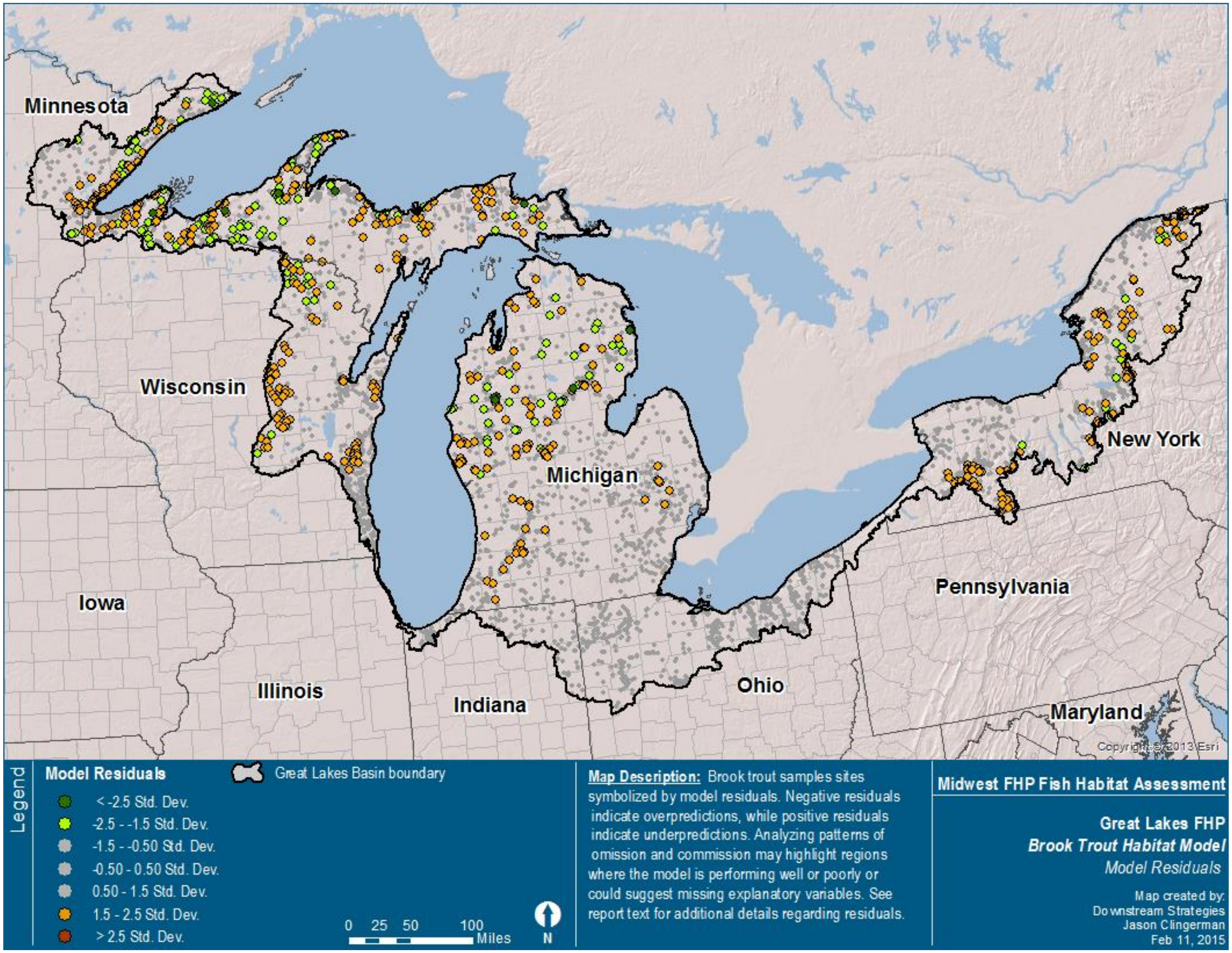
Figure 5: Expected brook trout distribution



2.4.2 *Spatial variability in predictive performance*

Analyzing patterns of omission and commission may highlight regions where the model is performing well or poorly or could suggest missing explanatory variables (Figure 6). To assess omission and commission, residuals are also calculated by the BRT model. The residuals are a measure of the difference in the measured and modeled values (measured value *minus* modeled value). Negative residuals indicate overpredictions (predicting higher values than are true), while positive residuals indicate underpredictions (predicting lower values than are true).

Figure 6: Distribution of brook trout model residuals by sampling site



2.4.3 *Indices of stress and natural quality*

Maps of HQI and ASI illustrate the spatial distribution of natural habitat potential (i.e., HQI score) and anthropogenic stress (i.e., ASI score) in the GLBFHP. HQI and ASI scores are mapped in Figure 7 and Figure 8, respectively. The three metrics contributing toward the calculation of ASI are mapped in Figure 9 and Figure 10. HQI, ASI, and their metrics are all scaled on a 0-1 scale (see Section 2.3 for more details on HQI and ASI calculation). For HQI, higher values indicate higher natural quality, while higher values for ASI indicate higher levels of anthropogenic stress.

At first glance, it may seem that regional stress conditions are overly optimistic, but it is necessary to consider that the stress index is showing areas where probability of presence for this response is reduced because of stressors. It is likely that stress on aquatic systems in general is much more widespread than is indicated in any individual model's stress maps. For all stress and natural quality indices, all catchments are shown, even in areas where the probability of presence is low. This is necessary and useful to consider areas outside of the current expect range where stress could have caused a historic population to be extirpated.

Figure 7: Habitat quality index for brook trout

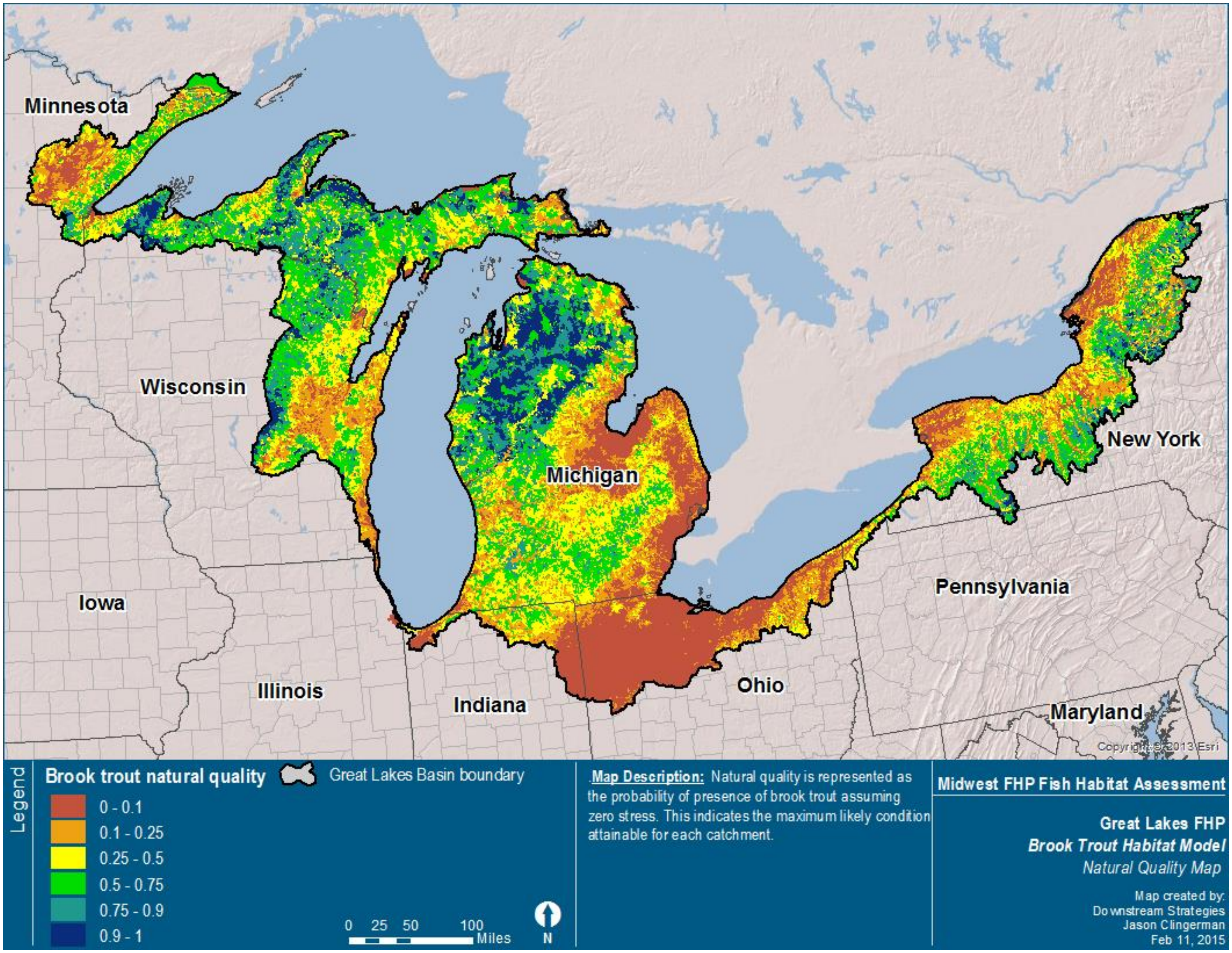


Figure 8: Anthropogenic stress index for brook trout

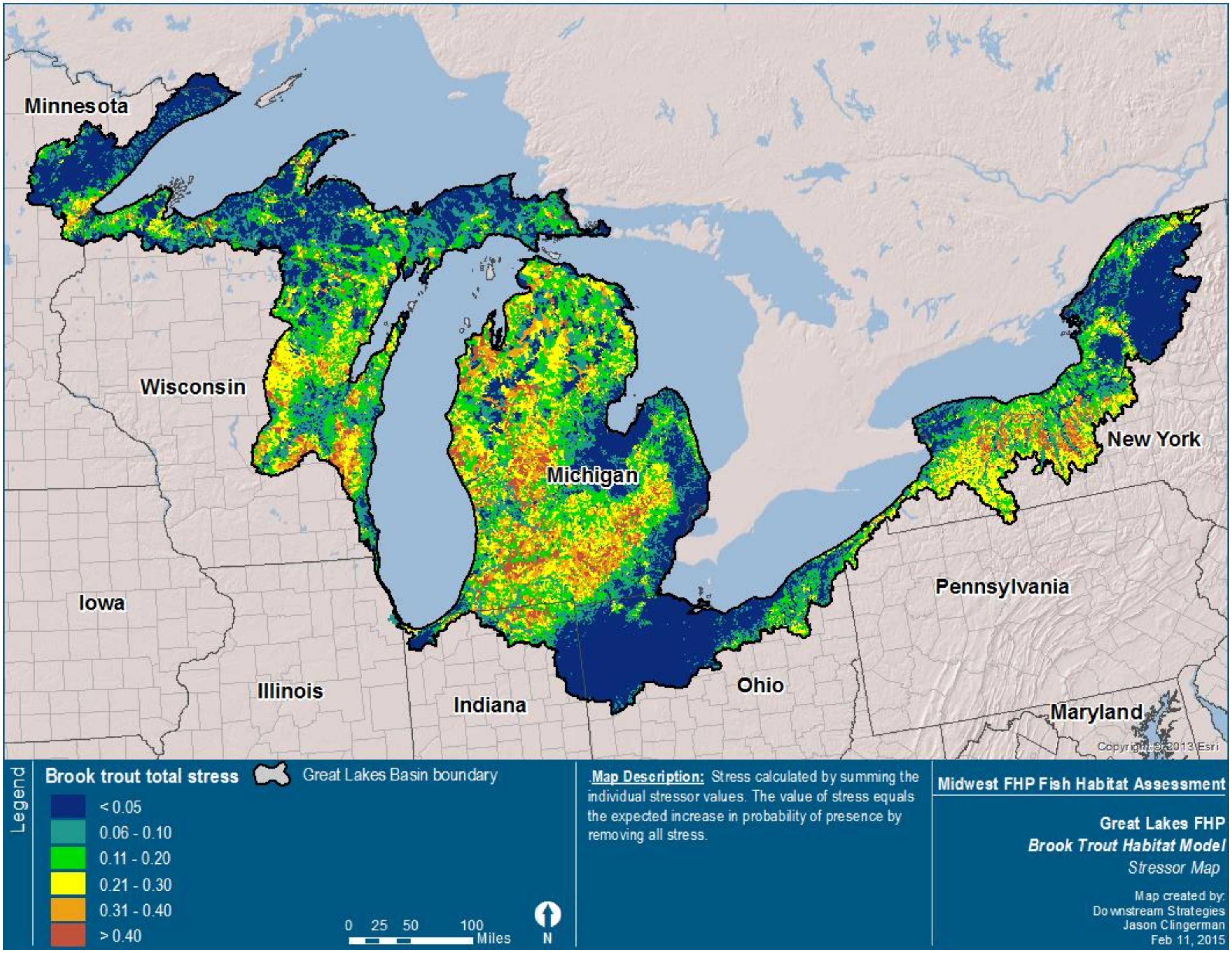


Figure 9: Most influential anthropogenic index metric for brook trout

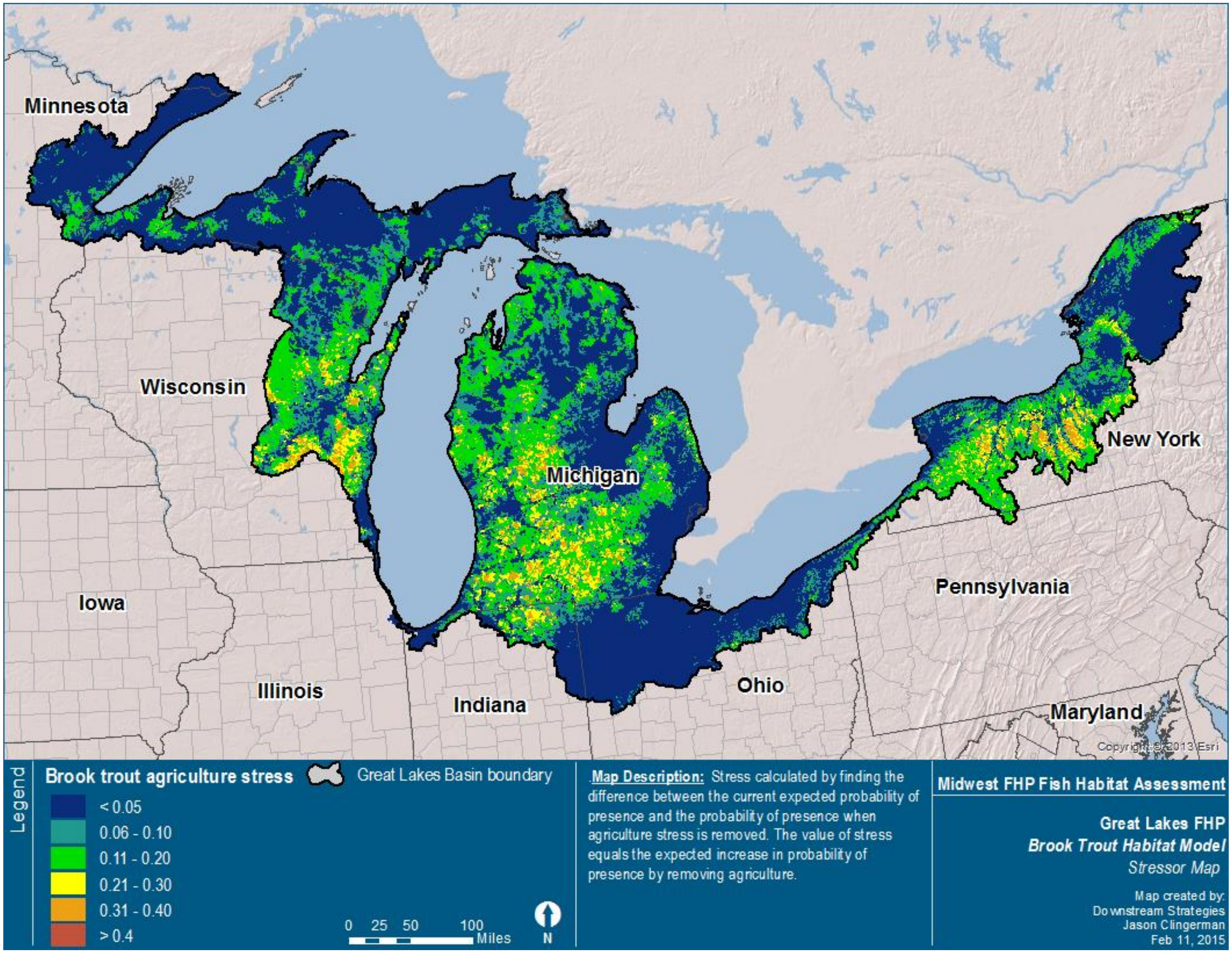
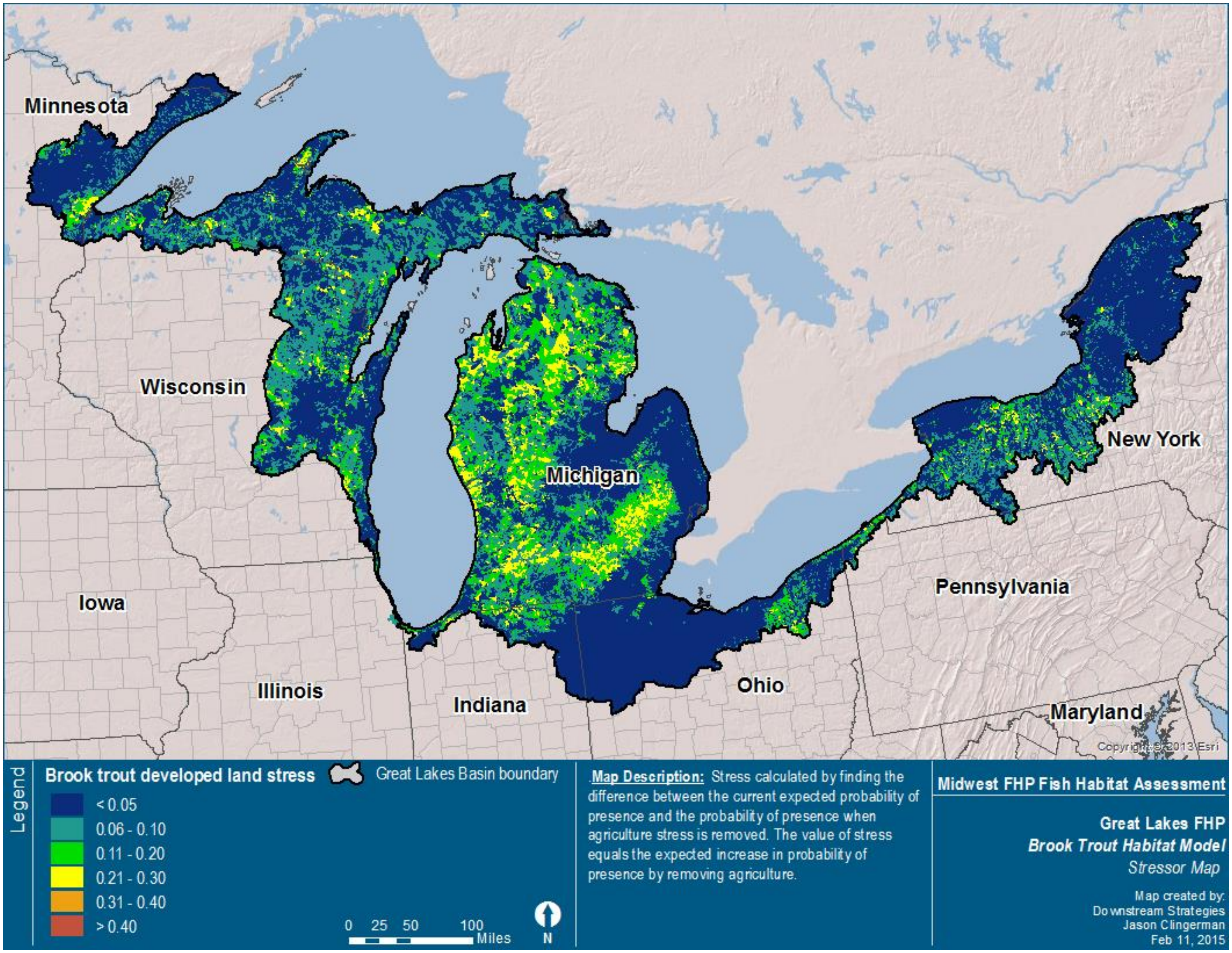


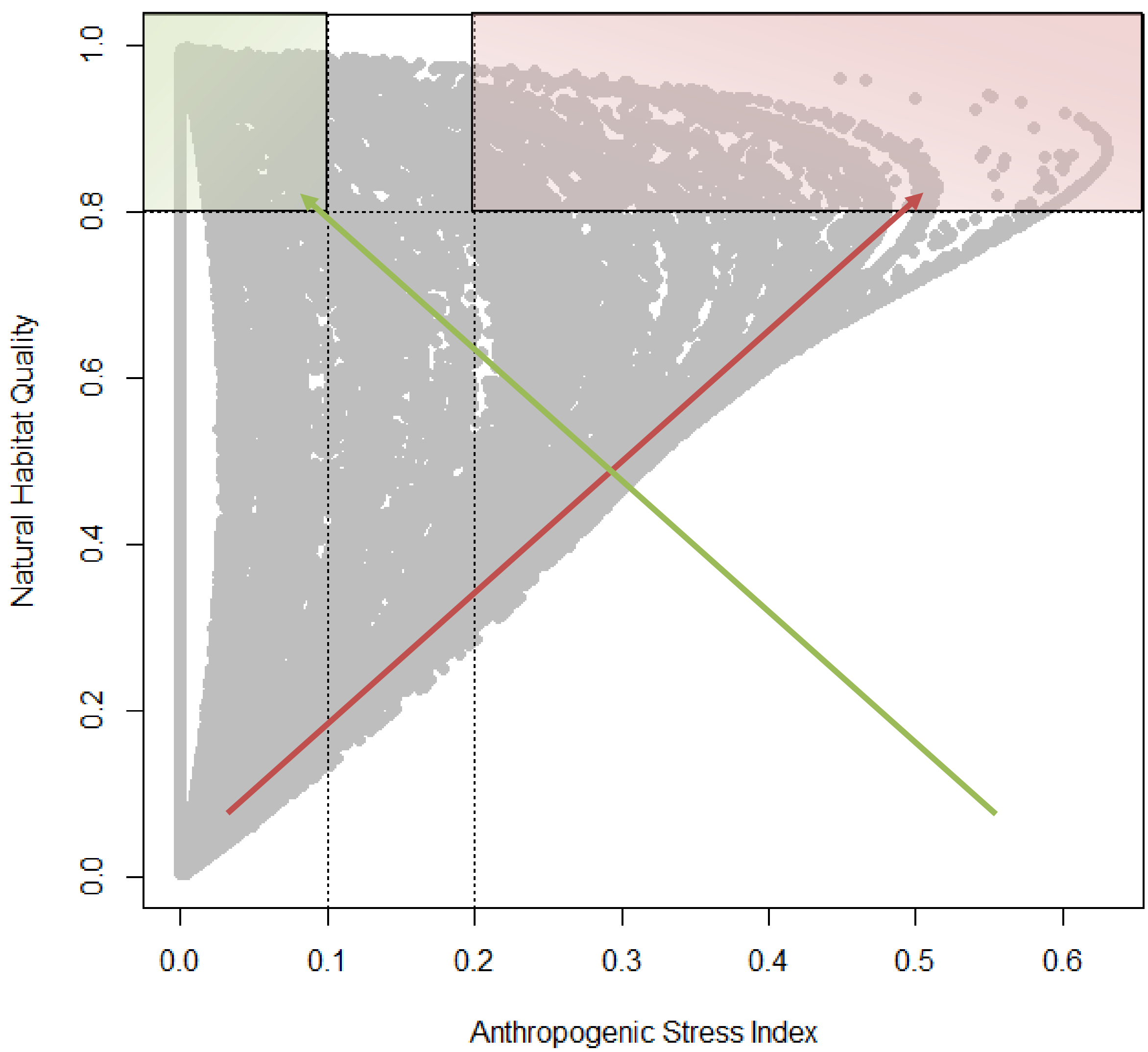
Figure 10: Second most influential anthropogenic index metric for brook trout



2.4.4 *Restoration and protection priorities*

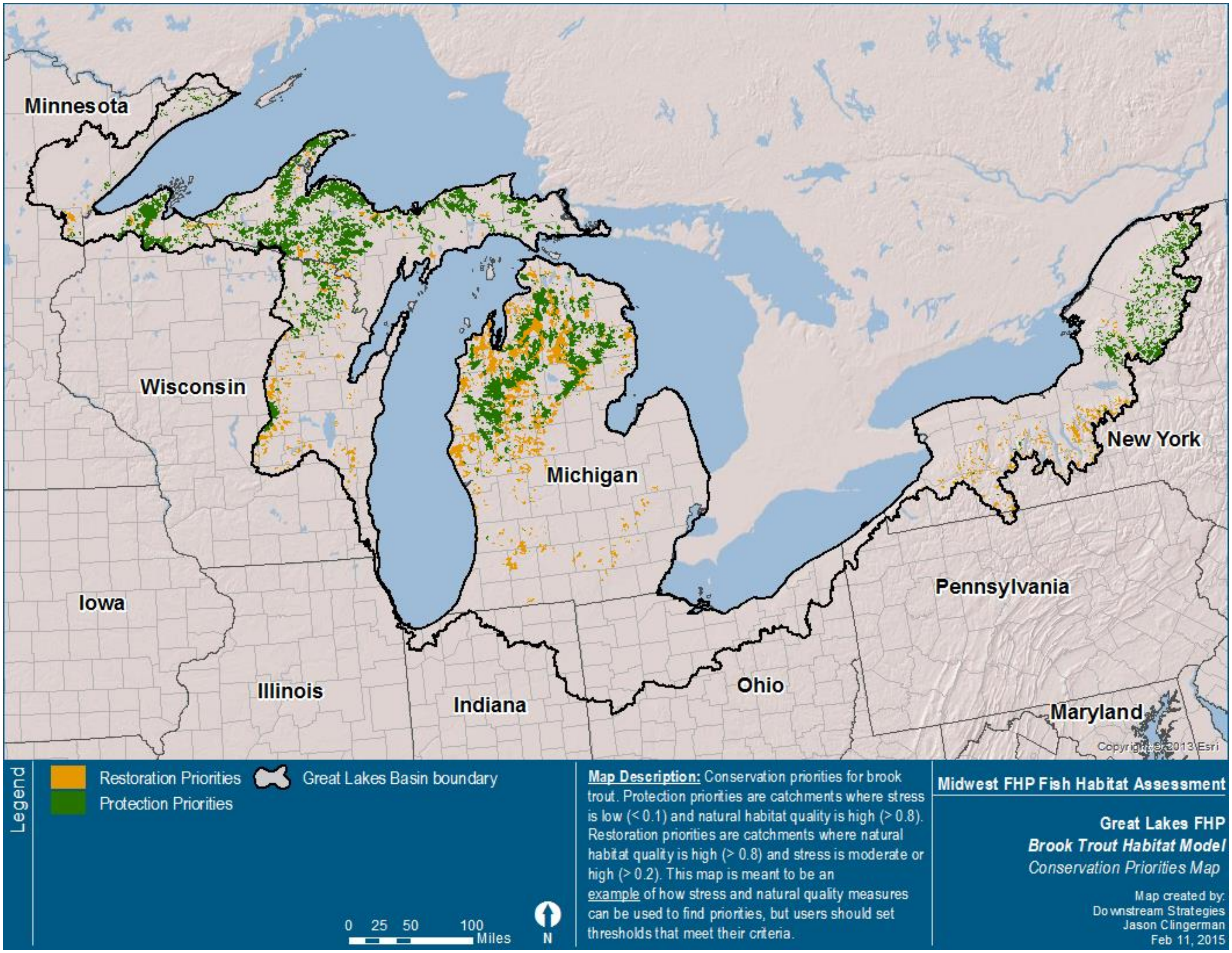
A plot of HQI versus ASI values for all catchments in the study area can be used as a reference to define HQI and ASI thresholds when evaluating restoration and protection priorities (Figure 11). In the example shown, thresholds for protection priorities were defined as catchments with high natural habitat quality and low anthropogenic stress; these thresholds were based on HQI greater than 0.8 and ASI less than 0.1. The thresholds used to identify restoration priorities were defined as catchments with high natural habitat quality and moderate to high anthropogenic stress; these thresholds were based on HQI greater than 0.8 and ASI greater than 0.2. These classifications are mapped in Figure 12. These thresholds were solely based on the relative scores for natural quality and stress indices. Though this example scenario provides an informed set of criteria for identifying conservation priorities, it is only intended to demonstrate the functionality of querying catchments based on these attributes to identify areas that meet user-defined criteria to guide conservation, protection, and restoration planning.

Figure 11: HQI versus ASI values for all catchments for brook trout



Note: Breakpoints for HQI and ASI classes in this example are denoted by dashed lines. The arrows indicate the directions of increasing potential protection (green arrow) or restoration (red arrow) priority. The red box in the upper right corner indicates catchments defined as restoration priorities under the example scenario. The green box in the upper left corner indicates catchments defined as protection priorities under the same scenario.

Figure 12: Restoration and protection priorities for brook trout



3. COLDWATER SPECIES

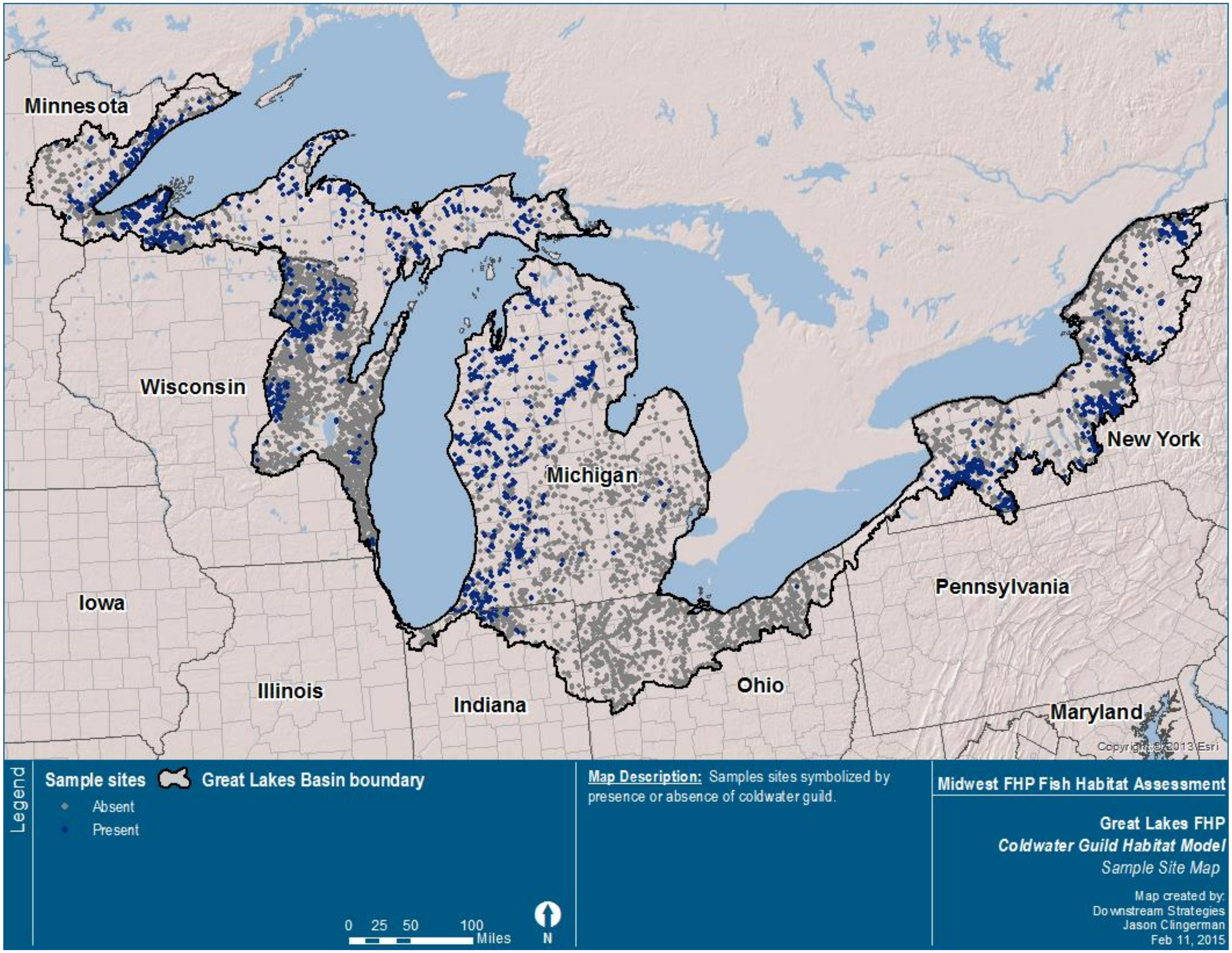
3.1 Modeling inputs

DS used a list of predictor variables selected by GLBFHP to develop a ten-fold CV BRT model for brook trout at the 1:100k catchment scale. The model was used to produce maps of expected brook trout distribution and maps of expected natural habitat quality and anthropogenic stress at the 1:100k scale throughout the extent of the GLBFHP.

DS cooperated with GLBFHP to arrive at a list of landscape-based habitat variables used to predict brook trout throughout the region; those variables were also used to characterizing habitat quality and anthropogenic stress. From an initial suite of 516 catchment attributes, DS and the GLBFHP compiled a list of 72 predictors for evaluation. From that list, 62 variables were removed due to statistical redundancy (r > 0.6), logical redundancy, or because of lack of model influence, resulting in a final list of 10 predictor variables for the BRT model and assessment. See Appendix A for a full data dictionary and the metadata document for variable processing notes.

GLBFHP provided DS with a presence-absence dataset for coldwater species comprised of 9,368 observations collected in streams over a time frame spanning 1995 to 2006. Figure 13 maps all of the sampling sites that were used to construct the model and the 1:100k catchments used in the model.

Figure 13: Coldwater species modeling area and sampling sites



3.2 Modeling process

3.2.1 *Predictive performance*

The final selected model was comprised of 5,050 trees. The model had a CV correlation statistic of 0.706±0.005 and a CV ROC score of 0.932±0.001.

3.2.2 *Variable influence*

The BRT output includes a list of the predictor variables used in the model ordered and scored by their relative importance. The relative importance values are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman and Meulman, 2003). The relative influence score is scaled so that the sum of the scores for all variables is 100, where higher numbers indicate greater influence.

Modeled stream temperature, the single most important variable in terms of relative influence, contributed over 74% of the total influence.

Table 3: Relative influence of all variables in the final coldwater species model

Variable Code	Variable Description	Relative influence
Merge_temp	Modeled stream temperature	74.58
Soil1pc	Network soil hydrologic group A	5.73
Precip	Local mean annual precipitation	4.40
Lu_agpc	Network agriculture land cover	3.37
Slope	Slope of catchment flowline	2.99
Minelevraw	Minimum elevation of catchment	2.48
Lf1pc	Network outwash landform	2.20
Soil4pc	Network soil hydrologic group D	1.62
Lu_wetpc	Network wetland land cover	1.33
Lf10pc	Network attenuated drift landform	1.30

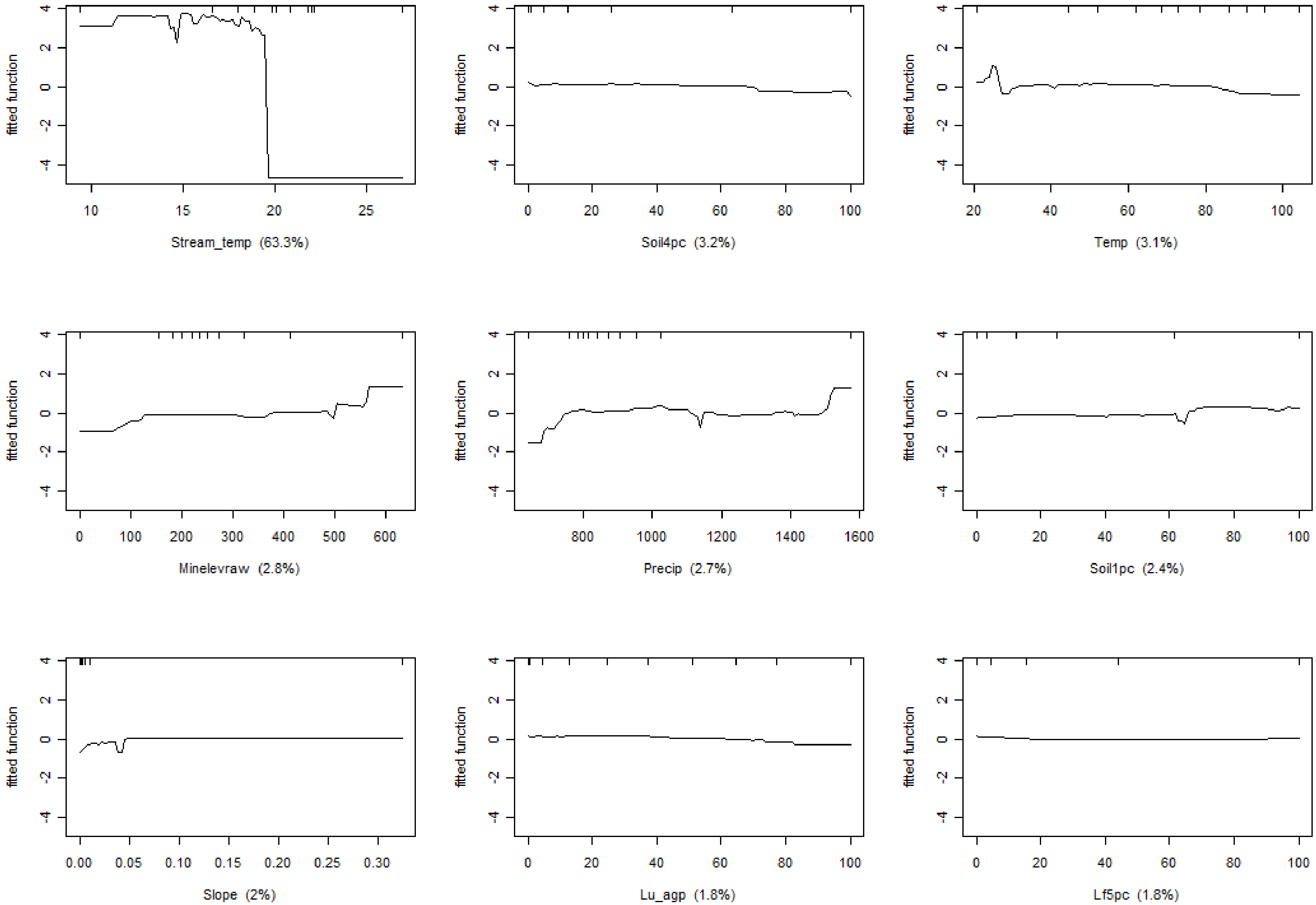
Note: Individual variables are highlighted according to whether they were determined to be anthropogenic in nature (gray shading) or natural (no shading).

3.2.3 *Variable functions*

The BRT output also contains quantitative information on partial dependence functions that can be plotted to visualize the effect of each individual predictor variable on the response after accounting for all other variables in the model. Similar to the interpretation of traditional regression coefficients, the function plots are not always a perfect representation of the relationship for each variable, particularly if interactions are strong or predictors are strongly correlated. However, they do provide a useful and objective basis for interpretation (Friedman, 2001; Friedman and Meulman, 2003).

These plots show the trend of the response variable (y-axis) as the predictor variable (x-axis) changes. The response variable is transformed (usually to the logit scale) so that the magnitude of trends for each predictor variable's function plot can be accurately compared. The dash marks at the top of each function represent the deciles of the data used to build the model. The function plots for the nine most influential variables in the coldwater species model (Table 3) are illustrated in Figure 14. The plots for all 10 variables are shown in Appendix B.

Figure 14: Functional responses of the dependent variable to individual predictors of coldwater species



Note: Only the top nine predictors, based on relative influence (shown in parentheses; see Appendix A for descriptions of variable codes), are shown here. See Appendix B for plots of remaining predictor variables.

3.3 Post-modeling

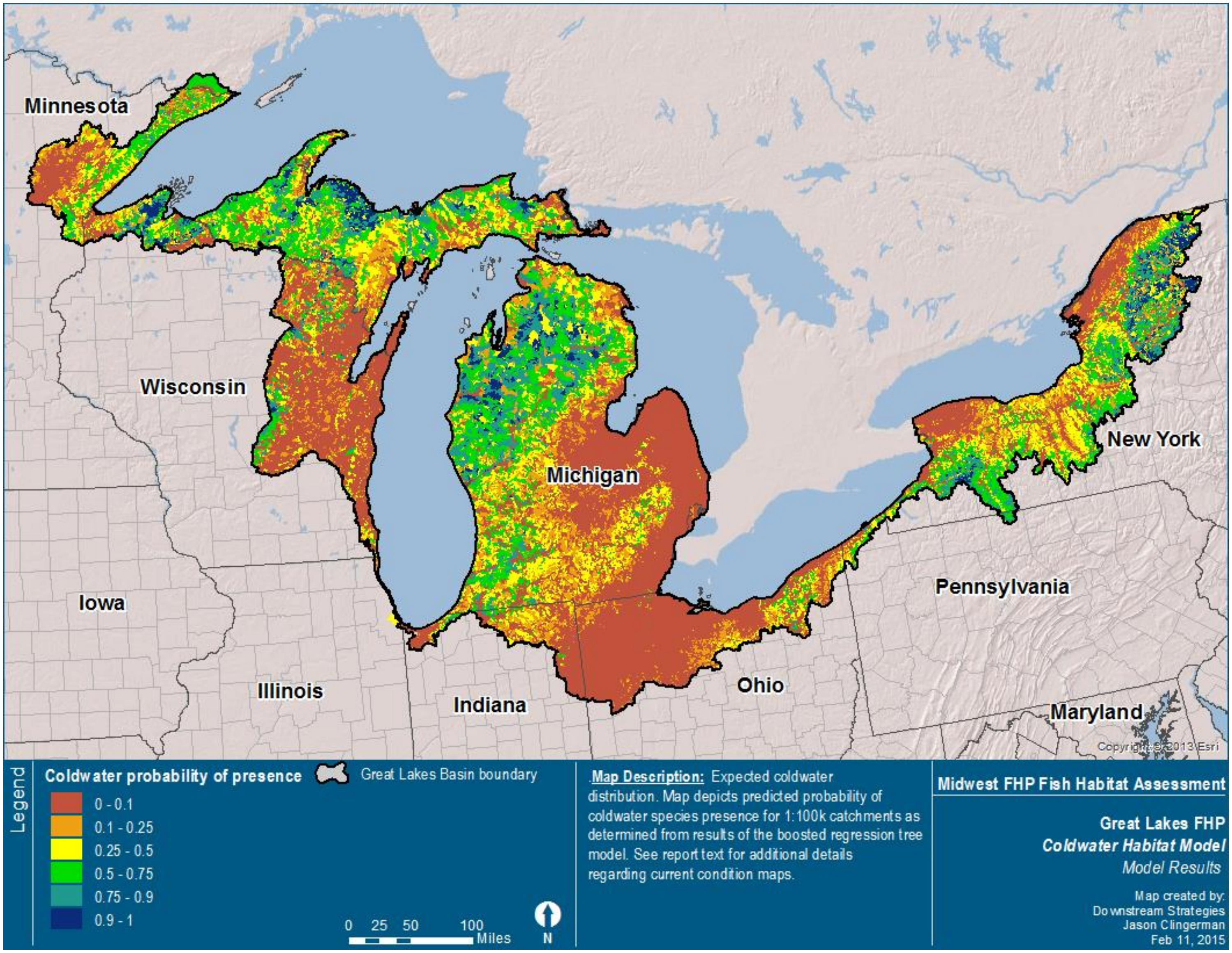
The variable importance table and partial dependence functions of the final BRT model were used to assess the potential stressors for the cold water species model. Within the model, there were two variables considered anthropogenic in nature (Table 3). After reviewing the functional relationships of these two potential stressors, one stressor was removed from ASI calculations. This variable ('Lu_wetpc') had a function plot that was unintuitive: their relationships to the response likely captured some sort of regional variation in the model rather than a mechanistic relationship with the response. The remaining stressor, network agriculture land cover (Lu_agpc) was used to calculate ASI for the coldwater model. Section 1.3.2 details how ASI and HQI were calculated for each model.

3.4 Mapped results

3.4.1 *Expected current conditions*

Coldwater species probability of presence was calculated for all 1:100k stream catchments in the study area using the BRT model. The predicted probability values ranged from 0 to 1. The mean predicted probability was 0.255. Of the total 104,343 catchments, there were 9,706 catchments with a predicted probability of presence greater than 0.75, and 14,213 catchments where the probability of presence was between 0.5 and 0.75. These results are mapped in Figure 15.

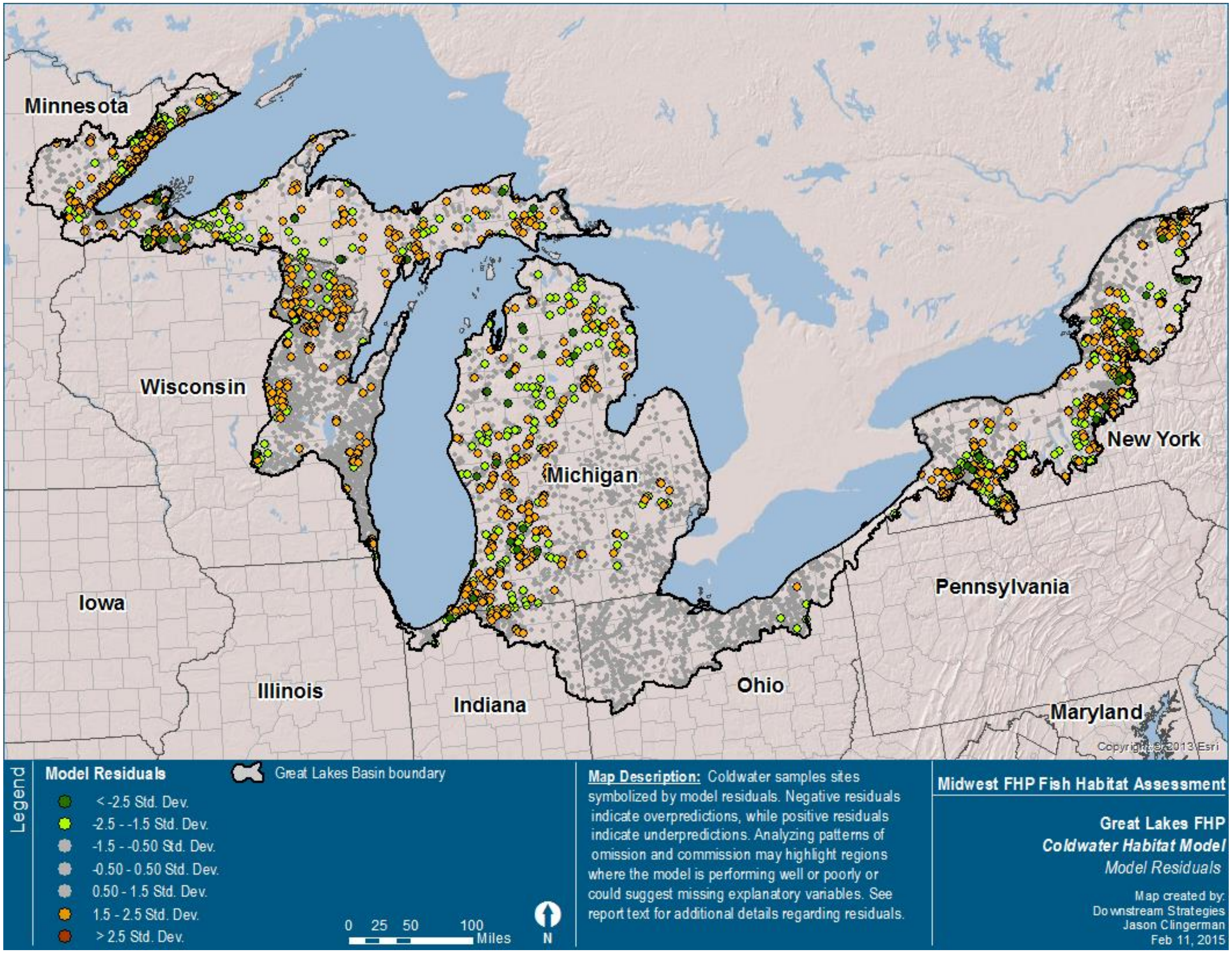
Figure 15: Expected coldwater species distribution



3.4.2 *Spatial variability in predictive performance*

Analyzing patterns of omission and commission may highlight regions where the model is performing well or poorly or could suggest missing explanatory variables (Figure 16). To assess omission and commission, residuals are also calculated by the BRT model. The residuals are a measure of the difference in the measured and modeled values (measured value *minus* modeled value). Negative residuals indicate overpredictions (predicting higher values than are true), while positive residuals indicate underpredictions (predicting lower values than are true).

Figure 16: Distribution of coldwater species model residuals by sampling site



3.4.3 *Indices of stress and natural quality*

Maps of HQI and ASI illustrate the spatial distribution of natural habitat potential (i.e., HQI score) and anthropogenic stress (i.e., ASI score) in the GLBFHP. HQI scores are mapped in Figure 17. The sole contributing variable toward the calculation of ASI, network agricultural land cover, is mapped in Figure 18. HQI and ASI are all scaled on a 0-1 scale (see Section 3.3 for more details on HQI and ASI calculation). For HQI, higher values indicate higher natural quality, while higher values for ASI indicate higher levels of anthropogenic stress.

At first glance, it may seem that regional stress conditions are overly optimistic, but it is necessary to consider that the stress index is showing areas where probability of presence for this response is reduced because of stressors. It is likely that stress on aquatic systems in general is much more widespread than is indicated in any individual model's stress maps. For all stress and natural quality indices, all catchments are shown, even in areas where the probability of presence is low. This is necessary and useful to consider areas outside of the current expect range where stress could have caused a historic population to be extirpated.

Figure 17: Natural quality index for coldwater species

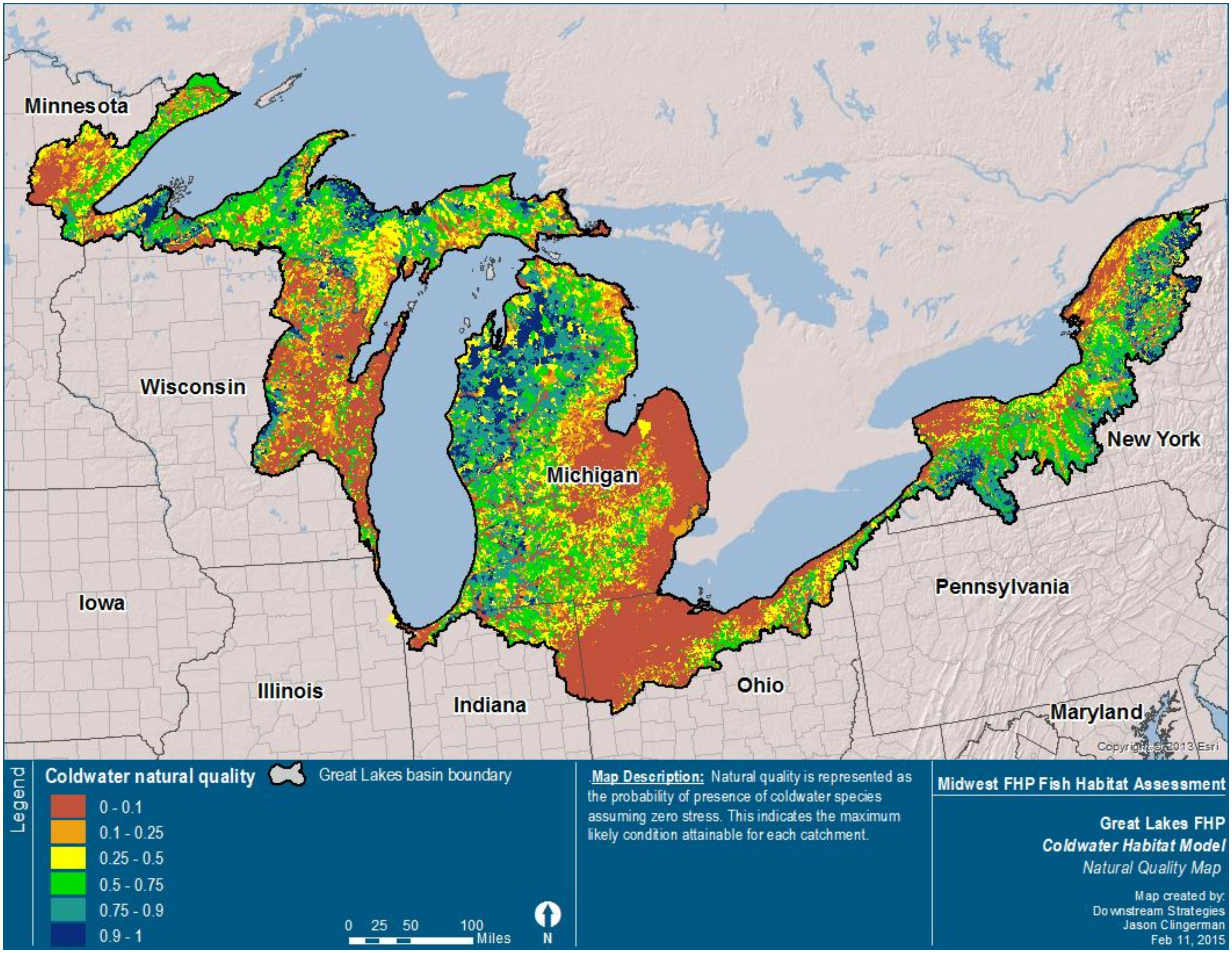
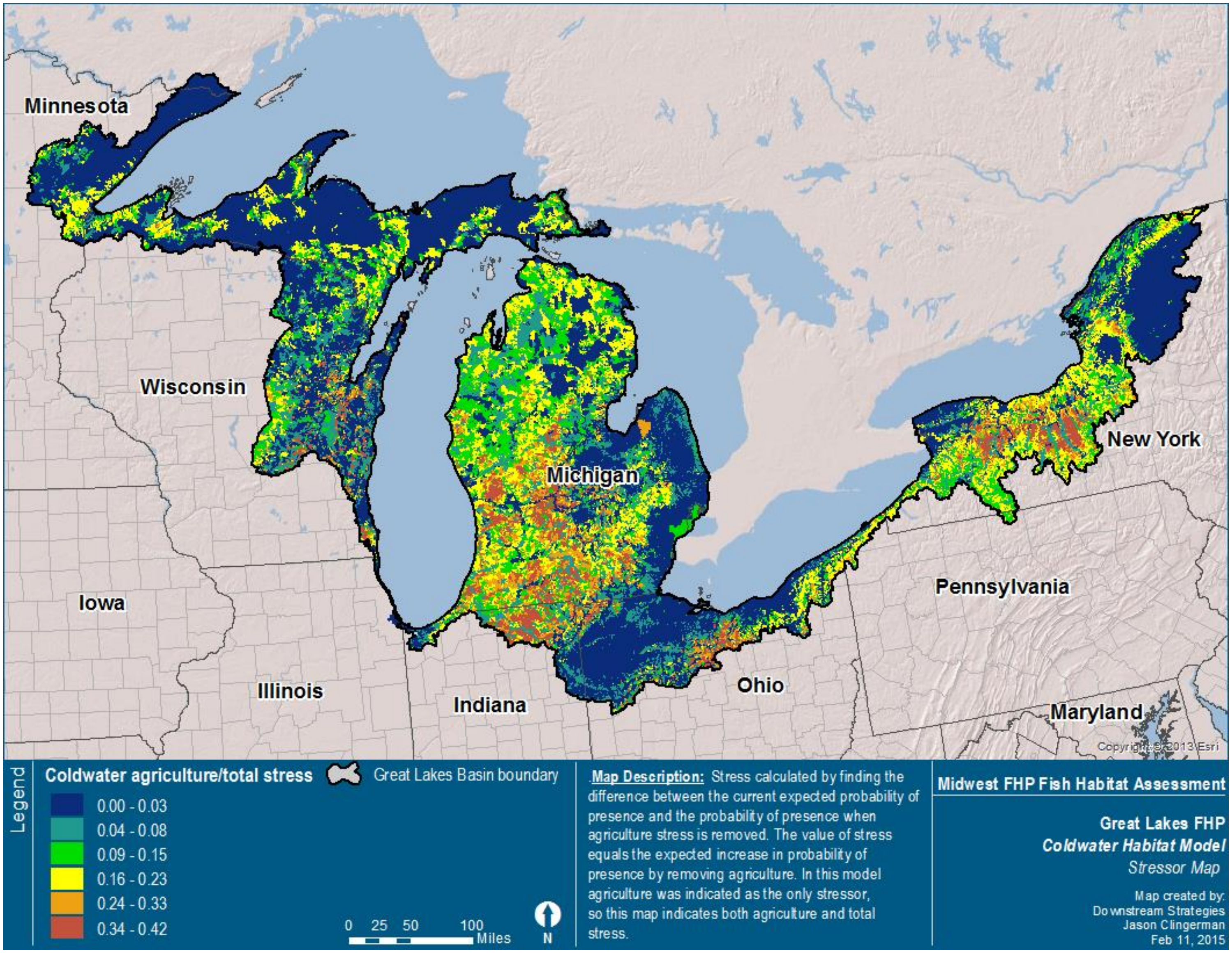


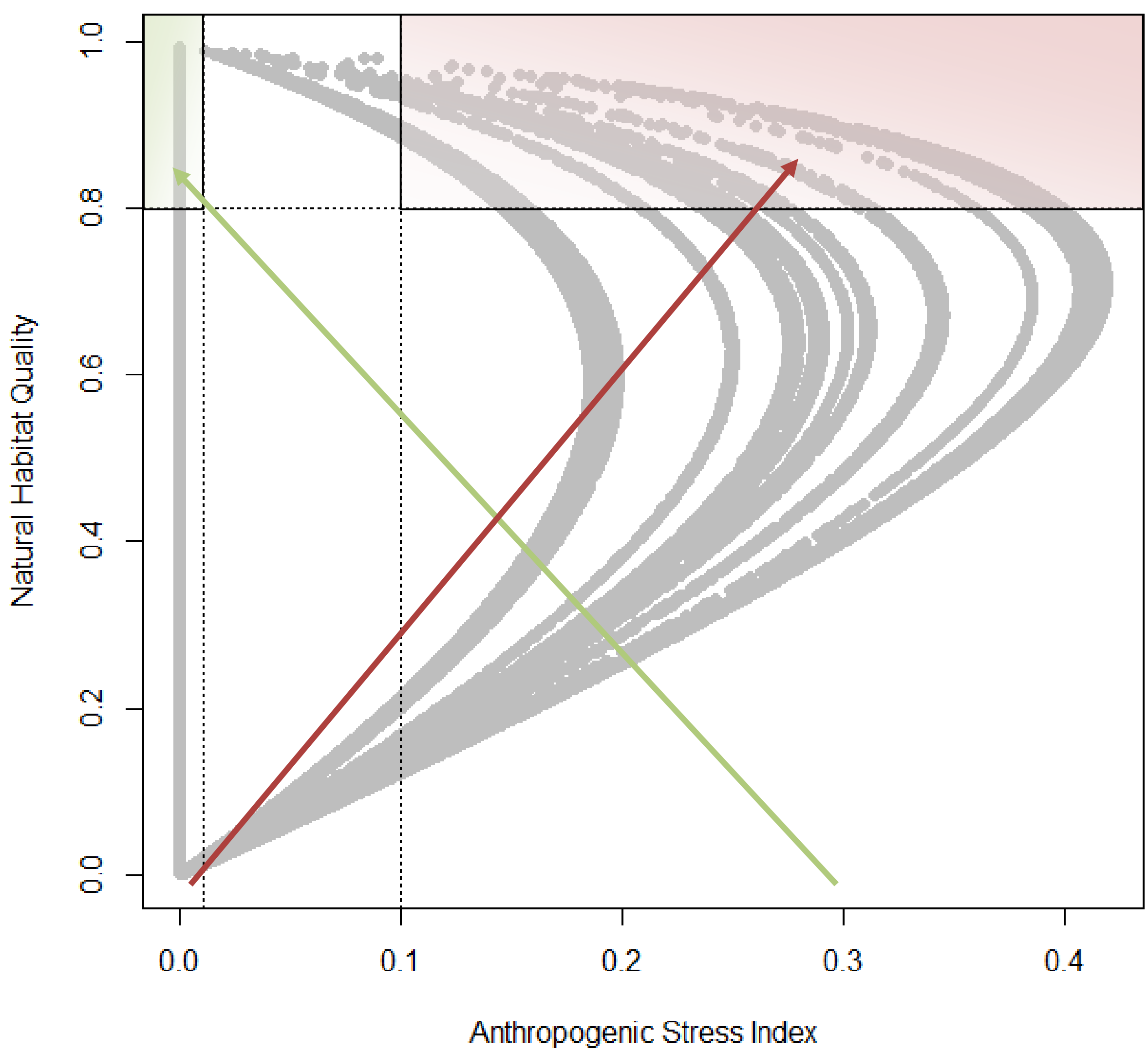
Figure 18: Anthropogenic stress index (total and network agriculture) for coldwater species



3.4.4 *Restoration and protection priorities*

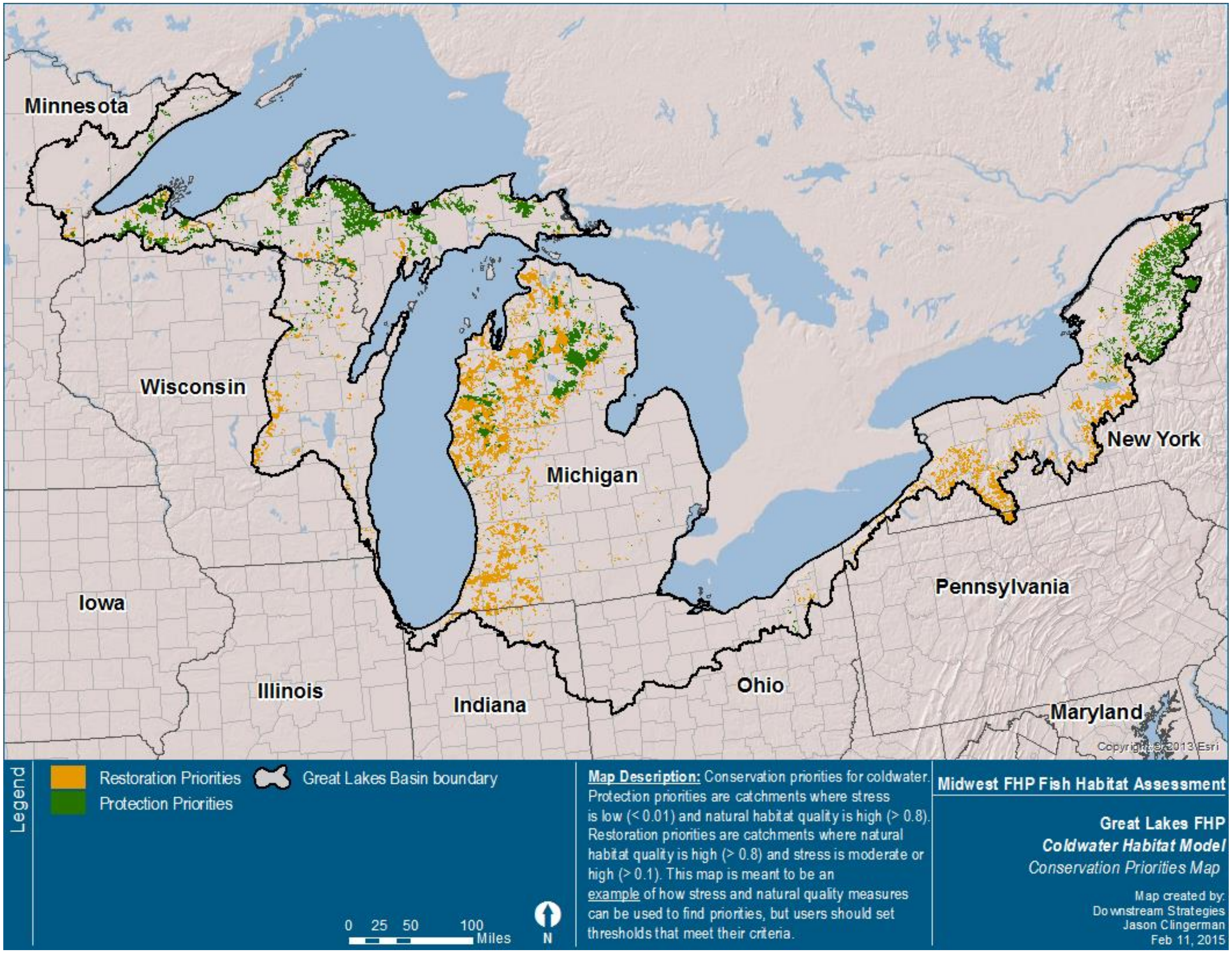
A plot of HQI versus ASI values for all catchments in the study area can be used as a reference to define HQI and ASI thresholds when evaluating restoration and protection priorities (Figure 19). In the example shown, thresholds for protection priorities were defined as catchments with high natural habitat quality and low anthropogenic stress; these thresholds were based on HQI greater than 0.8 and ASI less than 0.01. The thresholds used to identify restoration priorities were defined as catchments with high natural habitat quality and moderate to high anthropogenic stress; these thresholds were based on HQI greater than 0.8 and ASI greater than 0.1. These classifications are mapped in Figure 20. These thresholds were solely based on the relative scores for natural quality and stress indices. Though this example scenario provides an informed set of criteria for identifying conservation priorities, it is only intended to demonstrate the functionality of querying catchments based on these attributes to identify areas that meet user-defined criteria to guide conservation, protection, and restoration planning.

Figure 19: HQI versus ASI values for all catchments for coldwater species



Note: Breakpoints for HQI and ASI classes in this example are denoted by dashed lines. The arrows indicate the directions of increasing potential protection (green arrow) or restoration (red arrow) priority. The red box in the upper right corner indicates catchments defined as restoration priorities under the example scenario. The green box in the upper left corner indicates catchments defined as protection priorities under the same scenario.

Figure 20: Restoration and protection priorities for coldwater species



4. LITHOPHILIC SPECIES RICHNESS

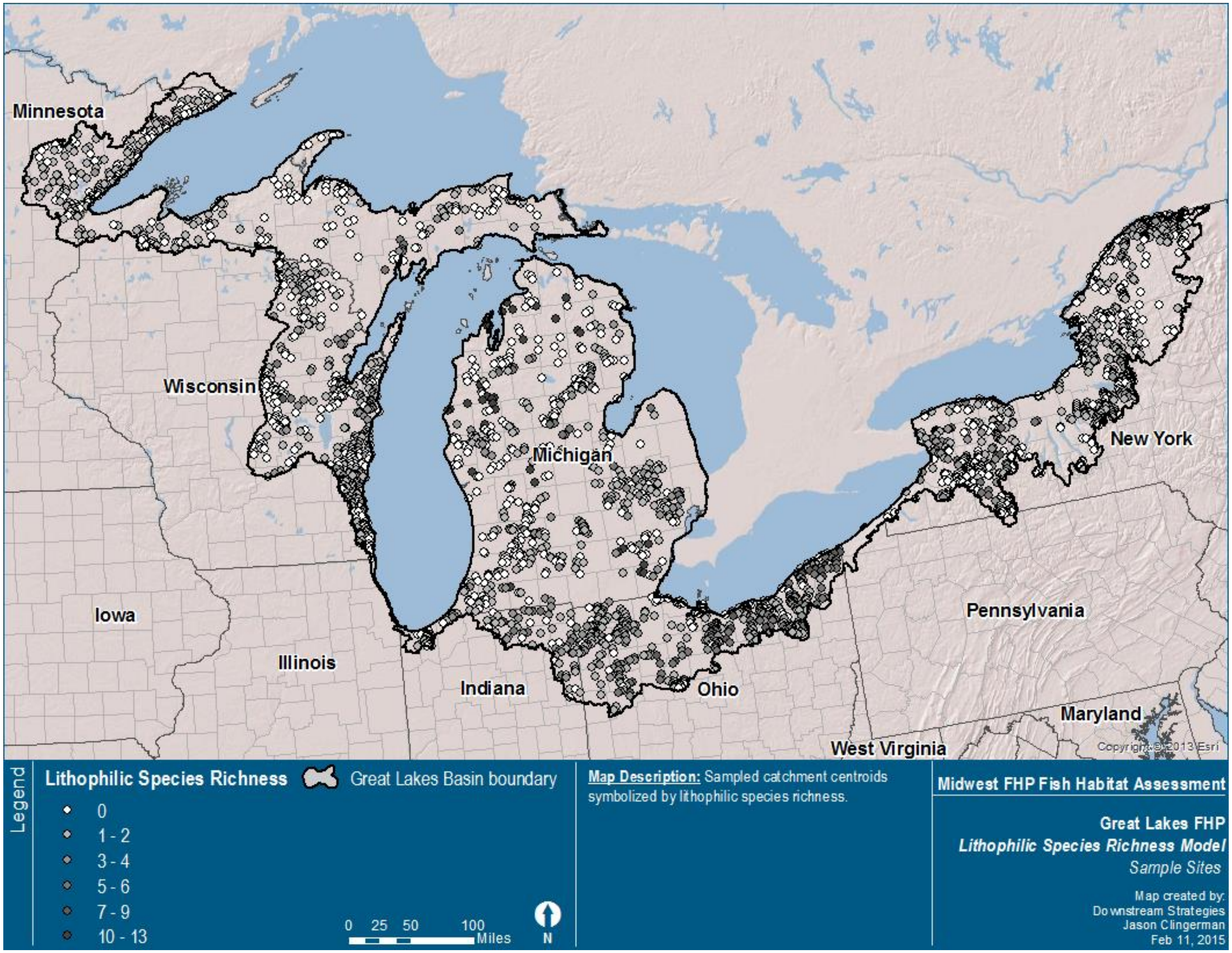
4.1 Modeling inputs

DS used a list of predictor variables selected by GLBFHP to develop a ten-fold CV BRT model for brook trout at the 1:100k catchment scale. The model was used to produce maps of expected brook trout distribution and maps of expected natural habitat quality and anthropogenic stress at the 1:100k scale throughout the extent of the GLBFHP.

DS cooperated with GLBFHP to arrive at a list of landscape-based habitat variables used to predict brook trout throughout the region; those variables were also used to characterizing habitat quality and anthropogenic stress. From an initial suite of 516 catchment attributes, DS and the GLBFHP compiled a list of 72 predictors for evaluation. From that list, 57 variables were removed due to statistical redundancy ($r > 0.6$), logical redundancy, or because of lack of model influence, resulting in a final list of 15 predictor variables for the BRT model and assessment. See Appendix A for a full data dictionary and the metadata document for variable processing notes.

GLBFHP provided DS with fish data collected in streams over a time frame spanning 1995 to 2006. Using that data, DS created a dataset for lithophilic species richness comprised of 2,762 sample points. Figure 21 maps all of the sampling sites that were used to construct the model and the 1:100k catchments used in the model.

Figure 21: Lithophilic species richness modeling area and sampling sites



4.2 Modeling process

4.2.1 *Predictive performance*

The final selected model was comprised of 3,800 trees. The model had a CV correlation statistic of 0.503±0.016.

4.2.2 *Variable influence*

The BRT output includes a list of the predictor variables used in the model ordered and scored by their relative importance. The relative importance values are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman and Meulman, 2003). The relative influence score is scaled so that the sum of the scores for all variables is 100, where higher numbers indicate greater influence.

Network drainage area and modeled stream temperature combined to contribute over 47% of the total influence.

Table 4: Relative influence of all variables in the final lithophilic species richness model

Variable	Variable Description	Relative Influence
Cumdrainag	Network drainage area	25.01
Merge_temp	Modeled stream temperature	22.40
Lu_agpc	Network agricultural land cover	8.78
Soil1pc	Network soil hydrologic group A	7.27
Minelevraw	Minimum catchment elevation	6.51
Dam_countc_den	Network dam density	6.45
Soil4pc	Network soil hydrologic group D	4.39
Lu_devpc	Network developed land cover	3.78
Precip	Mean annual precipitation	3.48
Slope	Slope of catchment flowline	3.34
Roadcr_den	Catchment road crossing density	2.68
Roadlen_den	Catchment road density	1.98
Lu_wetpc	Network wetland land cover	1.79
Lf3pc	Network high density residential area	1.32
Lf11pc	Network mixed forest area	0.82

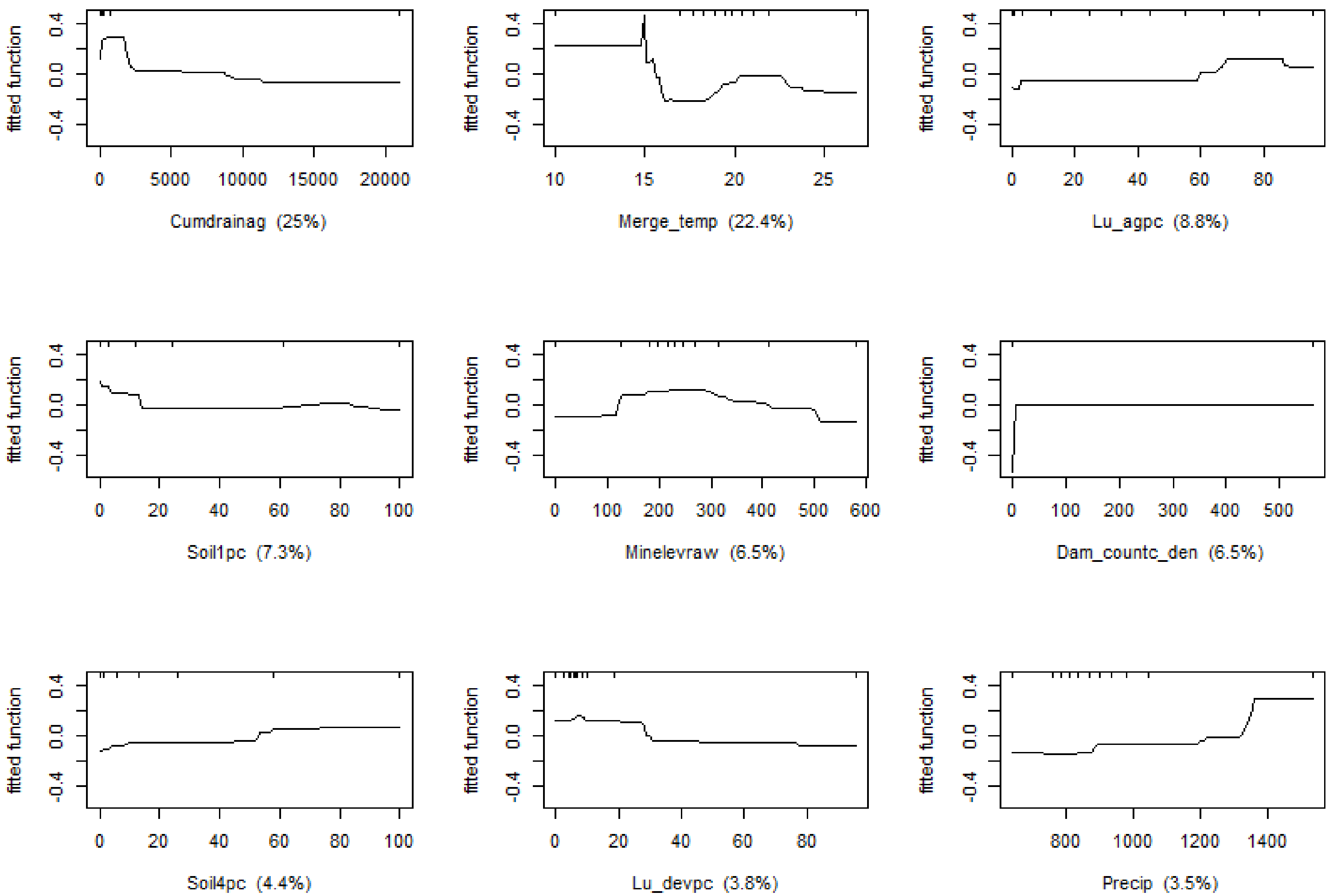
Note: Individual variables are highlighted according to whether they were determined to be anthropogenic in nature (gray shading) or natural (no shading).

4.2.3 *Variable functions*

The BRT output also contains quantitative information on partial dependence functions that can be plotted to visualize the effect of each individual predictor variable on the response after accounting for all other variables in the model. Similar to the interpretation of traditional regression coefficients, the function plots are not always a perfect representation of the relationship for each variable, particularly if interactions are strong or predictors are strongly correlated. However, they do provide a useful and objective basis for interpretation (Friedman, 2001; Friedman and Meulman, 2003).

These plots show the trend of the response variable (y-axis) as the predictor variable (x-axis) changes. The response variable is transformed (usually to the logit scale) so that the magnitude of trends for each predictor variable's function plot can be accurately compared. The dash marks at the top of each function represent the deciles of the data used to build the model. The function plots for the nine most influential variables in the lithophilic species richness model (Table 4) are illustrated in Figure 22. The plots for all 15 variables are shown in Appendix B.

Figure 22: Functional responses of the dependent variable to individual predictors of lithophilic species richness



Note: Only the top nine predictors, based on relative influence (shown in parentheses; see Appendix A for descriptions of variable codes), are shown here. See Appendix B for plots of remaining predictor variables.

4.3 Post-modeling

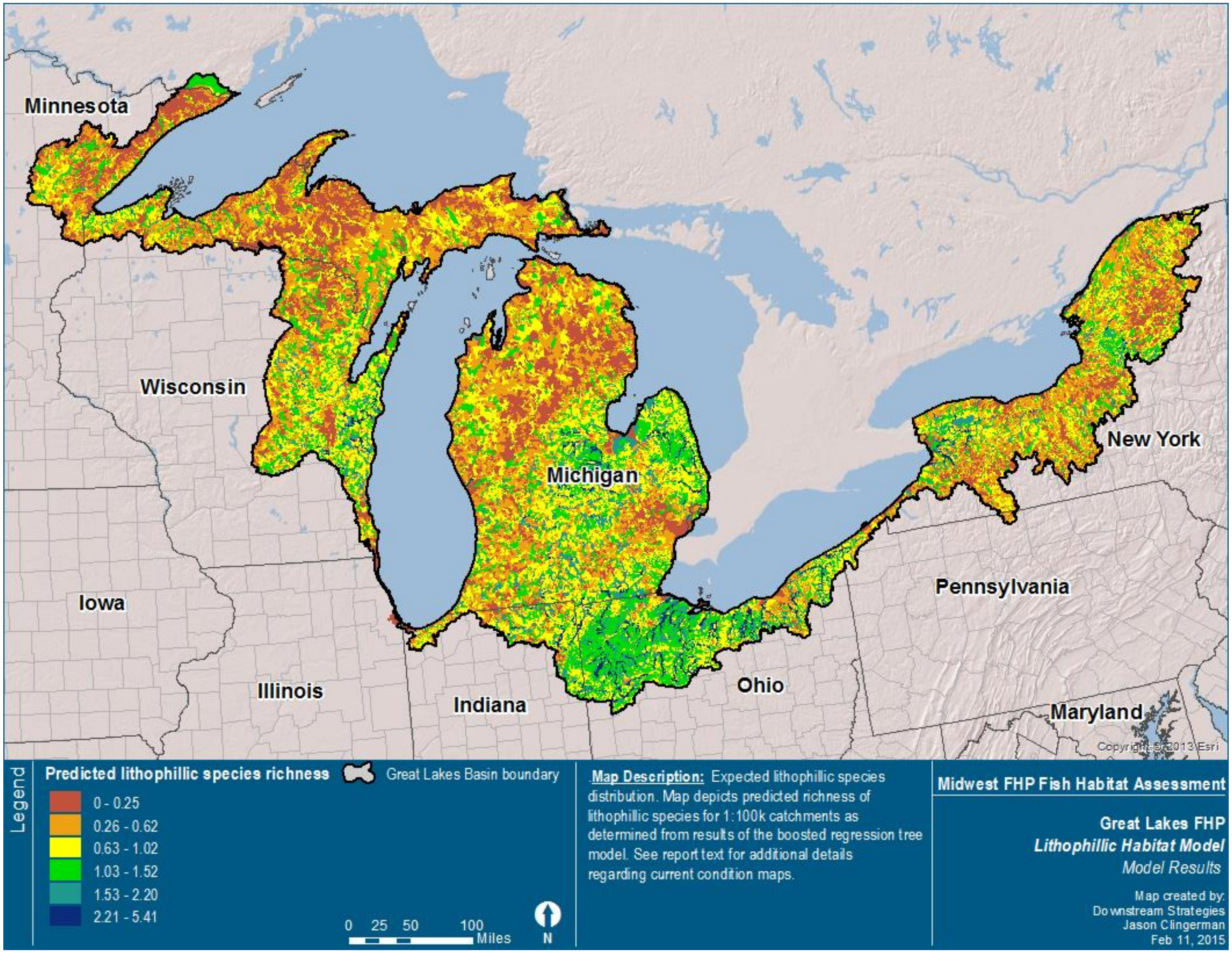
The variable importance table and partial dependence functions of the final BRT model were used to assess the potential stressors for the lithophilic species model. Within the model, there were nine variables considered anthropogenic in nature (Table 3). After reviewing the functional relationships of these potential stressors, only two stressors were retained for ASI calculations ('Lu_devpc' and 'Roadlen_den'). The removed variables had function plots that were unintuitive: their relationships to the response likely captured some sort of regional variation in the model rather than a mechanistic relationship with the response. Section 1.3.2 details how ASI and HQI were calculated for each model.

4.4 Mapped results

4.4.1 *Expected current conditions*

Predicted lithophilic species richness was calculated for all 1:100k stream catchments in the study area using the BRT model. The predicted values ranged from 0 to 5.41. The mean value of predictions was 0.73. These results are mapped in Figure 23.

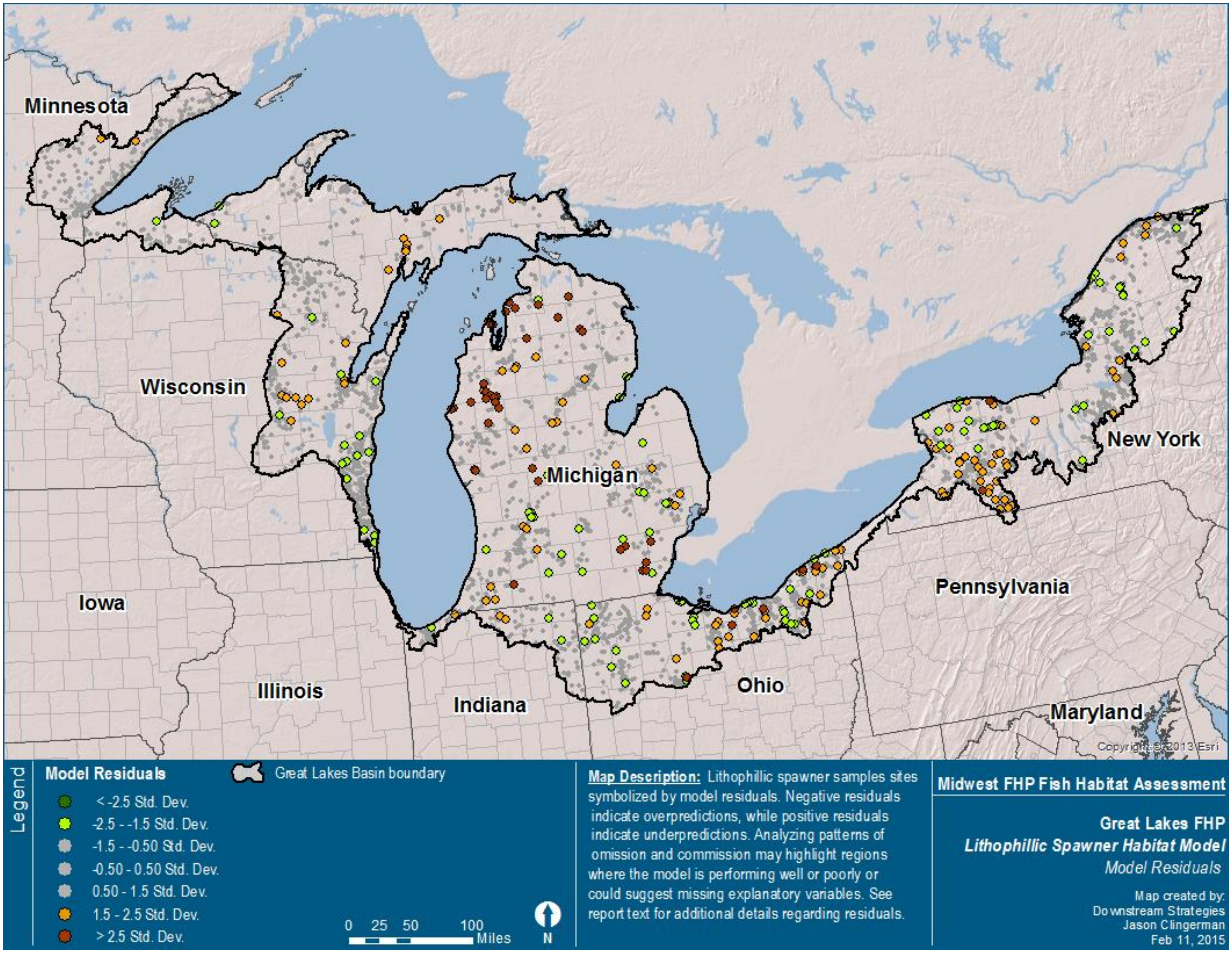
Figure 23: Expected lithophilic species richness distribution



4.4.2 *Spatial variability in predictive performance*

Analyzing patterns of omission and commission may highlight regions where the model is performing well or poorly or could suggest missing explanatory variables (Figure 24). To assess omission and commission, residuals are also calculated by the BRT model. The residuals are a measure of the difference in the measured and modeled values (measured value *minus* modeled value). Negative residuals indicate overpredictions (predicting higher values than are true), while positive residuals indicate underpredictions (predicting lower values than are true).

Figure 24: Distribution of lithophilic species richness model residuals by sampling site



4.4.3 *Indices of stress and natural quality*

Maps of HQI and ASI illustrate the spatial distribution of natural habitat potential (i.e., HQI score) and anthropogenic stress (i.e., ASI score) in the GLBFHP. HQI and ASI scores are mapped in Figure 25 and Figure 26, respectively. The two metrics contributing toward the calculation of ASI are mapped in Figure 27 andFigure 28. HQI, ASI, and their metrics are all scaled on the same scale as the predicted values (see Section 1.3.2 for more details on HQI and ASI calculation). For HQI, higher values indicate higher natural quality, while higher values for ASI indicate higher levels of anthropogenic stress.

At first glance, it may seem that regional stress conditions are overly optimistic, but it is necessary to consider that the stress index is showing areas where probability of presence for this response is reduced because of stressors. It is likely that stress on aquatic systems in general is much more widespread than is indicated in any individual model's stress maps. For all stress and natural quality indices, all catchments are shown, even in areas where the probability of presence is low. This is necessary and useful to consider areas outside of the current expect range where stress could have caused a historic population to be extirpated.

Figure 25: Natural quality index for lithophilic species richness

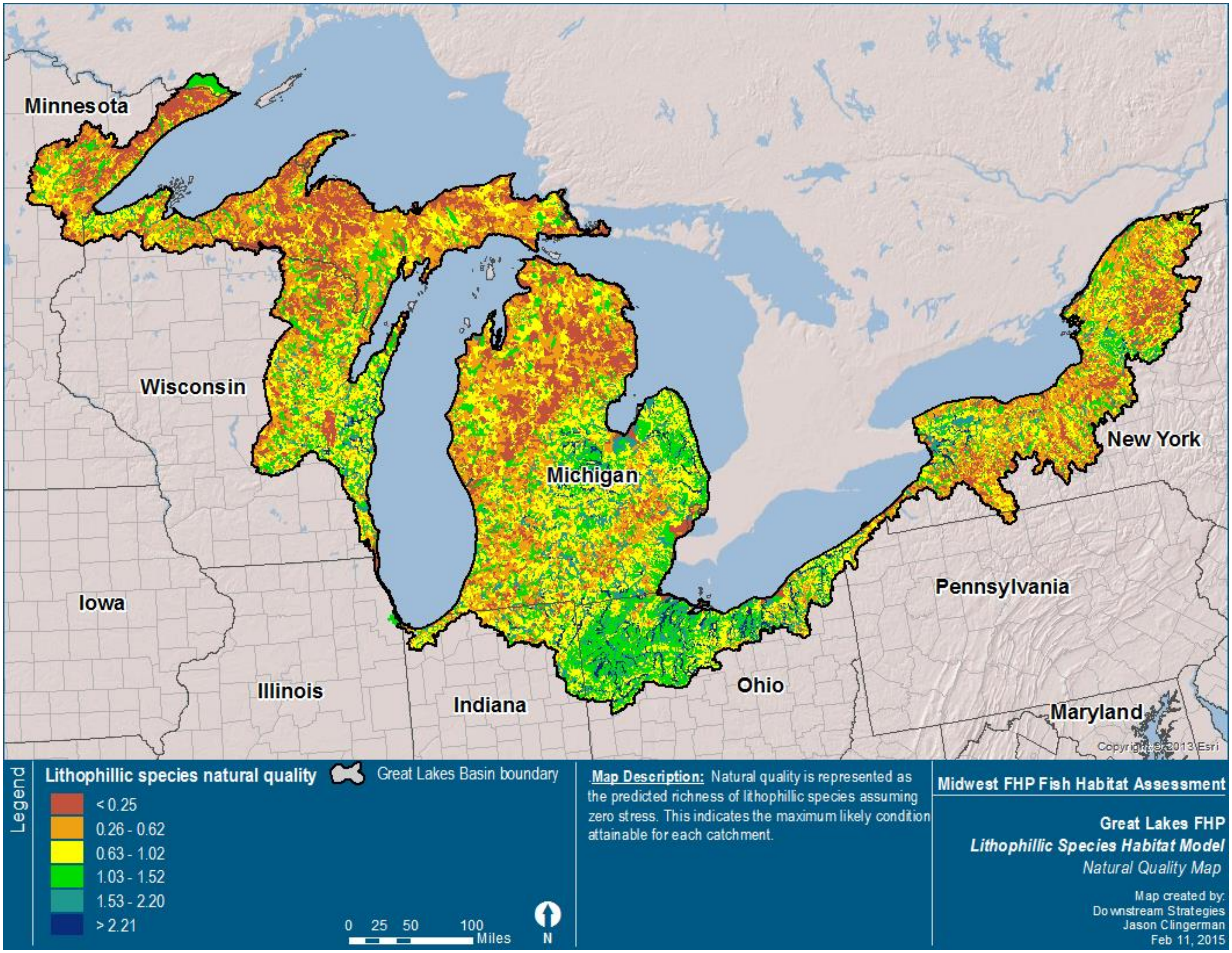


Figure 26: Anthropogenic stress index for lithophilic species richness

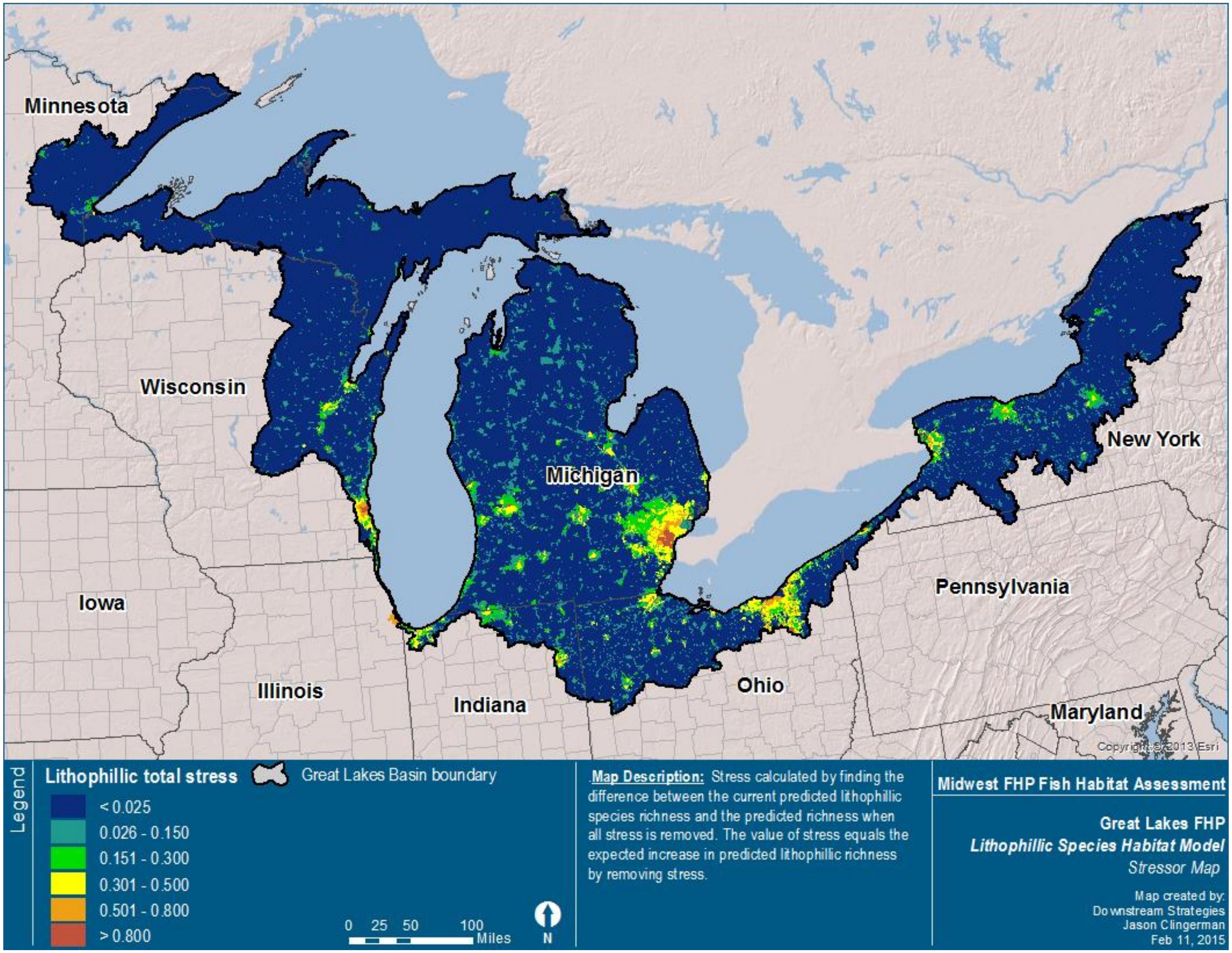


Figure 27: Most influential anthropogenic index metric for lithophilic species richness

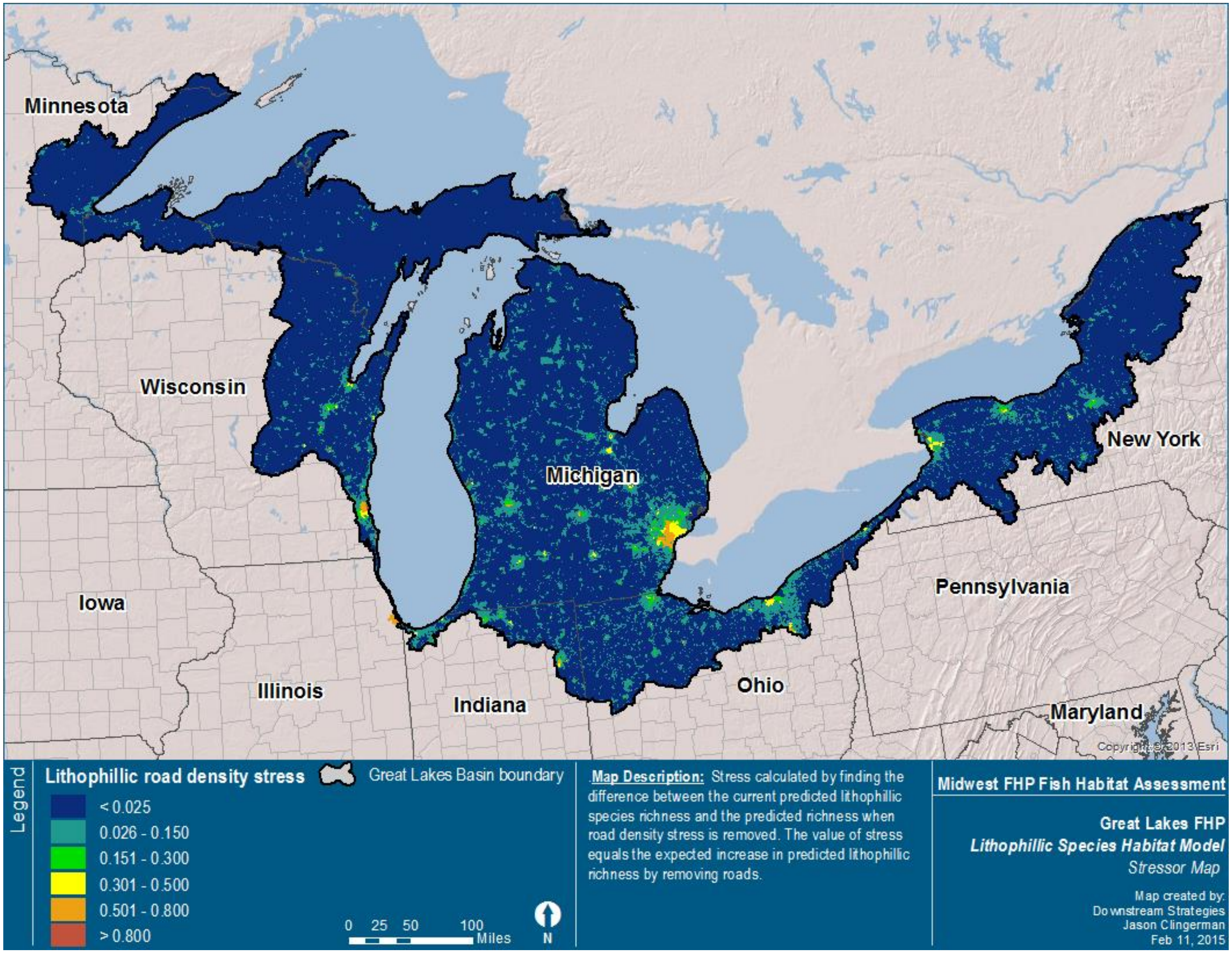
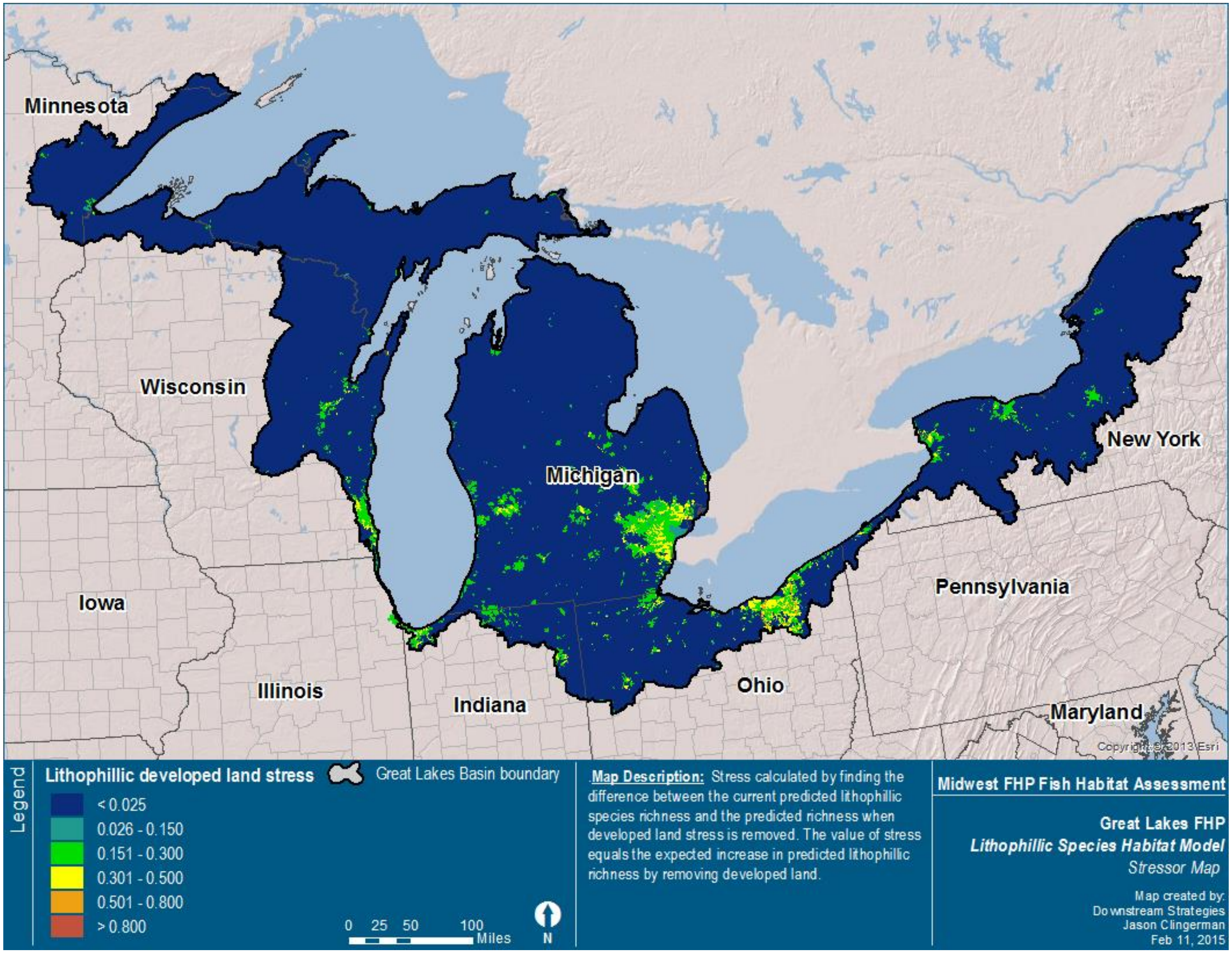


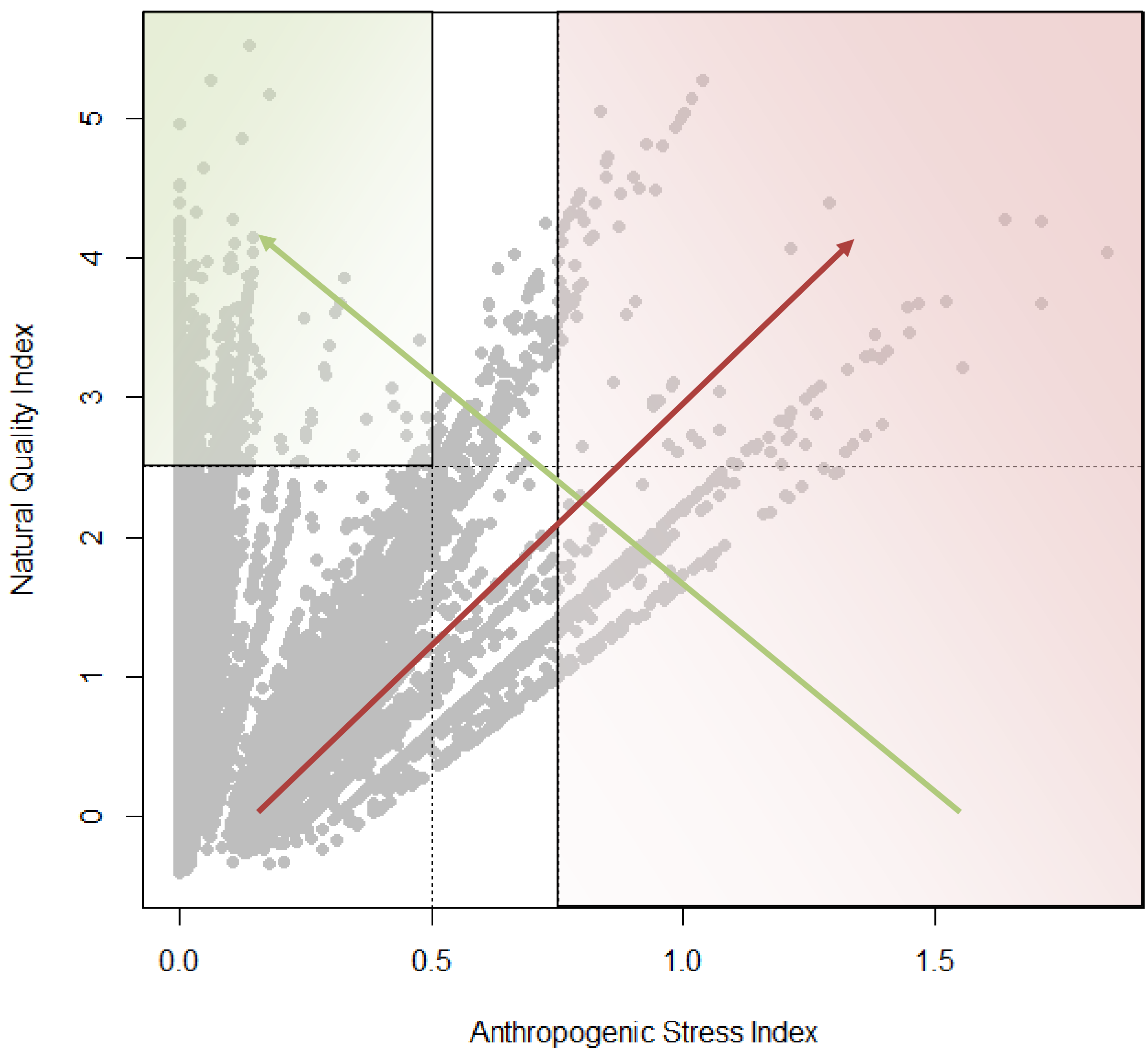
Figure 28: Second most influential anthropogenic index metric for lithophilic species richness



4.4.4 *Restoration and protection priorities*

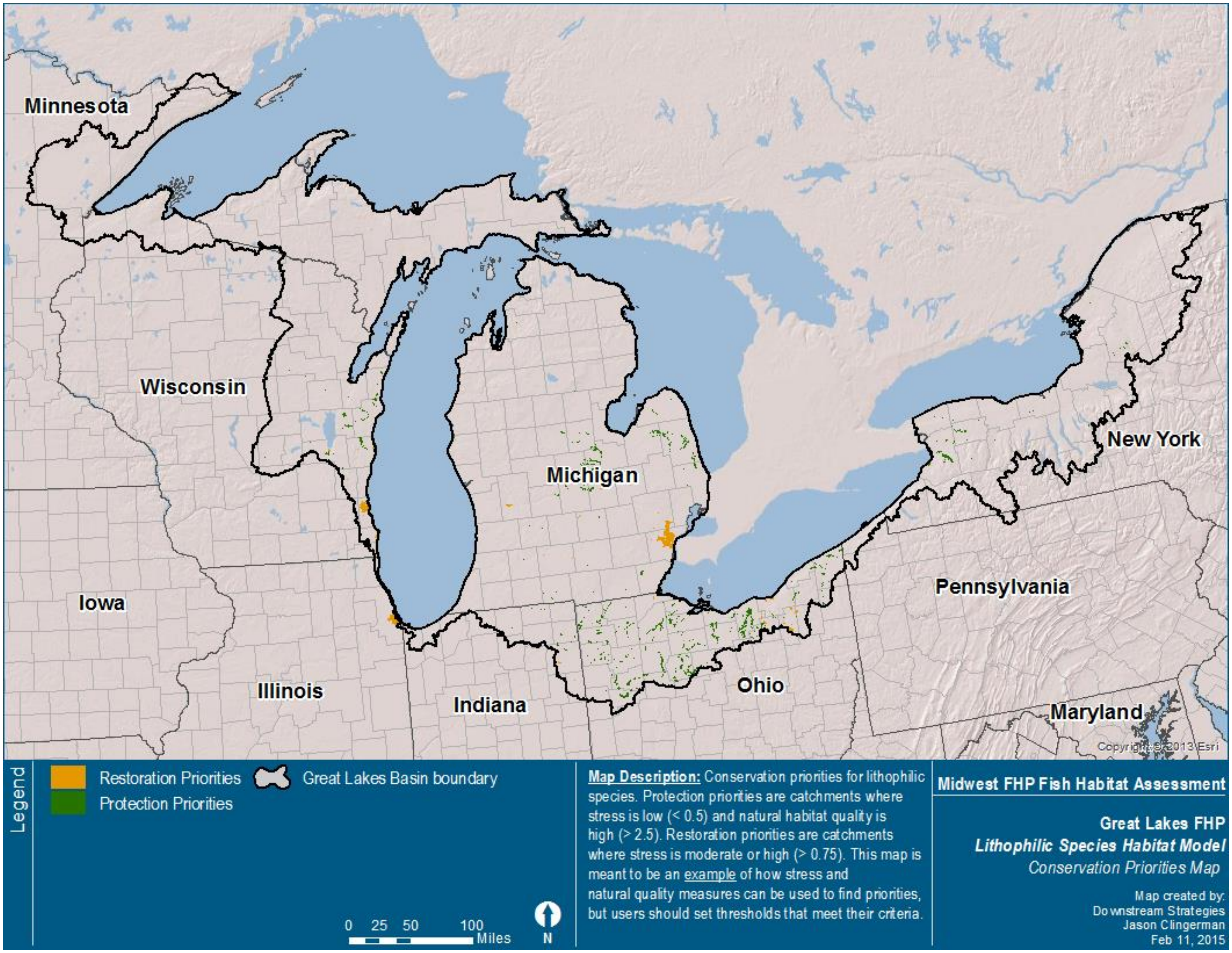
A plot of HQI versus ASI values for all catchments in the study area can be used as a reference to define HQI and ASI thresholds when evaluating restoration and protection priorities (Figure 29Figure 19). In the example shown, thresholds for protection priorities were defined as catchments with high natural habitat quality and low anthropogenic stress; these thresholds were based on HQI greater than 2.5 and ASI less than 0.5. The thresholds used to identify restoration priorities were defined as catchments with moderate to high anthropogenic stress; this threshold was based on ASI greater than 0.75. These classifications are mapped in Figure 30. These thresholds were solely based on the relative scores for natural quality and stress indices. Though this example scenario provides an informed set of criteria for identifying conservation priorities, it is only intended to demonstrate the functionality of querying catchments based on these attributes to identify areas that meet user-defined criteria to guide conservation, protection, and restoration planning.

Figure 29: HQI versus ASI values for all catchments for lithophilic species richness



Note: Breakpoints for HQI and ASI classes in this example are denoted by dashed lines. The arrows indicate the directions of increasing potential protection (green arrow) or restoration (red arrow) priority. The red box on the right indicates catchments defined as restoration priorities under the example scenario. The green box in the upper left corner indicates catchments defined as protection priorities under the same scenario.

Figure 30: Restoration and protection priorities for lithophilic species richness



5. WALLEYE

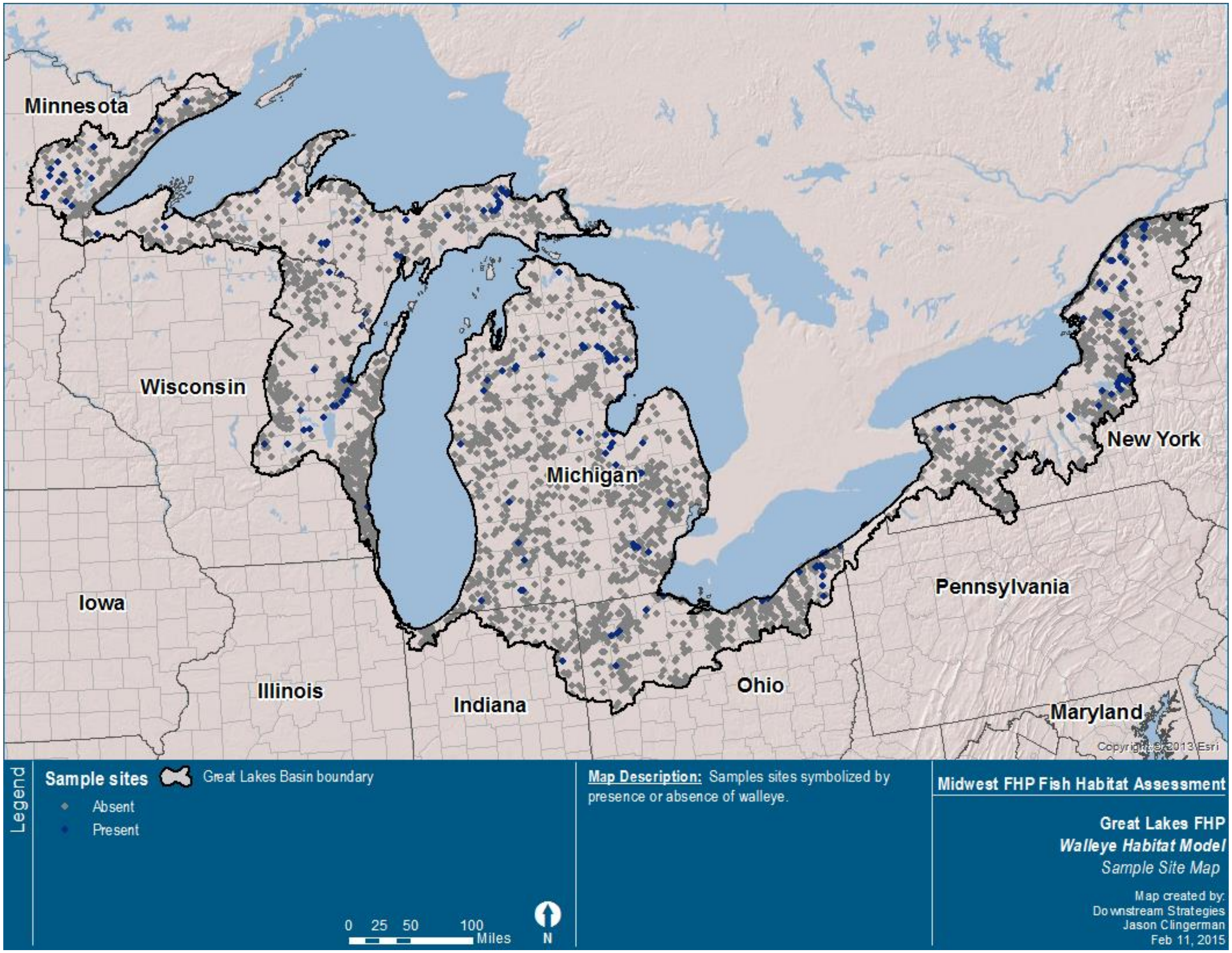
5.1 Modeling inputs

DS used a list of predictor variables selected by GLBFHP to develop a ten-fold CV BRT model for brook trout at the 1:100k catchment scale. The model was used to produce maps of expected brook trout distribution and maps of expected natural habitat quality and anthropogenic stress at the 1:100k scale throughout the extent of the GLBFHP.

DS cooperated with GLBFHP to arrive at a list of landscape-based habitat variables used to predict brook trout throughout the region; those variables were also used to characterizing habitat quality and anthropogenic stress. From an initial suite of 516 catchment attributes, DS and the GLBFHP compiled a list of 72 predictors for evaluation. From that list, 62 variables were removed due to statistical redundancy ($r > 0.6$), logical redundancy, or because of lack of model influence, resulting in a final list of 10 predictor variables for the BRT model and assessment. See Appendix A for a full data dictionary and the metadata document for variable processing notes.

GLBFHP provided DS with fish data collected in streams over a time frame spanning 1995 to 2006. Using that data, DS created a presence-absence dataset for walleye comprised of 3,240 observations. Figure 31 maps all of the sampling sites that were used to construct the model and the 1:100k catchments used in the model.

Figure 31: Walleye modeling area and sampling sites



5.2 Modeling process

5.2.1 *Predictive performance*

The final selected model was comprised of 2,500 trees. The model had a CV correlation statistic of 0.56±0.021 and a CV ROC score of 0.905±0.013.

5.2.2 *Variable influence*

The BRT output includes a list of the predictor variables used in the model ordered and scored by their relative importance. The relative importance values are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman and Meulman, 2003). The relative influence score is scaled so that the sum of the scores for all variables is 100, where higher numbers indicate greater influence. Of the 36 predictor variables used to develop the walleye model, 10 had a relative influence value greater than zero (Table 5).

Network drainage area, the single most important variable in terms of relative influence, contributed almost 65% of the total influence.

Table 5: Relative influence of all variables in the final walleye model

Variable Code	Variable Description	Relative Influence
Cumdrainag	Cumulative drainage area	64.8
Merge_temp	Modeled stream temperature	9.44
Lu_wetpc	Network wetlands	7.72
Precip	Mean annual precipitation	4.89
Slope	Slope of catchment flowline	4.50
Minelevraw	Minimum catchment elevation	2.56
Lf5pc	Landform code 4 (ground moraine), area (%), upstream cumulative	2.06
Dam_count_den	Density of dams in catchment	1.53
Soil1pc	Network soil hydrologic group A	1.26
Lu_devpc	Network developed land cover	1.21

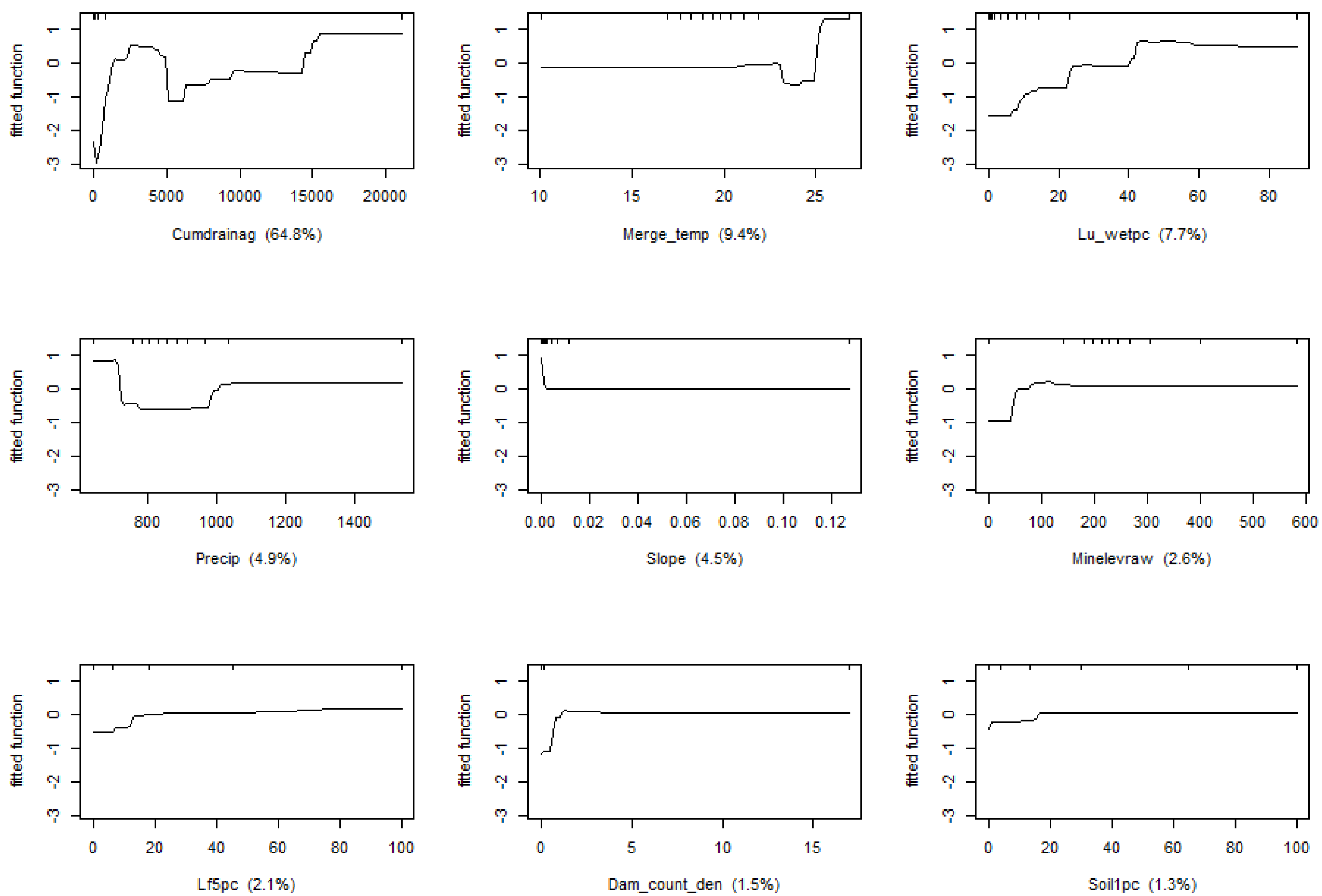
Note: Individual variables are highlighted according to whether they were determined to be anthropogenic in nature (gray shading) or natural (no shading).

5.2.3 *Variable functions*

The BRT output also contains quantitative information on partial dependence functions that can be plotted to visualize the effect of each individual predictor variable on the response after accounting for all other variables in the model. Similar to the interpretation of traditional regression coefficients, the function plots are not always a perfect representation of the relationship for each variable, particularly if interactions are strong or predictors are strongly correlated. However, they do provide a useful and objective basis for interpretation (Friedman, 2001; Friedman and Meulman, 2003).

These plots show the trend of the response variable (y-axis) as the predictor variable (x-axis) changes. The response variable is transformed (usually to the logit scale) so that the magnitude of trends for each predictor variable's function plot can be accurately compared. The dash marks at the top of each function represent the deciles of the data used to build the model. The function plots for the nine most influential variables in the walleye model (Table 5) are illustrated in Figure 32. The plots for all 10 variables are shown in Appendix B.

Figure 32: Functional responses of the dependent variable to individual predictors of walleye



Note: Only the top nine predictors, based on relative influence (shown in parentheses; see Appendix A for descriptions of variable codes), are shown here. See Appendix B for plots of remaining predictor variables.

5.3 Post-modeling

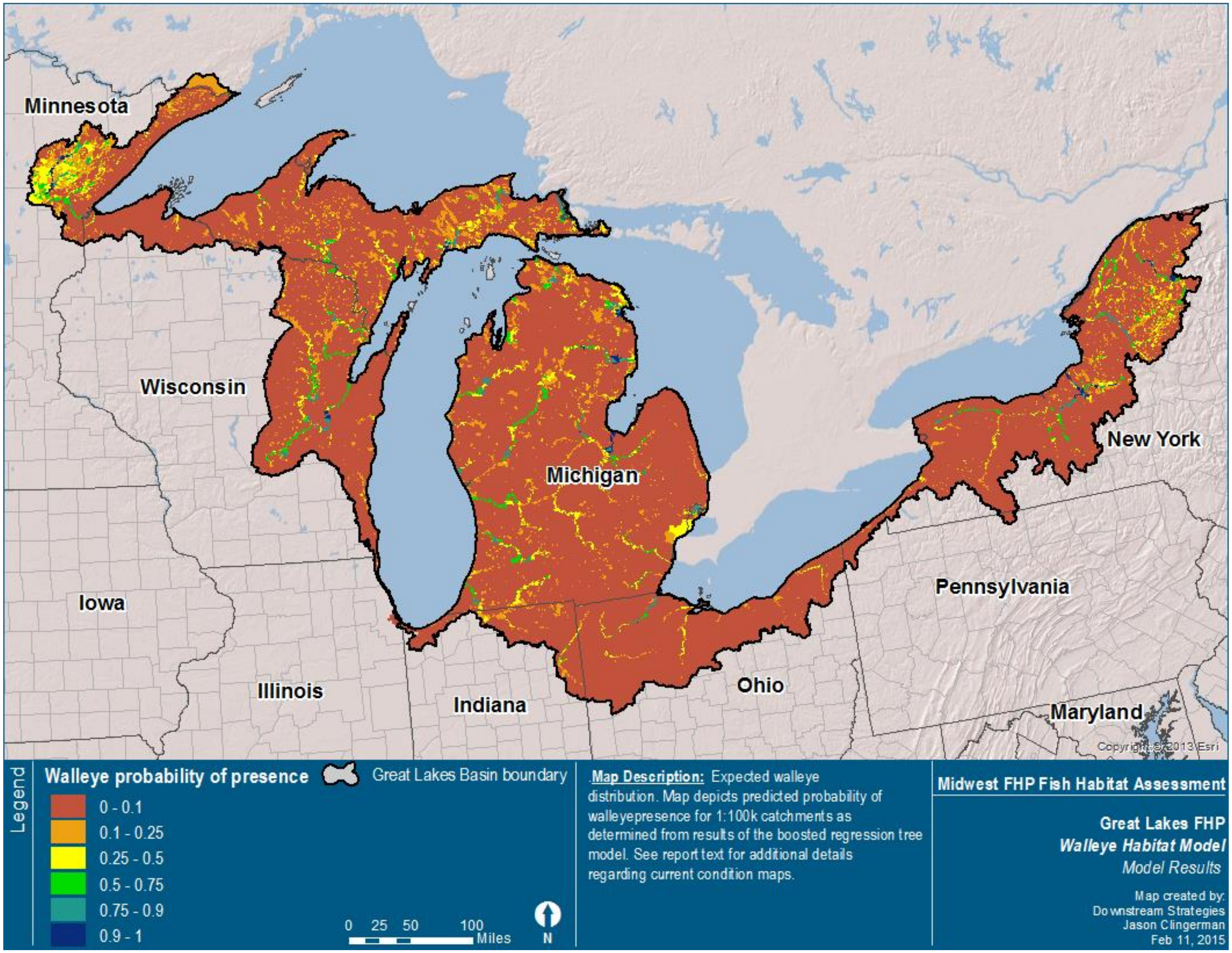
The variable importance table and partial dependence functions of the final BRT model were used to assess the potential stressors for the walleye model. Within the model, there were three variables considered anthropogenic in nature (Table 5). After reviewing the functional relationships of these three potential stressors, two stressors were removed from ASI calculations. These variables ('Lu_wetpc' and 'Dam_count_den') had a function plot that was unintuitive: their relationships to the response likely captured some sort of regional variation in the model rather than a mechanistic relationship with the response. The remaining stressor, network developed land cover (Lu_devpc) was used to calculate ASI for the walleye model. Section 1.3.2 details how ASI and HQI were calculated for each model.

5.4 Mapped results

5.4.1 *Expected current conditions*

Walleye probability of presence was calculated for all 1:100k stream catchments in the study area using the BRT model. The predicted probability values ranged from 0 to 1. The mean predicted probability was 0.067. Of the total 104,343 catchments, there were 850 catchments with a predicted probability of presence greater than 0.75 and 1,997 catchments where the probability of presence was between 0.5 and 0.75. These results are mapped in Figure 33.

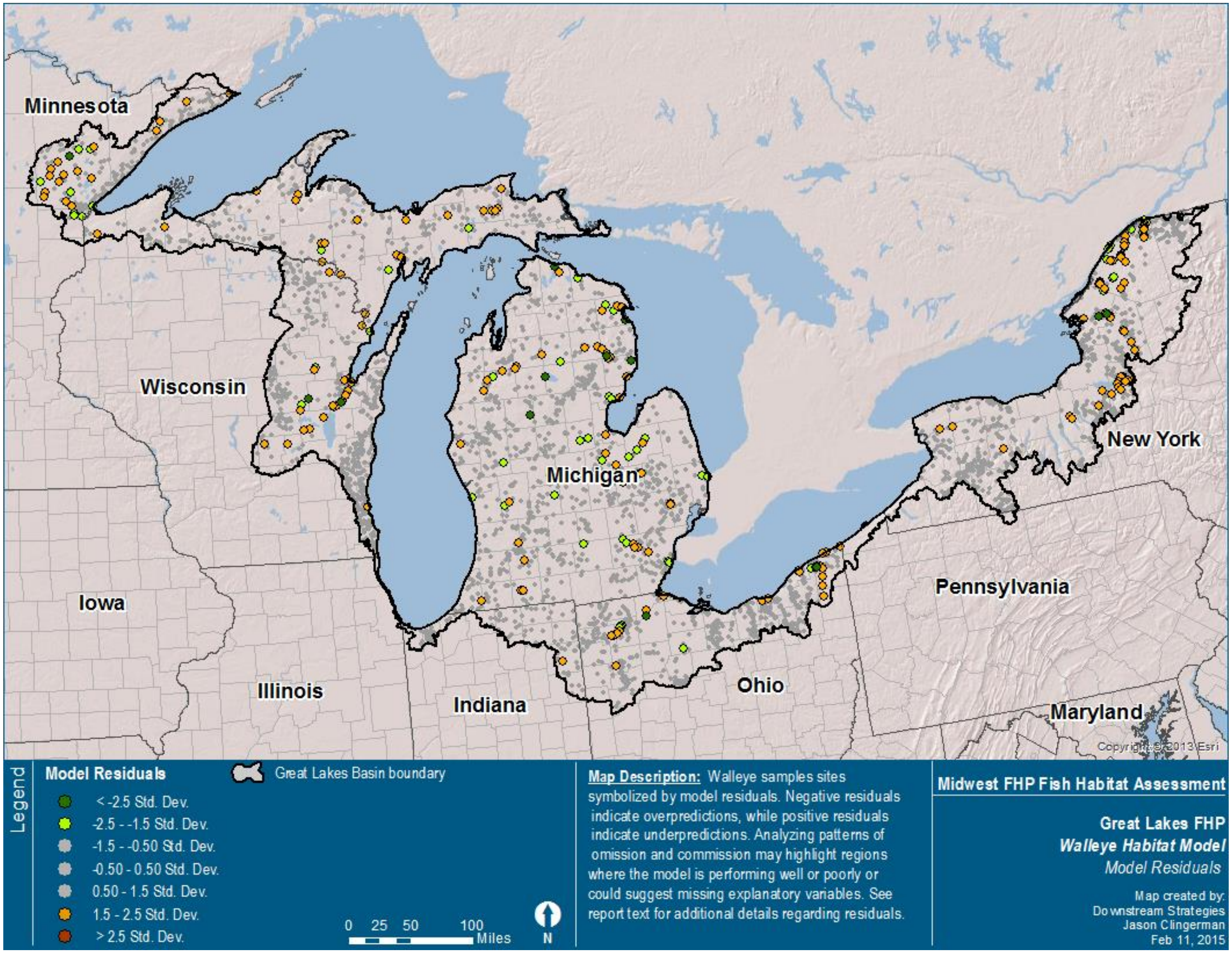
Figure 33: Expected walleye distribution



5.4.2 *Spatial variability in predictive performance*

Analyzing patterns of omission and commission may highlight regions where the model is performing well or poorly or could suggest missing explanatory variables (Figure 34). To assess omission and commission, residuals are also calculated by the BRT model. The residuals are a measure of the difference in the measured and modeled values (measured value *minus* modeled value). Negative residuals indicate overpredictions (predicting higher values than are true), while positive residuals indicate underpredictions (predicting lower values than are true).

Figure 34: Distribution of walleye model residuals by sampling site



5.4.3 *Indices of stress and natural quality*

Maps of HQI and ASI illustrate the spatial distribution of natural habitat potential (i.e., HQI score) and anthropogenic stress (i.e., ASI score) in the GLBFHP. HQI and ASI scores are mapped in Figure 35Figure 25 and Figure 36, respectively. HQI, ASI, and their metrics are all scaled on a 0-1 scale (see Section 2.3 for more details on HQI and ASI calculation). For HQI, higher values indicate higher natural quality, while higher values for ASI indicate higher levels of anthropogenic stress.

At first glance, it may seem that regional stress conditions are overly optimistic, but it is necessary to consider that the stress index is showing areas where probability of presence for this response is reduced because of stressors. It is likely that stress on aquatic systems in general is much more widespread than is indicated in any individual model's stress maps. For all stress and natural quality indices, all catchments are shown, even in areas where the probability of presence is low. This is necessary and useful to consider areas outside of the current expect range where stress could have caused a historic population to be extirpated.

Figure 35: Natural quality index for walleye

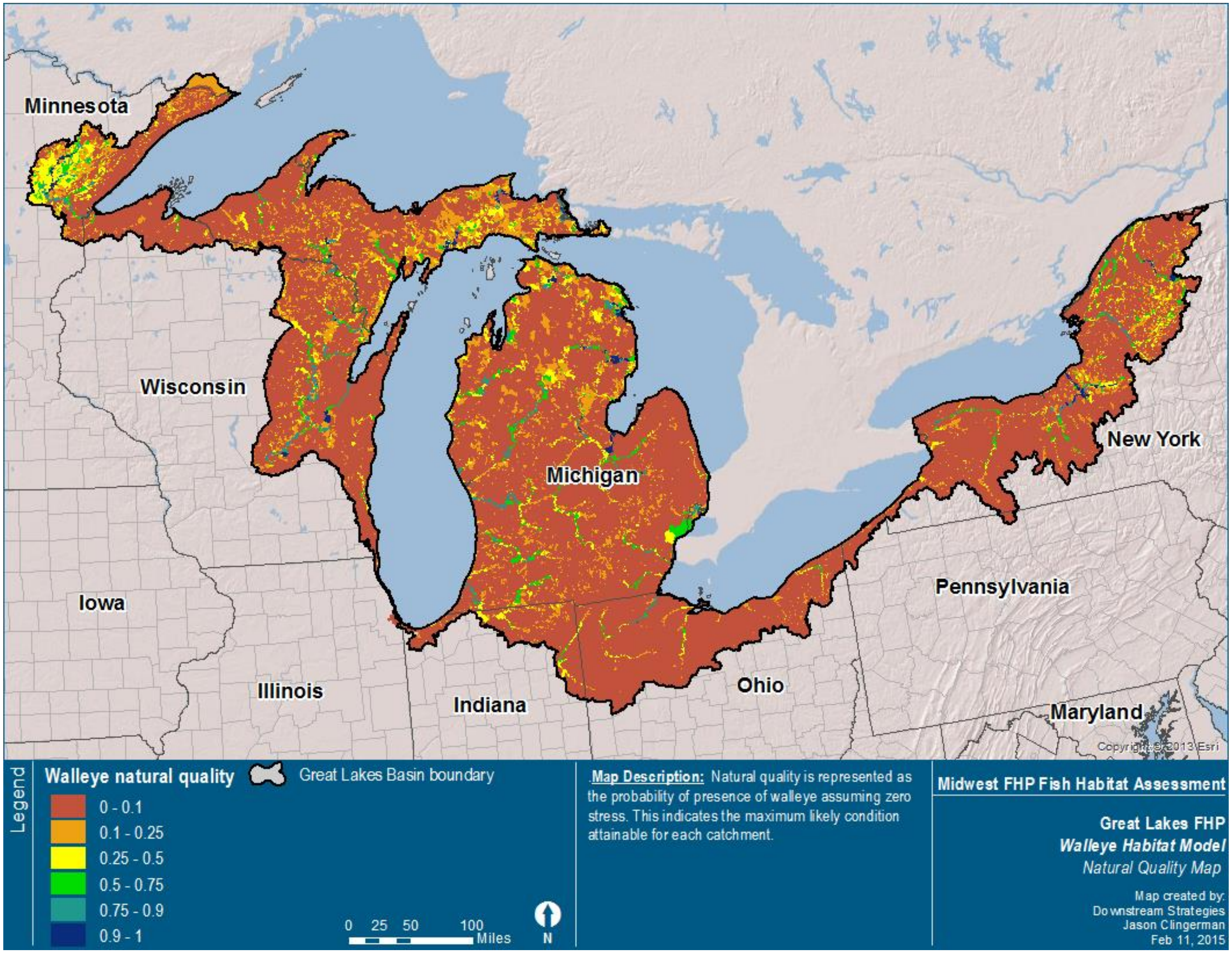
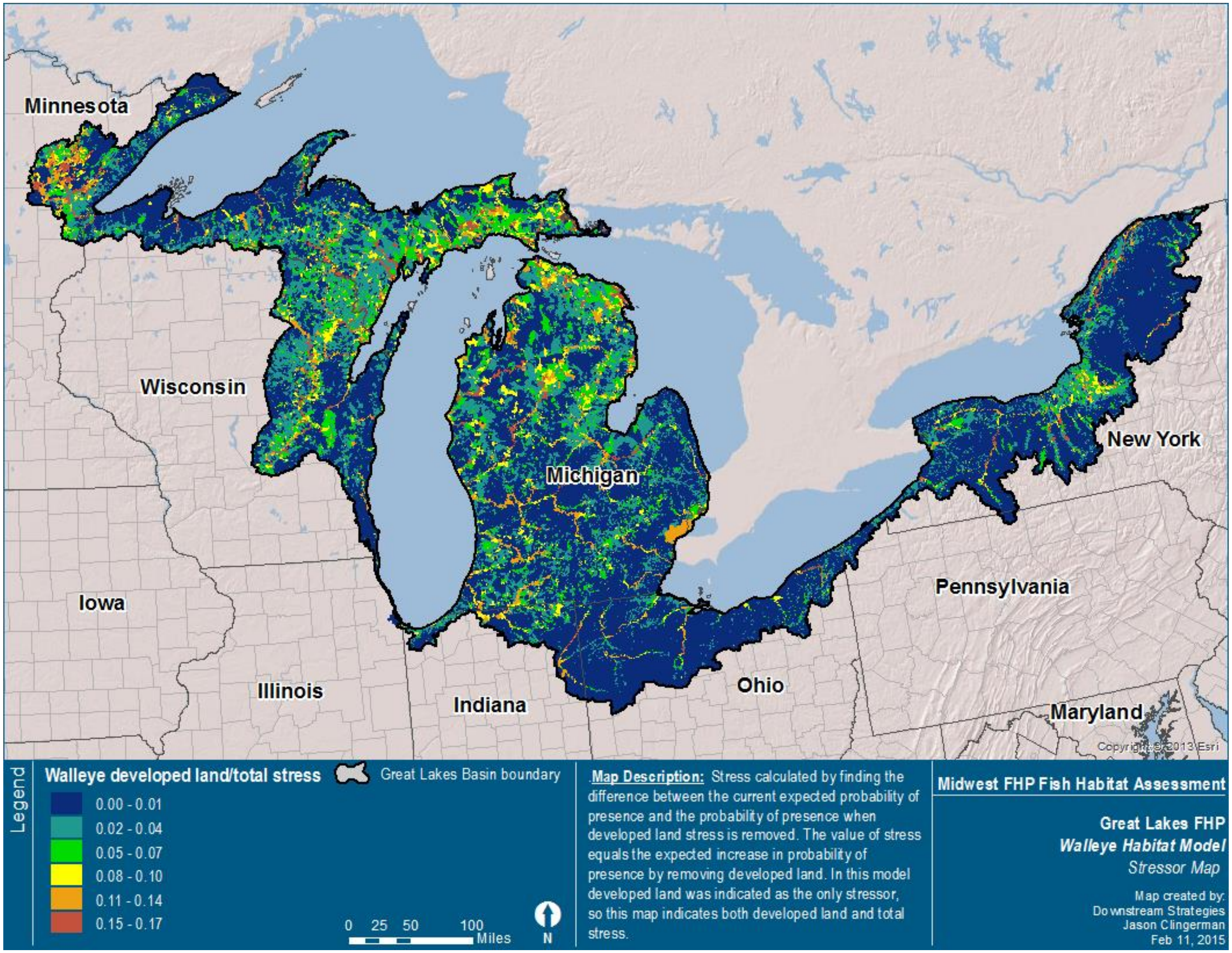


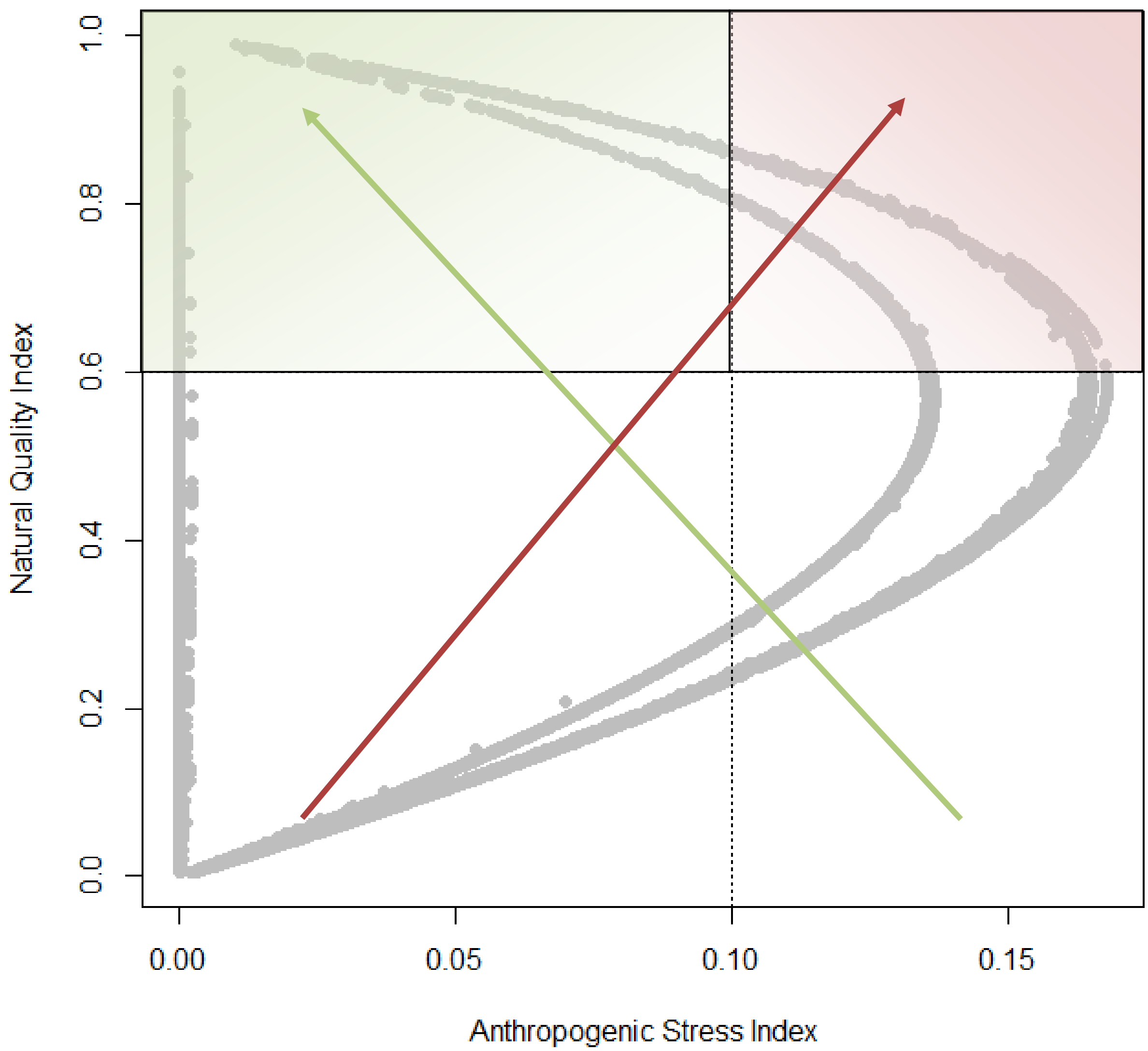
Figure 36: Anthropogenic stress index for walleye



5.4.4 *Restoration and protection priorities*

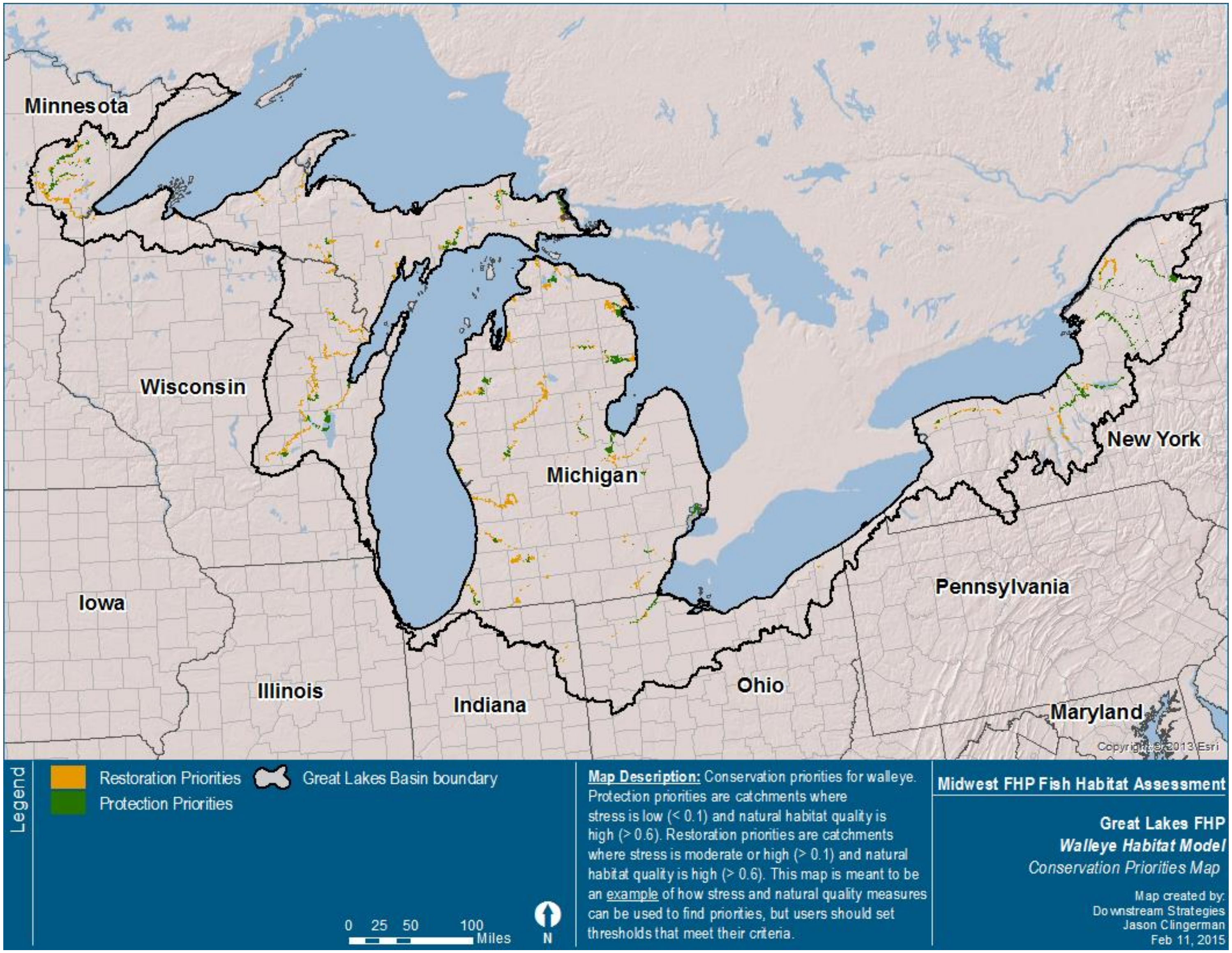
A plot of HQI versus ASI values for all catchments in the study area can be used as a reference to define HQI and ASI thresholds when evaluating restoration and protection priorities (Figure 37). In the example shown, thresholds for protection priorities were defined as catchments with high natural habitat quality and low anthropogenic stress; these thresholds were based on HQI greater than 0.6 and ASI less than 0.1. The thresholds used to identify restoration priorities were defined as catchments with moderate to high anthropogenic stress; this threshold was based on HQI greater than 0.6 and ASI greater than 0.1. These classifications are mapped in Figure 38. These thresholds were solely based on the relative scores for natural quality and stress indices. Though this example scenario provides an informed set of criteria for identifying conservation priorities, it is only intended to demonstrate the functionality of querying catchments based on these attributes to identify areas that meet user-defined criteria to guide conservation, protection, and restoration planning.

Figure 37: HQI versus ASI values for all catchments for walleye



Note: Breakpoints for HQI and ASI classes in this example are denoted by dashed lines. The arrows indicate the directions of increasing potential protection (green arrow) or restoration (red arrow) priority. The red box in the upper right indicates catchments defined as restoration priorities under the example scenario. The green box in the upper left indicates catchments defined as protection priorities under the same scenario.

Figure 38: Restoration and protection priorities for walleye



6. LARGE RIVER SPECIES

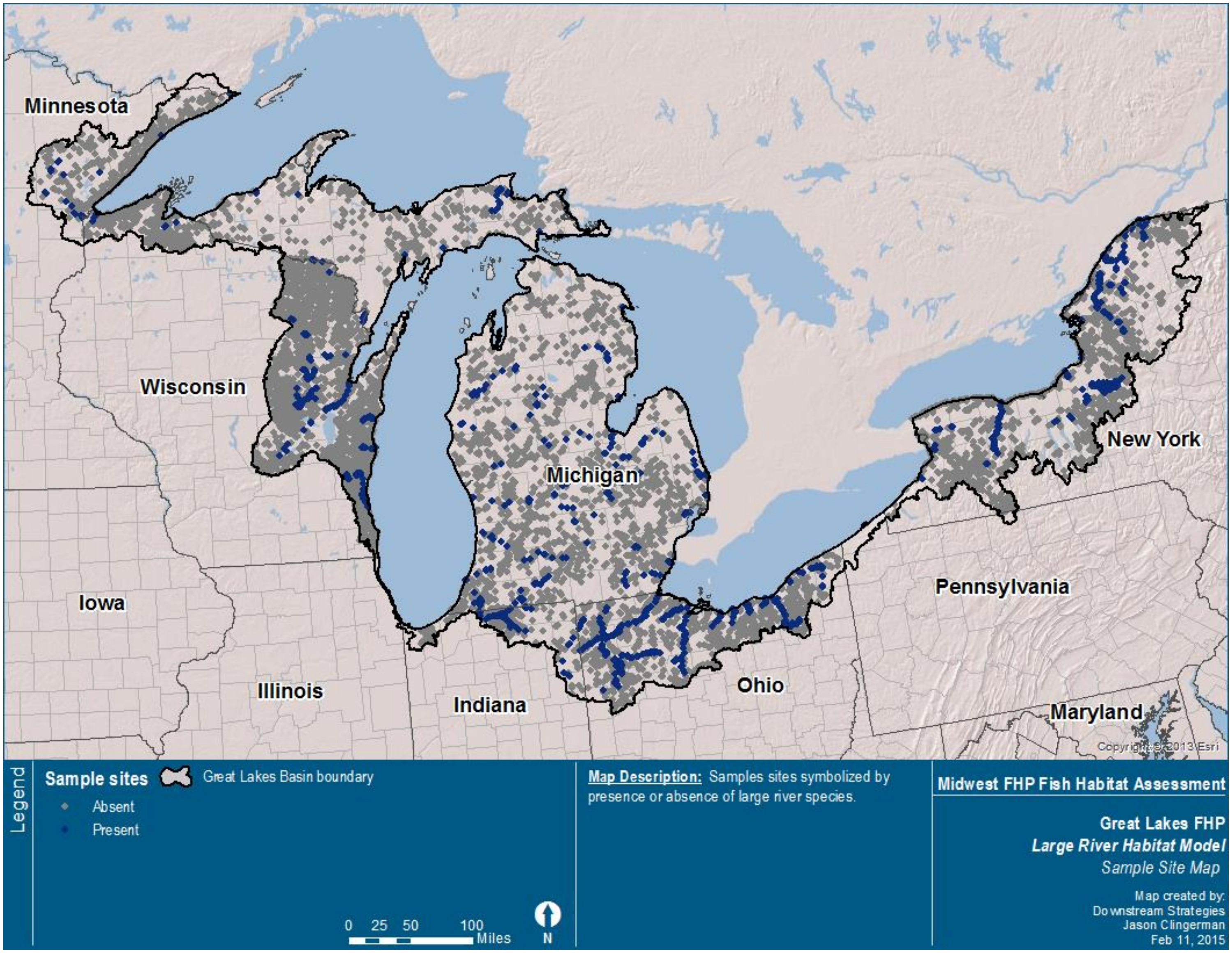
6.1 Modeling inputs

DS used a list of predictor variables selected by GLBFHP to develop a ten-fold CV BRT model for brook trout at the 1:100k catchment scale. The model was used to produce maps of expected brook trout distribution and maps of expected natural habitat quality and anthropogenic stress at the 1:100k scale throughout the extent of the GLBFHP.

DS cooperated with GLBFHP to arrive at a list of landscape-based habitat variables used to predict brook trout throughout the region; those variables were also used to characterizing habitat quality and anthropogenic stress. From an initial suite of 516 catchment attributes, DS and the GLBFHP compiled a list of 72 predictors for evaluation. From that list, 66 variables were removed due to statistical redundancy ($r > 0.6$), logical redundancy, or because of lack of model influence, resulting in a final list of 6 predictor variables for the BRT model and assessment. See Appendix A for a full data dictionary and the metadata document for variable processing notes.

GLBFHP provided DS with a presence-absence dataset for large river species comprised of 9,368 observations collected in streams over a time frame spanning 1995 to 2006. Figure 39 maps all of the sampling sites that were used to construct the model and the 1:100k catchments used in the model.

Figure 39: Large river species modeling area and sampling sites



6.2 Modeling process

6.2.1 *Predictive performance*

The final selected model was comprised of 5,600 trees. The model had a CV correlation statistic of 0.832±0.008 and a CV ROC score of 0.982±0.001.

6.2.2 *Variable influence*

The BRT output includes a list of the predictor variables used in the model ordered and scored by their relative importance. The relative importance values are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman and Meulman, 2003). The relative influence score is scaled so that the sum of the scores for all variables is 100, where higher numbers indicate greater influence. Of the 36 predictor variables used to develop the large river species model, six had a relative influence value greater than zero (Table 6).

Network drainage area, the single most important variable in terms of relative influence, contributed 64% of the total influence.

Table 6: Relative influence of all variables in the final large river species model

Variable Code	Variable Description	Relative Influence
Cumdrainag	Cumulative drainage area	64.00
Merge_temp	Modeled stream temperature	16.98
Roadcrc_den	Cumulative road crossing density	7.26
Cercc_den	Cumulative Density of Compensation and Liability Information System sites	6.85
Minelevraw	Minimum catchment elevation	3.55
Water_gw	Groundwater use by county (millions gallons per day/km2)	1.26

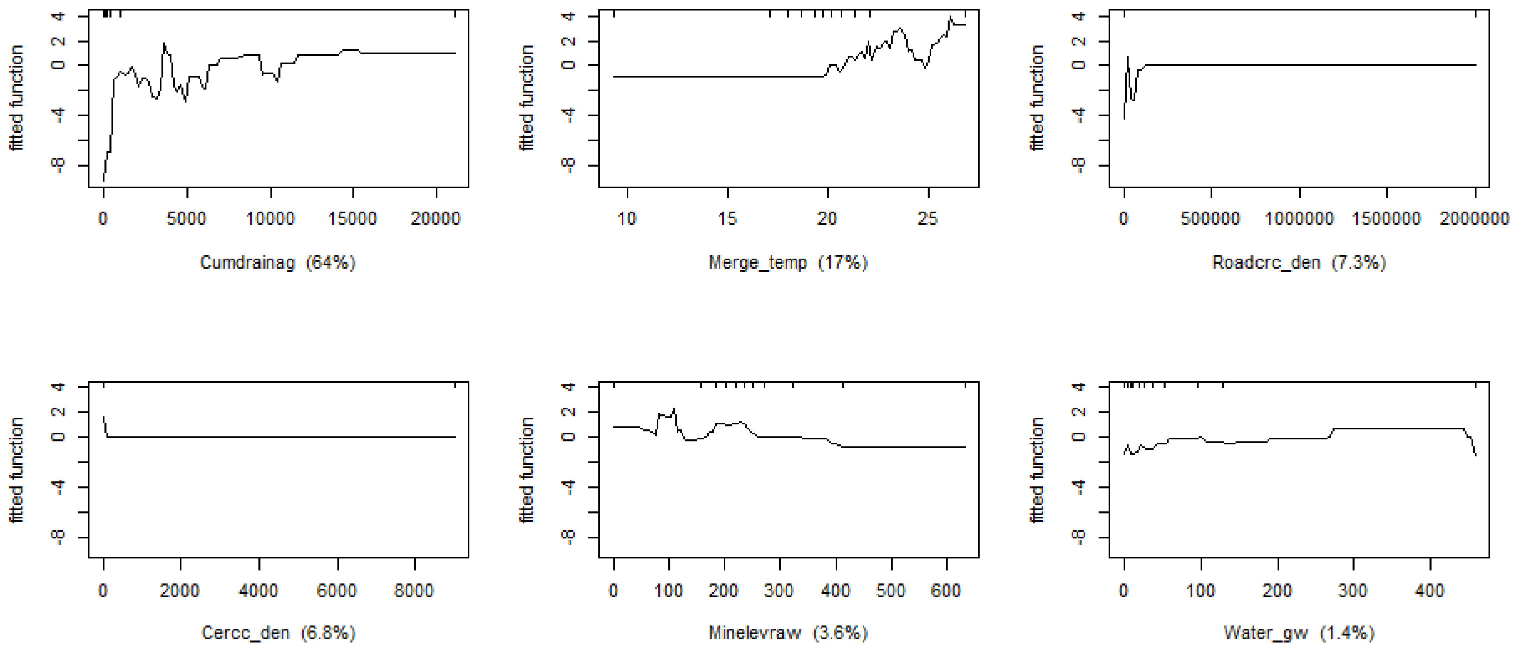
Note: Individual variables are highlighted according to whether they were determined to be anthropogenic in nature (gray shading) or natural (no shading).

6.2.3 *Variable functions*

The BRT output also contains quantitative information on partial dependence functions that can be plotted to visualize the effect of each individual predictor variable on the response after accounting for all other variables in the model. Similar to the interpretation of traditional regression coefficients, the function plots are not always a perfect representation of the relationship for each variable, particularly if interactions are strong or predictors are strongly correlated. However, they do provide a useful and objective basis for interpretation (Friedman, 2001; Friedman and Meulman, 2003).

These plots show the trend of the response variable (y-axis) as the predictor variable (x-axis) changes. The response variable is transformed (usually to the logit scale) so that the magnitude of trends for each predictor variable's function plot can be accurately compared. The dash marks at the top of each function represent the deciles of the data used to build the model. The function plots for the nine most influential variables in the large river species model (Table 6) are illustrated in Figure 40. The plots for all 38 variables are shown in Appendix B.

Figure 40: Functional responses of the dependent variable to individual predictors of large river species



Note: See Appendix A for descriptions of variable codes.

6.3 Post-modeling

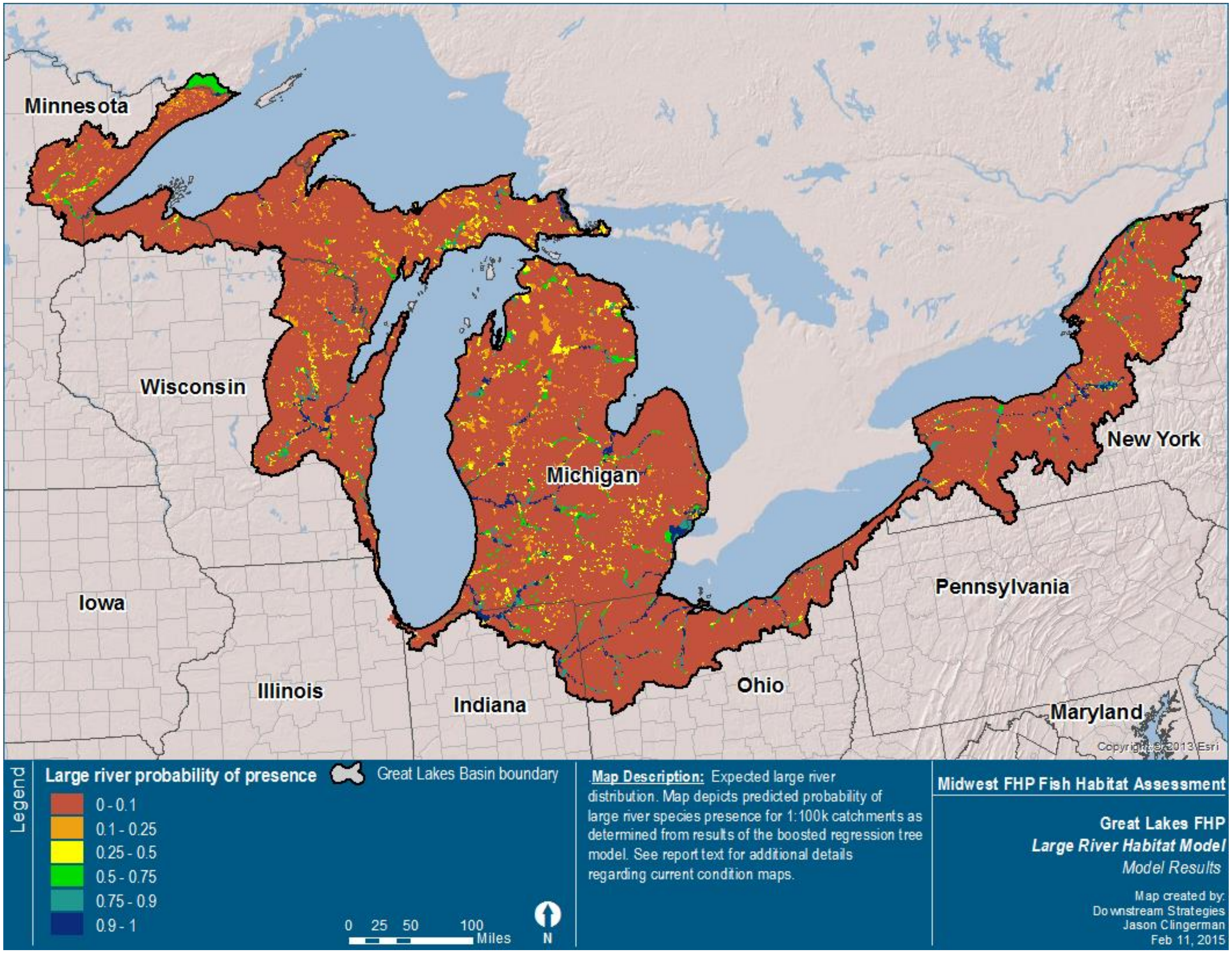
The variable importance table and partial dependence functions of the final BRT model were used to assess the potential stressors for the large river model. Within the model, there were three variables considered anthropogenic in nature (Table 6). After reviewing the functional relationships of these three potential stressors, two stressors were removed from ASI calculations. These variables (‘Roadcrc_den’ and ‘Water_gw’) had a function plot that was unintuitive: their relationships to the response likely captured some sort of regional variation in the model rather than a mechanistic relationship with the response. The remaining stressor, network density of superfund sites (‘Cercc_den’) was used to calculate ASI for the large river model. Section 1.3.2 details how ASI and HQI were calculated for each model.

6.4 Mapped results

6.4.1 *Expected current conditions*

Large river species probability of presence was calculated for all 1:100k stream catchments in the study area using the BRT model. The predicted probability values ranged from 0 to 1. The mean predicted probability was 0.063. Of the total 104,343 catchments, there were 3,207 catchments with a predicted probability of presence greater than 0.75 and 1,742 catchments where the probability of presence was between 0.5 and 0.75. These results are mapped in Figure 41.

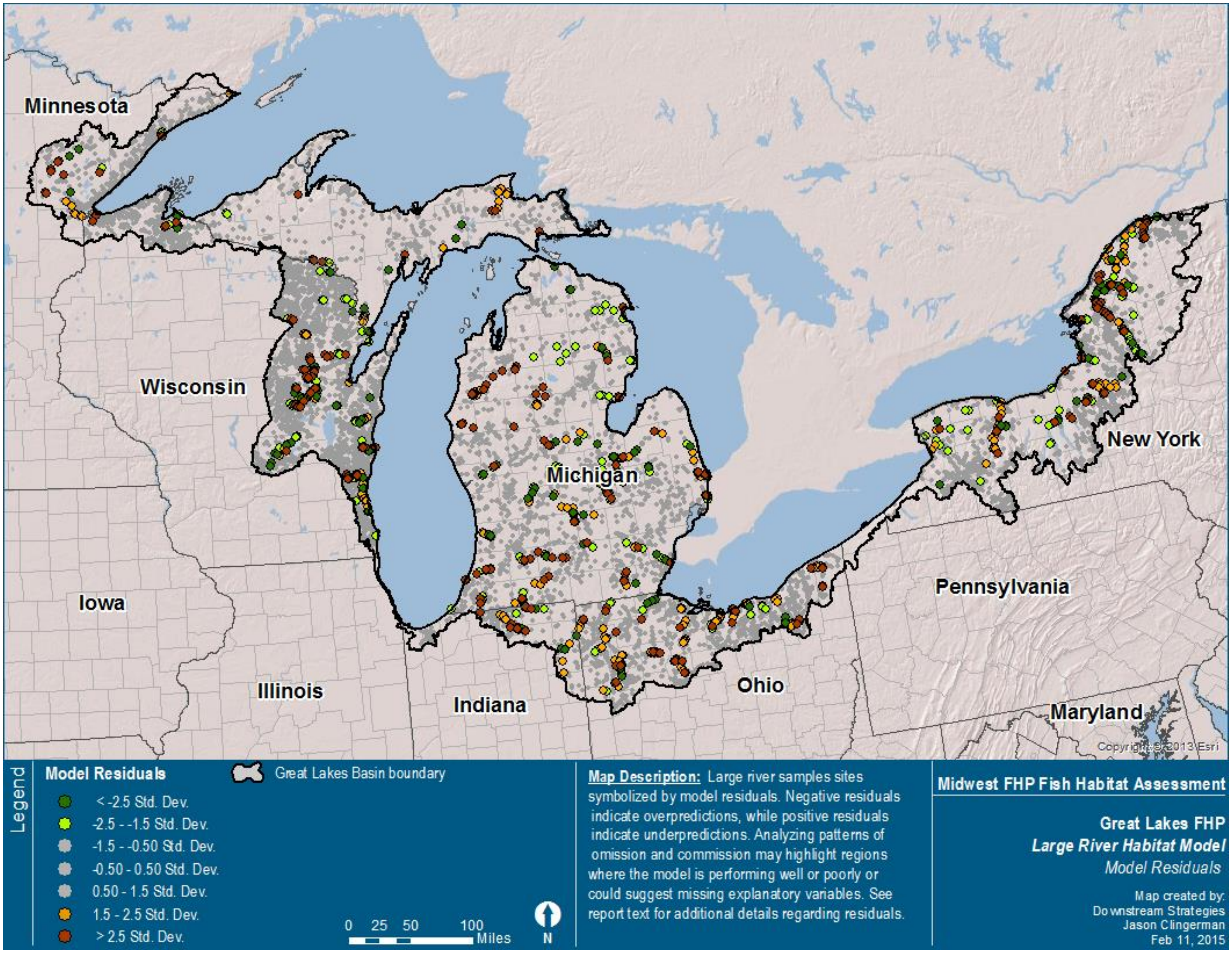
Figure 41: Expected large river species distribution



6.4.2 *Spatial variability in predictive performance*

Analyzing patterns of omission and commission may highlight regions where the model is performing well or poorly or could suggest missing explanatory variables (Figure 42). To assess omission and commission, residuals are also calculated by the BRT model. The residuals are a measure of the difference in the measured and modeled values (measured value *minus* modeled value). Negative residuals indicate overpredictions (predicting higher values than are true), while positive residuals indicate underpredictions (predicting lower values than are true).

Figure 42: Distribution of large river species model residuals by sampling site



6.4.3 *Indices of stress and natural quality*

Maps of HQI and ASI illustrate the spatial distribution of natural habitat potential (i.e., HQI score) and anthropogenic stress (i.e., ASI score) in the GLBFHP. HQI and ASI scores are mapped in Figure 43 and Figure 44, respectively. HQI and ASI are all scaled on a 0-1 scale (see Section 2.3 for more details on HQI and ASI calculation). For HQI, higher values indicate higher natural quality, while higher values for ASI indicate higher levels of anthropogenic stress.

At first glance, it may seem that regional stress conditions are overly optimistic, but it is necessary to consider that the stress index is showing areas where probability of presence for this response is reduced because of stressors. It is likely that stress on aquatic systems in general is much more widespread than is indicated in any individual model's stress maps. For all stress and natural quality indices, all catchments are shown, even in areas where the probability of presence is low. This is necessary and useful to consider areas outside of the current expect range where stress could have caused a historic population to be extirpated.

Figure 43: natural quality index for large river species

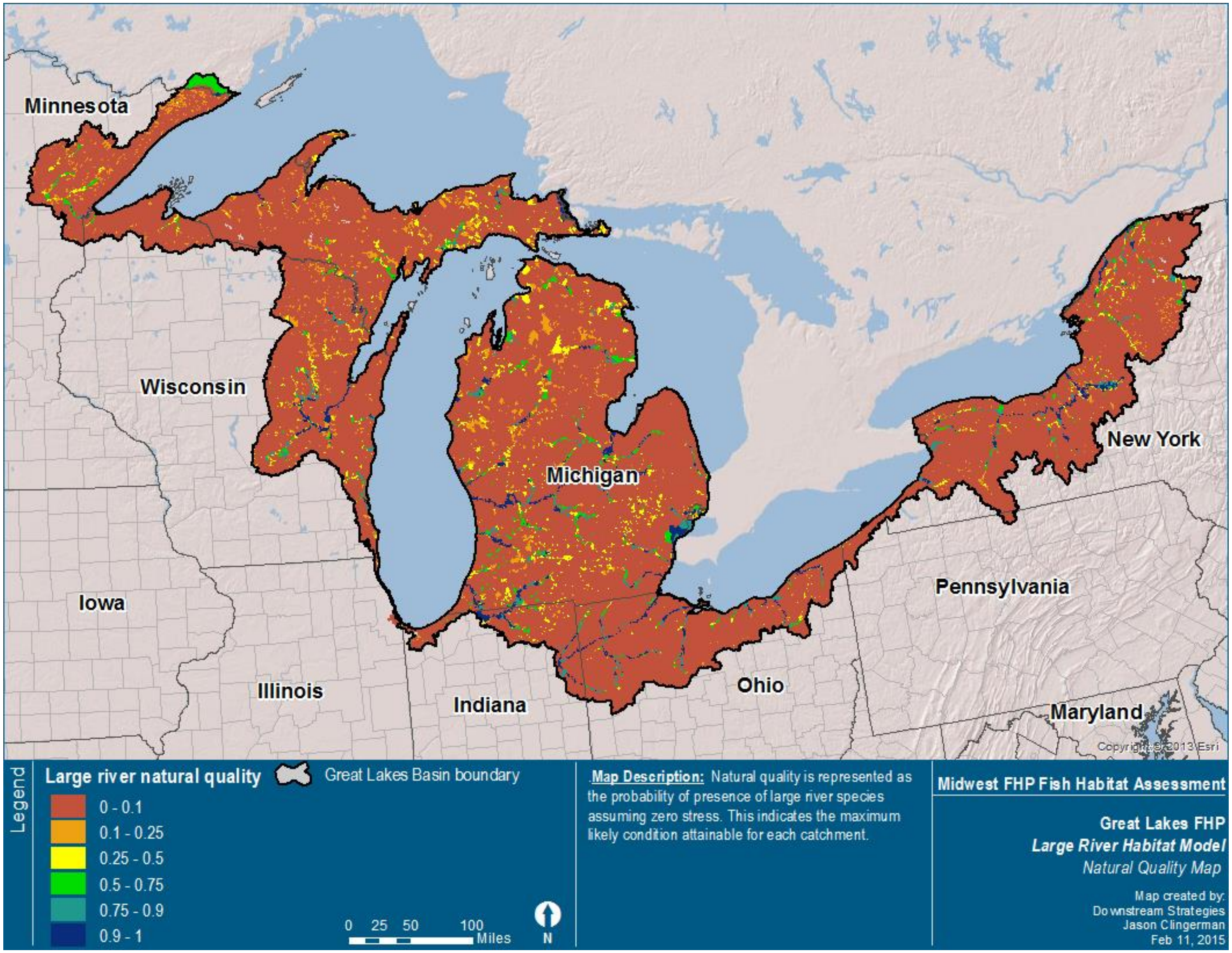
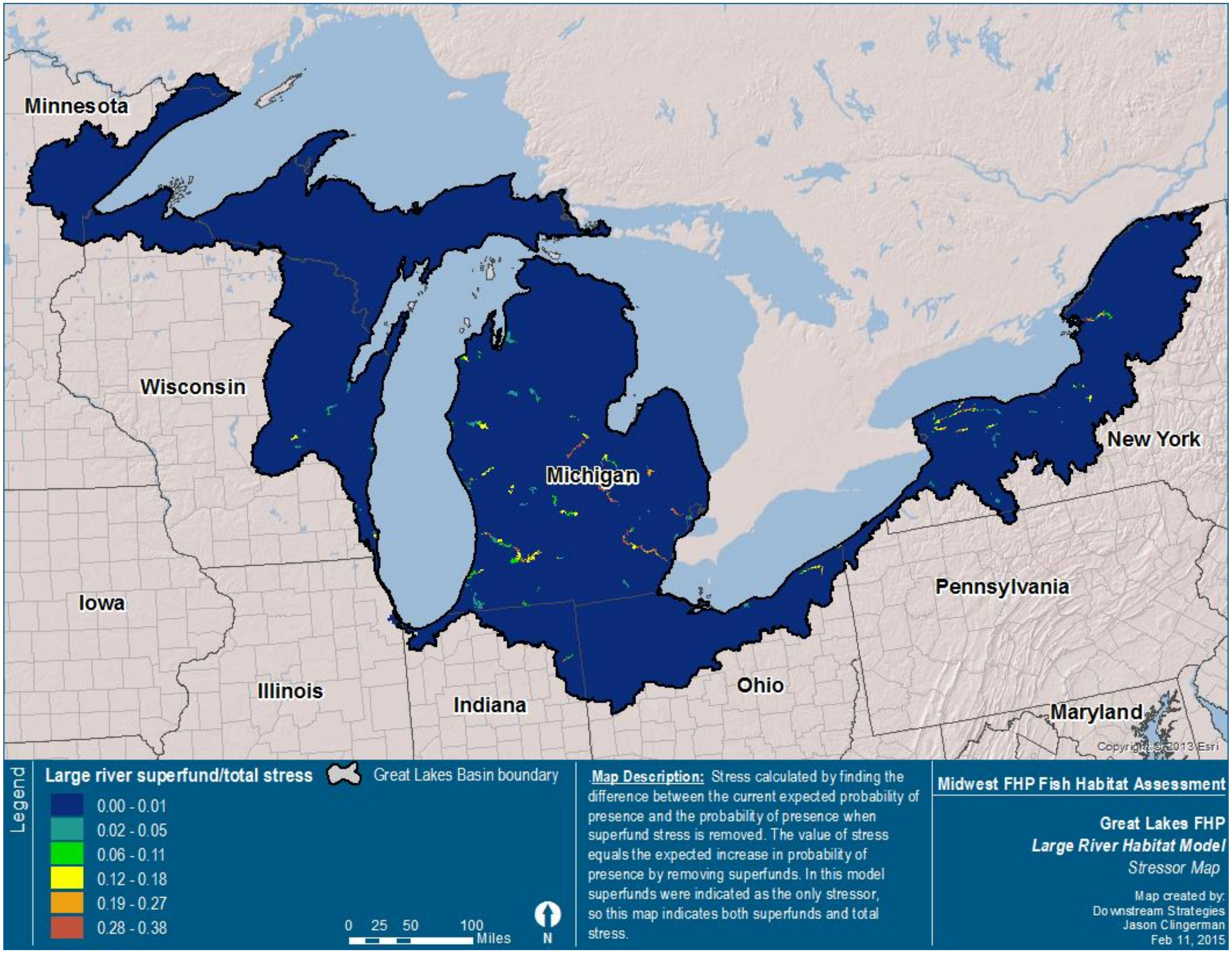


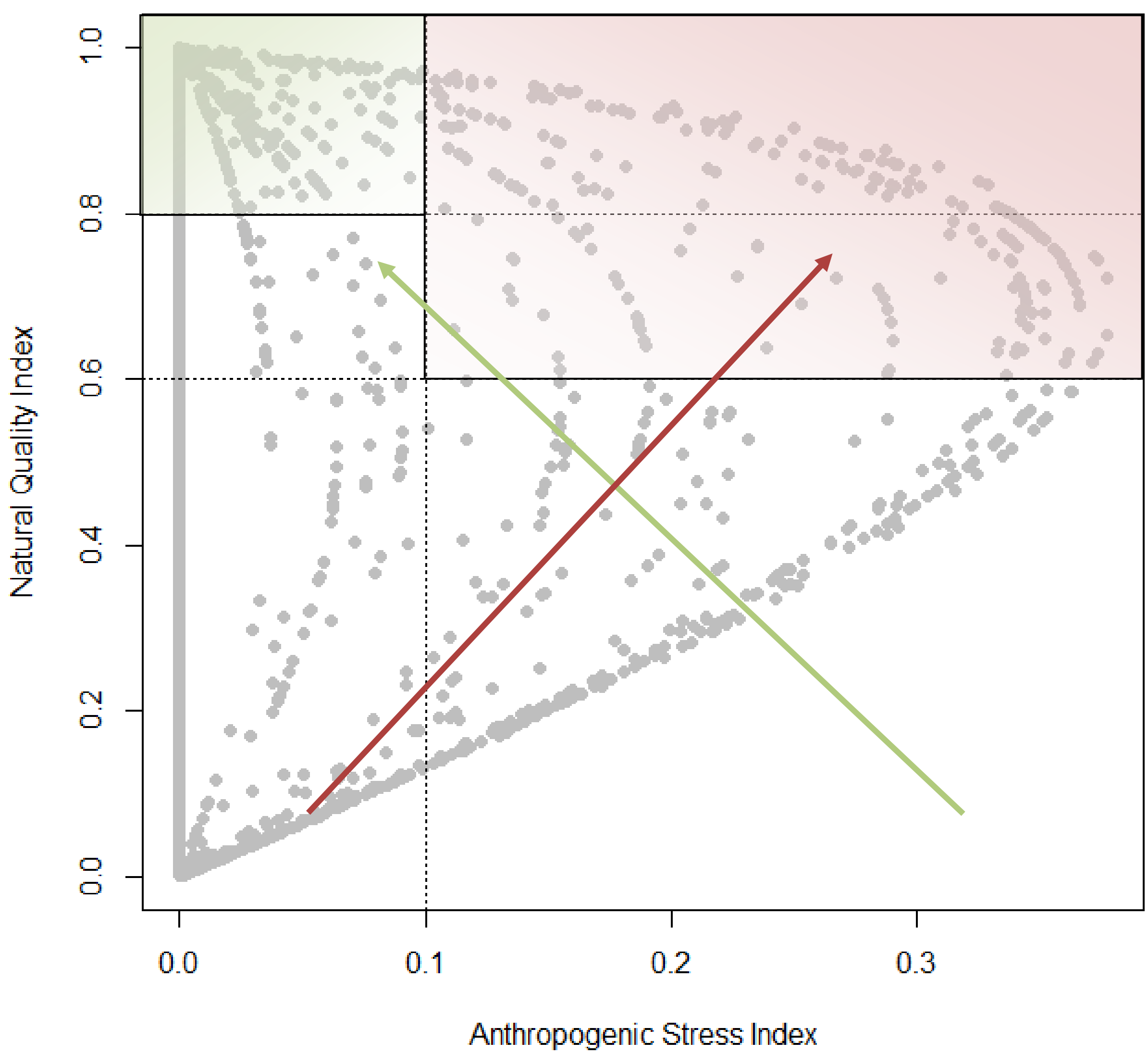
Figure 44: Superfund site and anthropogenic stress index for large river species



6.4.4 *Restoration and protection priorities*

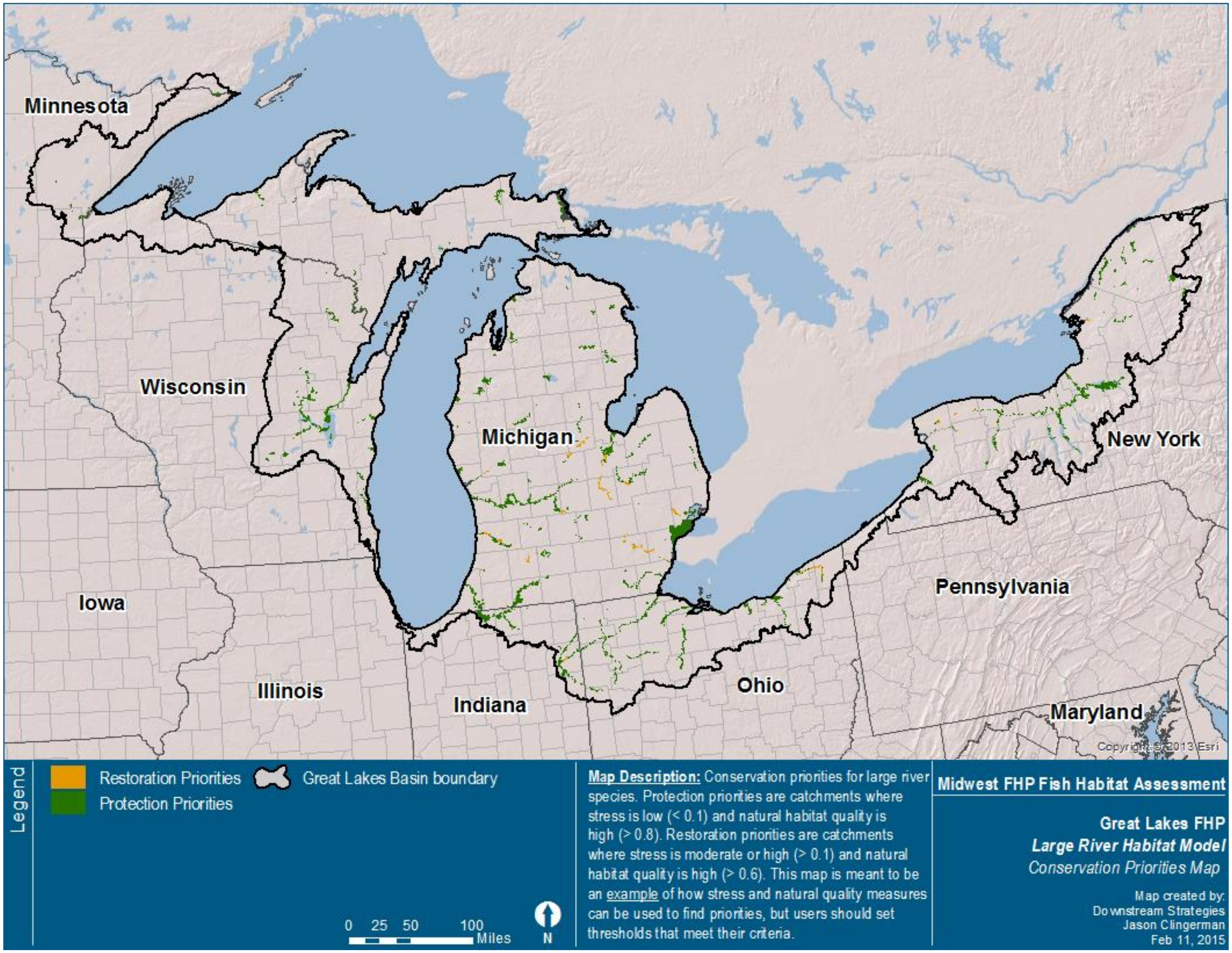
A plot of HQI versus ASI values for all catchments in the study area can be used as a reference to define HQI and ASI thresholds when evaluating restoration and protection priorities (Figure 45). In the example shown, thresholds for protection priorities were defined as catchments with high natural habitat quality and low anthropogenic stress; these thresholds were based on HQI greater than 0.8 and ASI less than 0.1. The thresholds used to identify restoration priorities were defined as catchments with moderate to high anthropogenic stress and moderate to high natural quality; this threshold was based on HQI greater than 0.6 and ASI greater than 0.1. These classifications are mapped in Figure 46. These thresholds were solely based on the relative scores for natural quality and stress indices. Though this example scenario provides an informed set of criteria for identifying conservation priorities, it is only intended to demonstrate the functionality of querying catchments based on these attributes to identify areas that meet user-defined criteria to guide conservation, protection, and restoration planning.

Figure 45: HQI versus ASI values for all catchments for large river species



Note: Breakpoints for HQI and ASI classes in this example are denoted by dashed lines. The arrows indicate the directions of increasing potential protection (green arrow) or restoration (red arrow) priority. The red box in the upper right indicates catchments defined as restoration priorities under the example scenario. The green box in the upper left indicates catchments defined as protection priorities under the same scenario.

Figure 46: Restoration and protection priorities for large river species



7. LIMITATIONS AND SUGGESTIONS FOR FUTURE WORK

In general, while the estimates of probability of presence, index scores, HQI, and ASI generated through this assessment represent a useful and objective means for assessing aquatic habitat and prioritizing habitats for restoration or protection, there are some limitations that are important to consider. Results generated through the modeling process are ultimately limited by the quality of data used to generate them. In the future, the model can be improved by improving the resolution and precision of the data. For example, some county-level data were used as predictor variables although the data likely generalize conditions at the catchment scale. In some cases, this resulted in generalizations in ASI or in the individual ASI metrics, which is evidenced by the unnatural hard break lines at some county boundaries. Although these variables—such as network cattle density and network surface water consumption—were limited in spatial resolution, they still had high relative influence in the BRT model and were important to retain for predictive performance. In the future, refinement of these county-level variables or inclusion of higher resolution surrogates could improve both the precision of the BRT model predictions and post-modeling indices.

A second limitation is that the data and maps represent only a snapshot in time. Therefore, the models may not represent conditions before or after the data were collected or created. For example, any habitat lost or gained due to increased impervious surface cover since the 2006 National Land Cover Database (NLCD) was not considered in this assessment. Similarly, a portion of the uncertainty can be attributable to the temporal mismatches between the fish collection data and landscape data. As such, improving the temporal match between those datasets for future work would be beneficial.

While continuous response variables can be modeled, binomial response variables can generally be modeled with greater precision in cases where the response data vary in collection method or date. Throughout this assessment, we have generally found that binomial (i.e., presence-absence) response variable models have performed better than continuous (i.e., abundance-based) variable models. In the future, basing diversity metrics on the presence-absence of targeted species, rather than relative abundance, may improve their precision.

There were also a few important issues that were beyond the scope of this project. Acid precipitation, biological interactions, and local habitat variation are all important in structuring fish communities. These variables were not directly used as predictor variables, although, when possible, surrogates were used to approximate variation in the model resulting from these processes.

Local habitat measures such as water quality (pH, alkalinity, instream temperature), physical habitat complexity, and substrate size are examples of local measures important to structuring fish communities. These measures could not be directly quantified in this analysis given the scope and scale of the project. However, since each catchment's land cover and geology was included in the analysis, some aspects of water quality were indirectly modeled. Likewise, habitat complexity and substrate size could be partially captured by the combination of stream slope and bedrock and surficial geology. Nonetheless, exclusion of detailed local measures likely accounts for some uncertainty in the model results. Thus, the results from this analysis should be combined with local expert knowledge and additional field data to arrive at the most accurate representation of habitat conditions.

In addition, inclusion of biological interactions in future models could improve the precision of the model and the ability to quantify its influence on the response variables. For this assessment, information relating to Asian carp was pursued to use as a predictor variable, but unfortunately adequate data did not exist prior to the initiation of the modeling process. The biological interactions between non-native species such as Asian carp, brown trout, and other exotic aquatic species may account for some local variability in model results that were beyond the scope of this project.

Finally, the effects of climate change were not considered in this analysis, such as altered thermal and stream flow regimes, and physical habitat. Particularly for coldwater species, such as trout, future warming could result in increased population isolation due to confinement to headwater habitats or more localized thermal refugia. Specifically, identifying catchments vulnerable to climate change for a particular species could represent an important and supplementary next step in the identification of restoration and protection priorities for targeted aquatic populations.

REFERENCES

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77: 802-13.

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of Statistics 29: 1189-1232.

Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. Statistics in Medicine 22: 1365-81.

Appendix A: DATA DICTIONARY

Field Name	Description	Source
Comid	catchment comid (unique identifier)	NHDPlus
HUC8	8 digit Hydrologic Unit Code	Midwest Fish Habitat Assessment Project
HUC12	12 digit Hydrologic Unit Code (NRCS WBD)	Midwest Fish Habitat Assessment Project
HUC12_Name	12 digit Hydrologic Unit Code Name (NRCS WBD)	Midwest Fish Habitat Assessment Project
Grid_code		NHDPlus
Grid_count	Number of cells in catchment grid, 30m	NHDPlus
Prod_unit	NHDPlus production unit (subdivides the region)	NHDPlus
Areasqkm	area of catchment, sq km	NHDPlus
WARNINGS	my warnings on potential problems with catchment for modeling	Jackie
geol_maj	geology, texture, majority type within catchment (coded value)	Sally's surficial geology dataset (see notes for values)
dam_inshed	dam(s) present in catchment (yes/no)	Sally's dam dataset
dam_count	number of dam(s) present in catchment	Sally's dam dataset
wet_sqkm	area of mapped wetland polygons in catchment, sq km	Sally's wetlands dataset
dam_countC	number of dam(s) in catchment and upstream area	Sally's dam dataset
wet_sqkmC	area of mapped wetland polygons in upstream area, sq km	Sally's wetlands dataset
lu_devA	calculated developed land use, area (sq km), catchment	Great Lakes land cover (2001)
lu_devP	calculated developed land use, percent, catchment	Great Lakes land cover (2001)
lu_agA	calculated agricultural land use, area (sq km), catchment	Great Lakes land cover (2001)
lu_agP	calculated agricultural land use, percent, catchment	Great Lakes land cover (2001)
lu_forA	calculated forested land use, area (sq km), catchment	Great Lakes land cover (2001)

lu_forP	calculated forested land use, percent, catchment	Great Lakes land cover (2001)
lu_wetA	calculated wetland land use, area (sq km), catchment	Great Lakes land cover (2001)
lu_wetP	calculated wetland land use, percent, catchment	Great Lakes land cover (2001)
Soil0a	Revised soil hydrologic group code 0 (Canada), area (sq km), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil0p	Revised soil hydrologic group code 0 (Canada), area (%), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil1a	Revised soil hydrologic group code 1 (A), area (sq km), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil1p	Revised soil hydrologic group code 1 (A), area (%), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil2a	Revised soil hydrologic group code 2 (B), area (sq km), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil2p	Revised soil hydrologic group code 2 (B), area (%), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil3a	Revised soil hydrologic group code 3 (C), area (sq km), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil3p	Revised soil hydrologic group code 3 (C), area (%), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil4a	Revised soil hydrologic group code 4 (D), area (sq km), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil4p	Revised soil hydrologic group code 4 (D), area (%), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil5a	Revised soil hydrologic group code 5 (urban areas/water), area (sq km), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil5p	Revised soil hydrologic group code 5 (urban areas/water), area (%), catchment	Revised soil hydrologic group dataset from Sally (based on STATSGO)

Soil0ac	Revised soil hydrologic group code 0 (Canada), area (sq km), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil0pc	Revised soil hydrologic group code 0 (Canada), area (%), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil1ac	Revised soil hydrologic group code 1 (A), area (sq km), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil1pc	Revised soil hydrologic group code 1 (A), area (%), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil2ac	Revised soil hydrologic group code 2 (B), area (sq km), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil2pc	Revised soil hydrologic group code 2 (B), area (%), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil3ac	Revised soil hydrologic group code 3 (C), area (sq km), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil3pc	Revised soil hydrologic group code 3 (C), area (%), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil4ac	Revised soil hydrologic group code 4 (D), area (sq km), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil4pc	Revised soil hydrologic group code 4 (D), area (%), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil5ac	Revised soil hydrologic group code 5 (urban areas/water), area (sq km), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
Soil5pc	Revised soil hydrologic group code 5 (urban areas/water), area (%), upstream cumulative	Revised soil hydrologic group dataset from Sally (based on STATSGO)
lu_devAC	calculated developed land use, area (sq km), upstream cumulative	Great Lakes land cover (2001)
lu_devPC	calculated developed land use, percent, upstream cumulative	Great Lakes land cover (2001)
lu_agAC	calculated agriculture land use, area (sq km), upstream cumulative	Great Lakes land cover (2001)

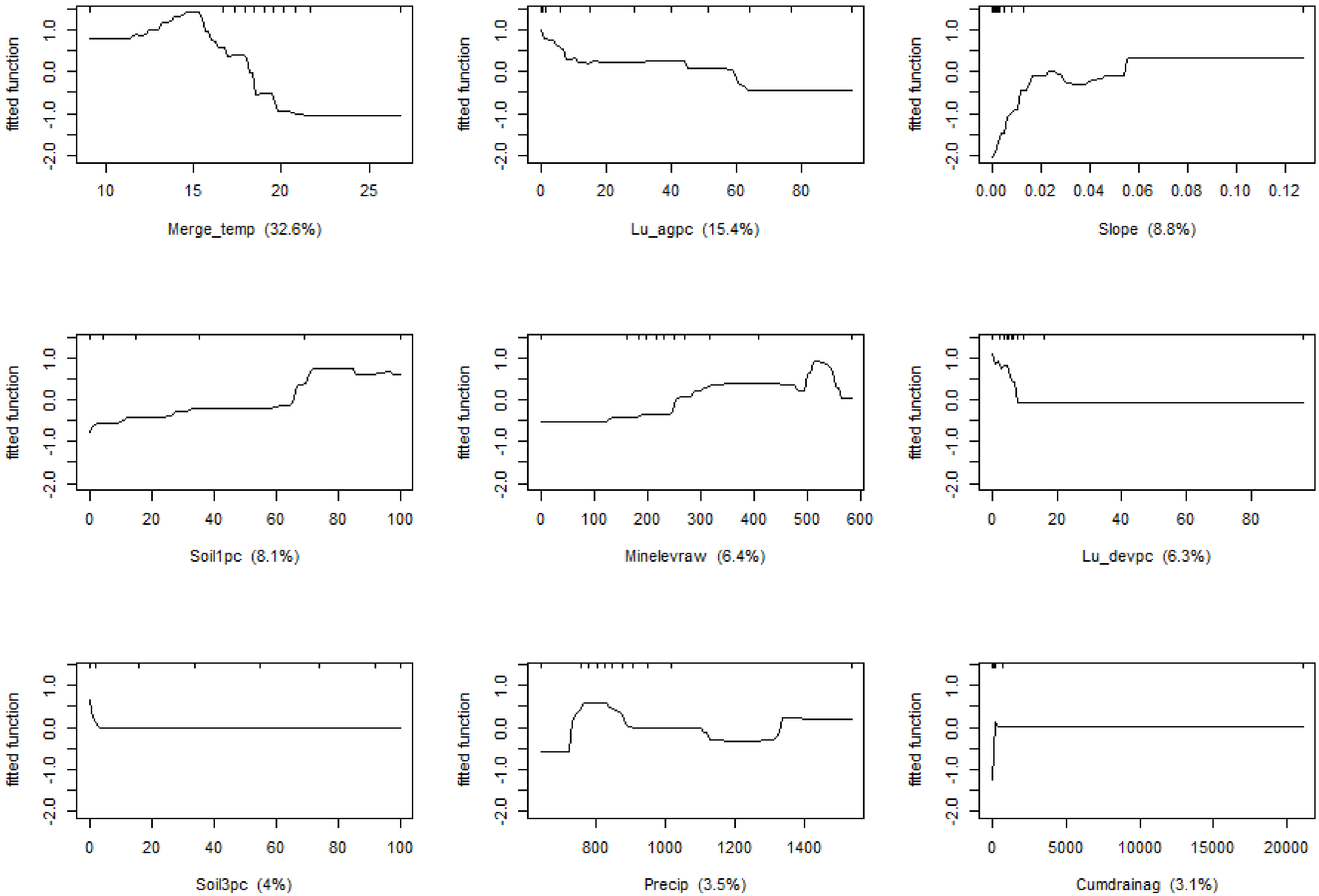
lu_agPC	calculated agriculture land use, percent, upstream cumulative	Great Lakes land cover (2001)
lu_forAC	calculated forested land use, area (sq km), upstream cumulative	Great Lakes land cover (2001)
lu_forPC	calculated forested land use, percent, upstream cumulative	Great Lakes land cover (2001)
lu_wetAC	calculated wetland land use, area (sq km), upstream cumulative	Great Lakes land cover (2001)
lu_wetPC	calculated wetland land use, percent, upstream cumulative	Great Lakes land cover (2001)
ROADCR	LOCAL: Census 2000 TIGER Roads, 1:100K scale, road crossings identified by INTERSECT, with points generated, #/km2	local_disturbance_variables.dbf
ROADLEN	LOCAL: Census 2000 TIGER Roads, 1:100K scale, units not given - m/km2	local_disturbance_variables.dbf
MINES	LOCAL: USGS Active Mines and Mineral Processing Plants, 2003, #/km2	local_disturbance_variables.dbf
ROADCRC	NETWORK: Census 2000 TIGER Roads, 1:100K scale, road crossings identified by INTERSECT, with points generated, #/km2	network_disturbance_variables.dbf
ROADLENC	NETWORK: Census 2000 TIGER Roads, 1:100K scale, units not given - m/km2	network_disturbance_variables.dbf
MINESC	NETWORK: USGS Active Mines and Mineral Processing Plants, 2003, #/km2	network_disturbance_variables.dbf
IMPERVS	LOCAL: Impervious surface area (allocation per segment): area (km2)	2001 Impervious Surface Area
IMPERVSC	NETWORK: Impervious surface area (accumulation of upstream segments): total upstream area (km2)	2001 Impervious Surface Area
CATCHTYPE	Catchment flowline feature type (flowline and waterbody/area combined)	based on NHD
GAP_TEMP	GAP regional temperature regime	Regional Aquatic GAP
Areasqkmc	Total area upstream (cumulative) sq km	Calculated (GLFHP project)
Eco_code3	Ecoregion code (majority), Level III, catchment	US EPA Omernik Ecoregions for North America, Level III
Water_gw	LOCAL: USGS National Atlas of the US: Ground Water Use by COUNTY 2000: Millions gallons per day/km2	local_disturbance_variables.dbf
Water_sw	LOCAL: USGS National Atlas of the US: Surface Water Use by COUNTY	local_disturbance_variables.dbf

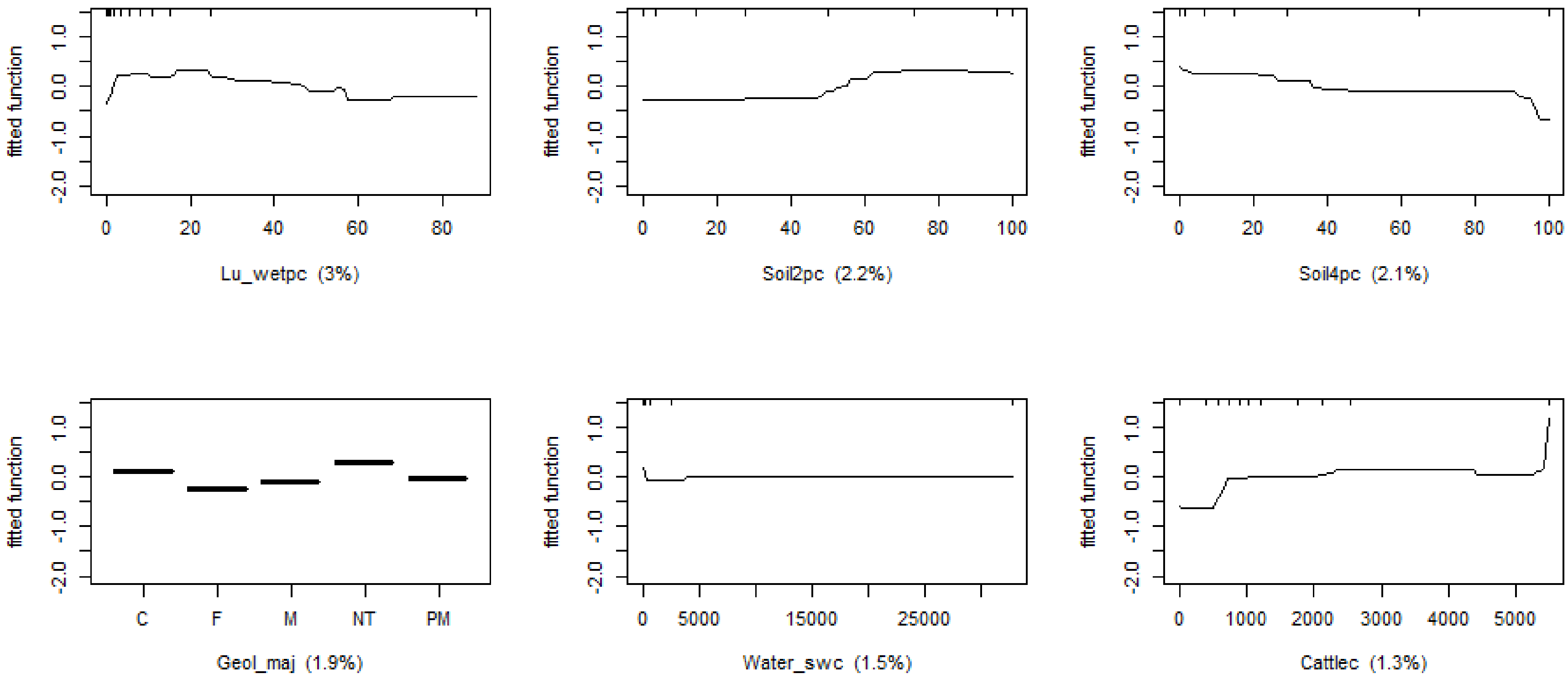
	2000: Millions gallons per day/km2	
Cattle	LOCAL: Agricultural Census 2002, 1:2M scale, INTEGER: average number of cattle/acre farmland	local_disturbance_variables.dbf
Water_gwc	NETWORK: USGS National Atlas of the US: Ground Water Use by COUNTY 2000: Millions gallons per day/km2	network_disturbance_variables.dbf
Water_swc	NETWORK: USGS National Atlas of the US: Surface Water Use by COUNTY 2000: Millions gallons per day/km2	network_disturbance_variables.dbf
Cattlec	NETWORK: Agricultural Census 2002, 1:2M scale, INTEGER: average number of cattle/acre farmland	network_disturbance_variables.dbf
Minelevraw	Minimum elevation (unsmoothed) in meters	catchmentattributesflow.dbf
Slope	Slope of flowline (cm/cm)	catchmentattributesflow.dbf
Precip	Mean annual precipitation in mm	catchmentattributestempprecip.dbf
Temp	Mean annual temperature in degrees centigrade * 10	catchmentattributestempprecip.dbf
Stream_temp	Modeled stream temperture, degrees C	GLB FHP
hJXnow	Modeled stream temperature, degrees C	USGS/GLBFHP
Merge_temp	Modeled stream temperature, degrees C (newer hJXnow values where available, otherwise, Stream_temp values were used)	USGS/GLBFHP
NLCD06DevA	NLCD 2006 Developed land cover classes (21, 22, 23, 24), area (sq km), catchment	NLCD 2006
NLCD06DevP	NLCD 2006 Developed land cover classes (21, 22, 23, 24), area (%), catchment	NLCD 2006
NLCD06ForA	NLCD 2006 Forested land cover classes (41, 42, 43), area (sq km), catchment	NLCD 2006
NLCD06ForP	NLCD 2006 Forested land cover classes (41, 42, 43), area (%), catchment	NLCD 2006
NLCD06AgA	NLCD 2006 Agriculture land cover classes (81, 82), area (sq km), catchment	NLCD 2006

NLCD06AgP	NLCD 2006 Agriculture land cover classes (81, 82), area (%), catchment	NLCD 2006
NLCD06WetA	NLCD 2006 Wetland land cover classes (90, 95), area (sq km), catchment	NLCD 2006
NLCD06WetP	NLCD 2006 Wetland land cover classes (90, 95), area (%), catchment	NLCD 2006
NLCD06DevAC	NLCD 2006 Developed land cover classes (21, 22, 23, 24), area (sq km), upstream cumulative	NLCD 2006
NLCD06DevPC	NLCD 2006 Developed land cover classes (21, 22, 23, 24), area (%), upstream cumulative	NLCD 2006
NLCD06ForAC	NLCD 2006 Forested land cover classes (41, 42, 43), area (sq km), upstream cumulative	NLCD 2006
NLCD06ForPC	NLCD 2006 Forested land cover classes (41, 42, 43), area (%), upstream cumulative	NLCD 2006
NLCD06AgAC	NLCD 2006 Agriculture land cover classes (81, 82), area (sq km), upstream cumulative	NLCD 2006
NLCD06AgPC	NLCD 2006 Agriculture land cover classes (81, 82), area (%), upstream cumulative	NLCD 2006
NLCD06WetAC	NLCD 2006 Wetland land cover classes (90, 95), area (sq km), upstream cumulative	NLCD 2006
NLCD06WetPC	NLCD 2006 Wetland land cover classes (90, 95), area (%), upstream cumulative	NLCD 2006

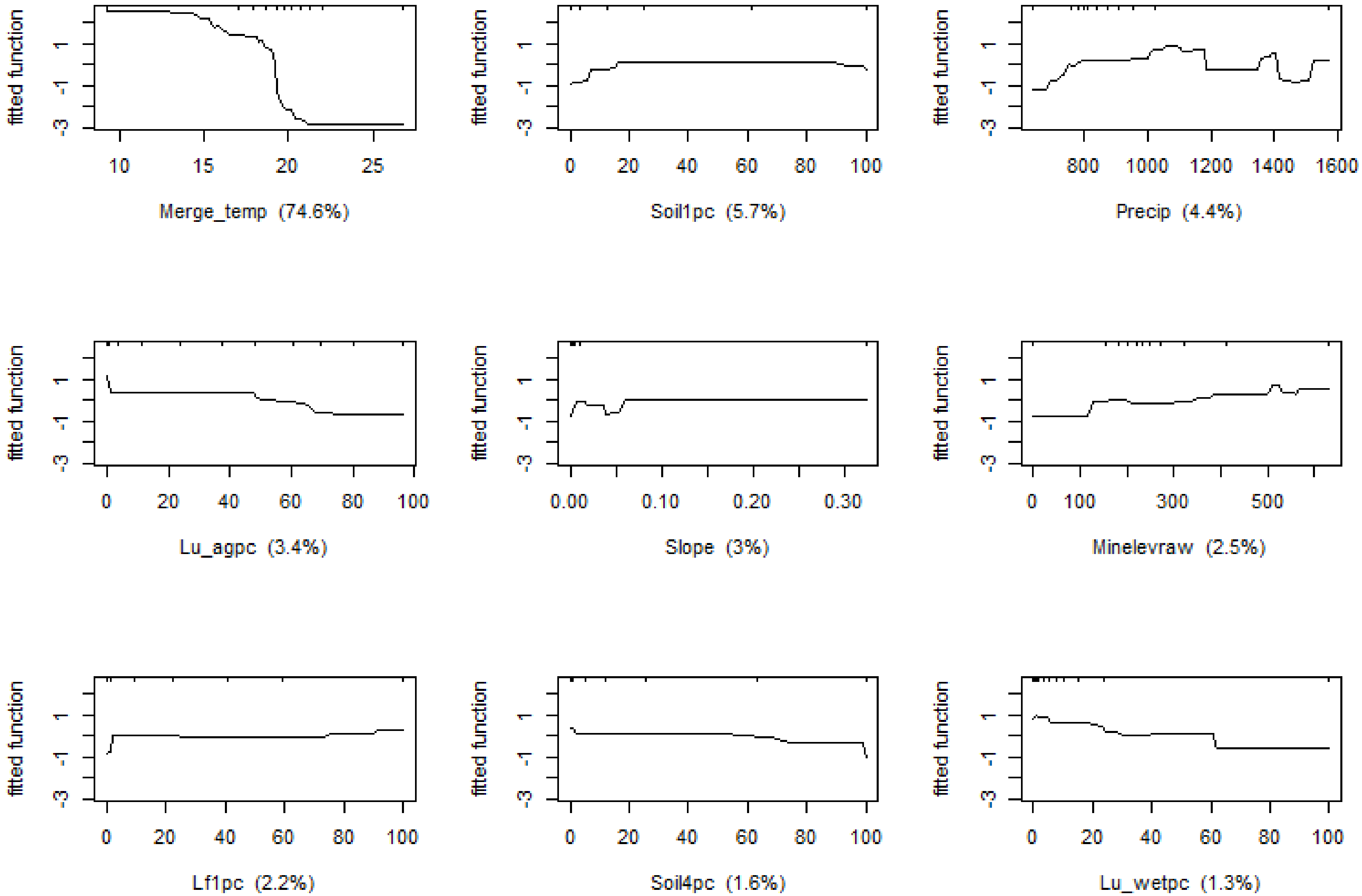
Appendix B: FUNCTIONAL RESPONSE PLOTS

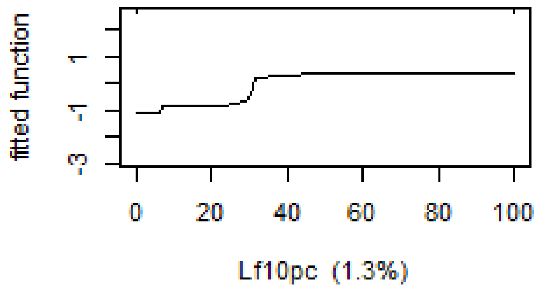
Brook trout



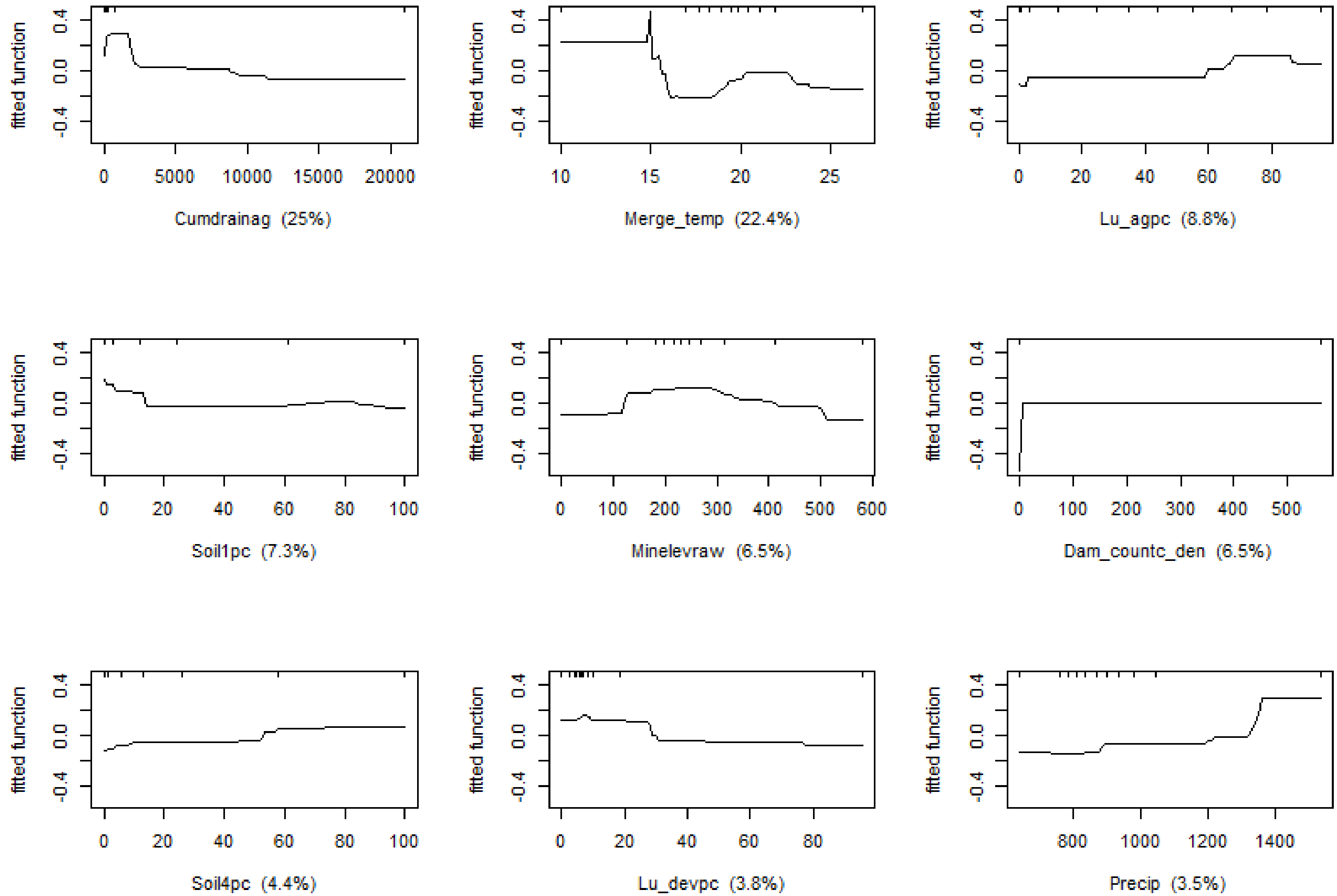

fitted function
Lu_wetpc (3%)
Soil2pc (2.2%)
Soil4pc (2.1%)
Geol_maj (1.9%)
C
F
M
NT
PM
Water_swc (1.5%)
Cattlec (1.3%)

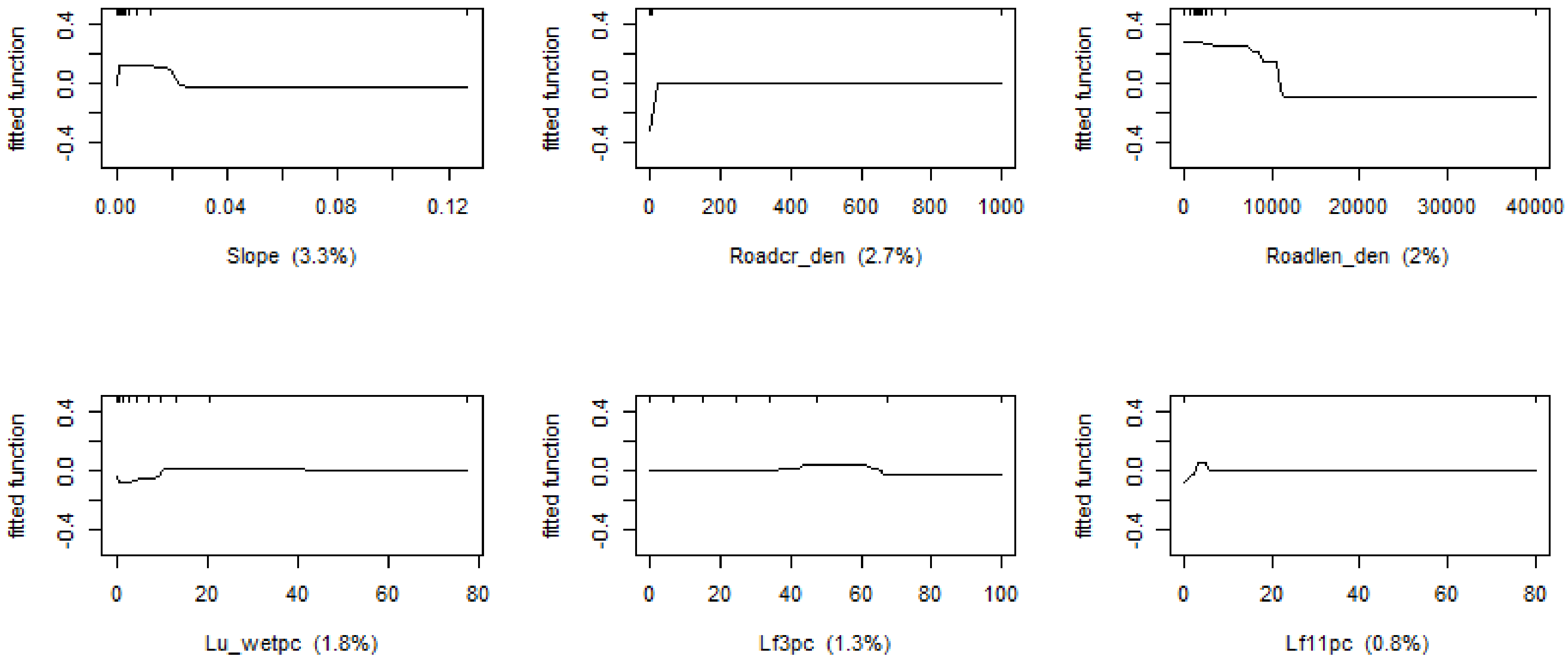
Coldwater species



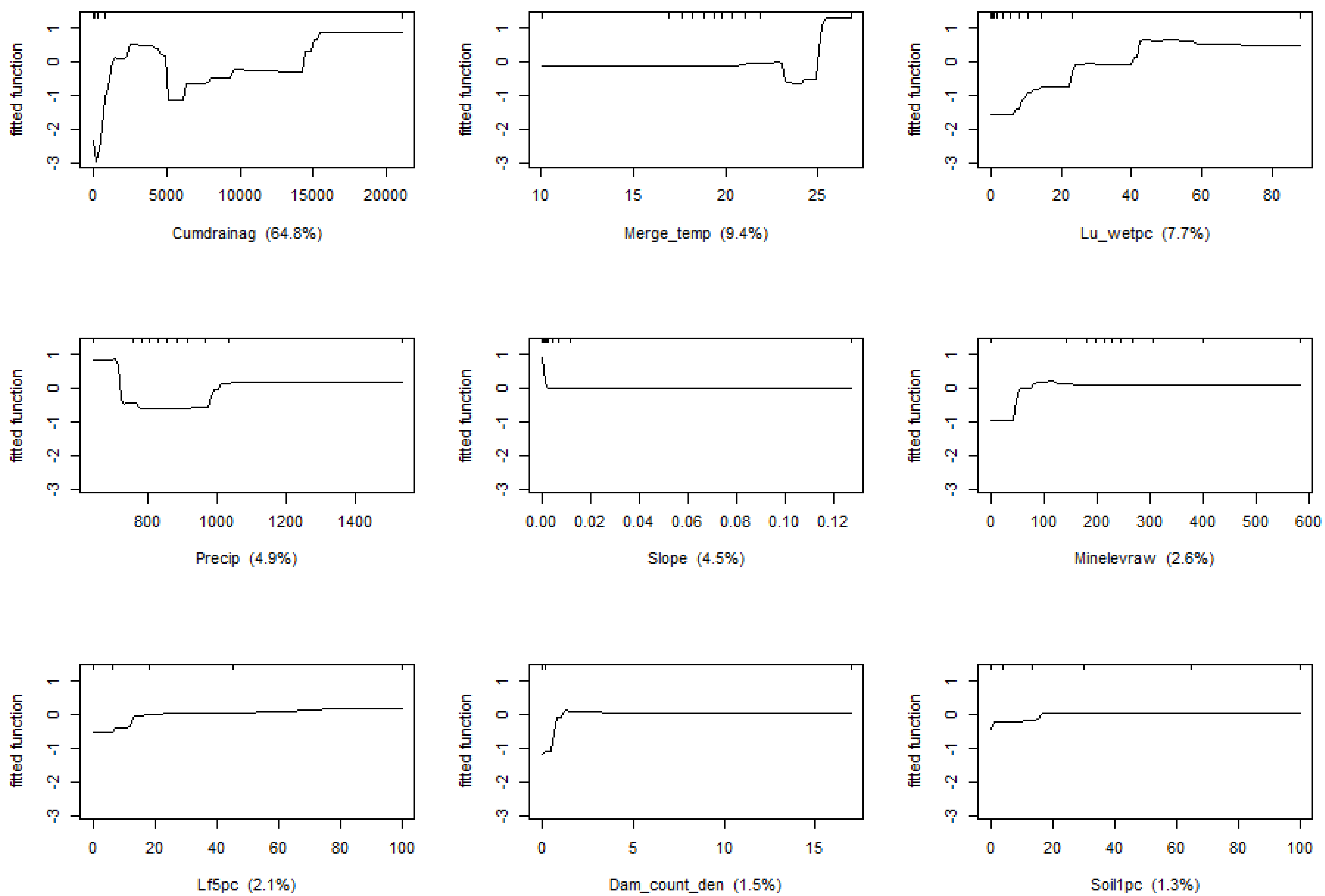

fitted function
-3
-1
1
0
20
40
60
80
100
Lf10pc (1.3%)

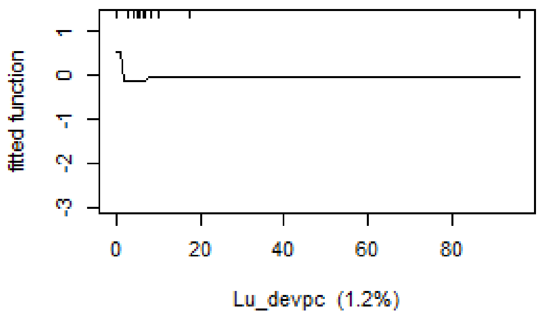
Lithophilic species richness



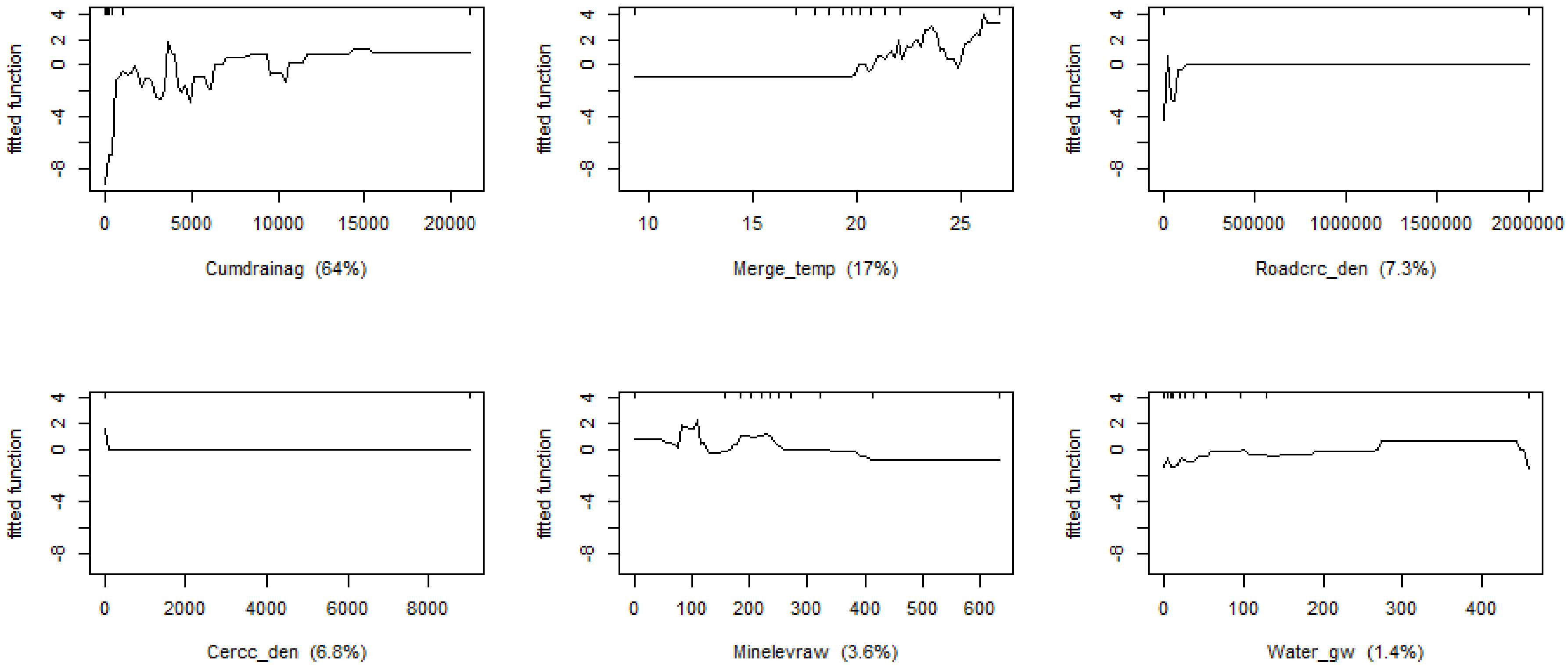

fitted function
-0.4
0.0
0.4
0.00
0.04
0.08
0.12
Slope (3.3%)
0
200
400
600
800
1000
Roadcr_den (2.7%)
0
10000
20000
30000
40000
Roadlen_den (2%)
0
20
40
60
80
Lu_wetpc (1.8%)
100
Lf3pc (1.3%)
Lf11pc (0.8%)

Walleye


fitted function
Cumdrainag (64.8%)
Merge_temp (9.4%)
Lu_wetpc (7.7%)
Precip (4.9%)
Slope (4.5%)
Minelevraw (2.6%)
Lf5pc (2.1%)
Dam_count_den (1.5%)
Soil1pc (1.3%)


fitted function
-3
-2
-1
0
1
0
20
40
60
80
Lu_devpc (1.2%)

Large river species



[i] http://www2.research.att.com/~phillips/pdf/Elith_et_al_ecography.pdf

[ii] http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2656.2008.01390.x/pdf