Case Study:

Analysis of scale on boosted regression tree fish habitat models

Date: 4-2-2013

Authors:

Jason Clingerman, Aquatic Ecologist, Downstream Strategies

J. Todd Petty, PhD, Professor of Wildlife and Fisheries Resources, West Virginia University

Fritz Boettner, Principal, Downstream Strategies

Downstream Strategies

building capacity for sustainability

1. INTRODUCTION

Downstream Strategies (DS) has produced predictive models for several fish habitat partnerships (FHPs) across the United States. These models utilized widely available landscape variables as predictors for instream aquatic responses, such as presence of certain guilds or species of fish. Boosted regression tree (BRT) models were chosen as the predictive statistical models for these analyses after careful consideration of their strengths and weaknesses compared with other available statistical methods. These models created a broad and unique understanding of the link between terrestrial and aquatic health and allowed for the determination of stressors for each response.

Thus far, the models have been built at broad scales that encompassed thousands of square miles and stretched across many states. These large-scale models offered valuable insight into which landscape-level stressors and natural conditions were structuring aquatic responses. However, the determination of more local-level stressors proved problematic as the broad patterns overshadowed those stressors that may structure aquatic responses at finer scales.

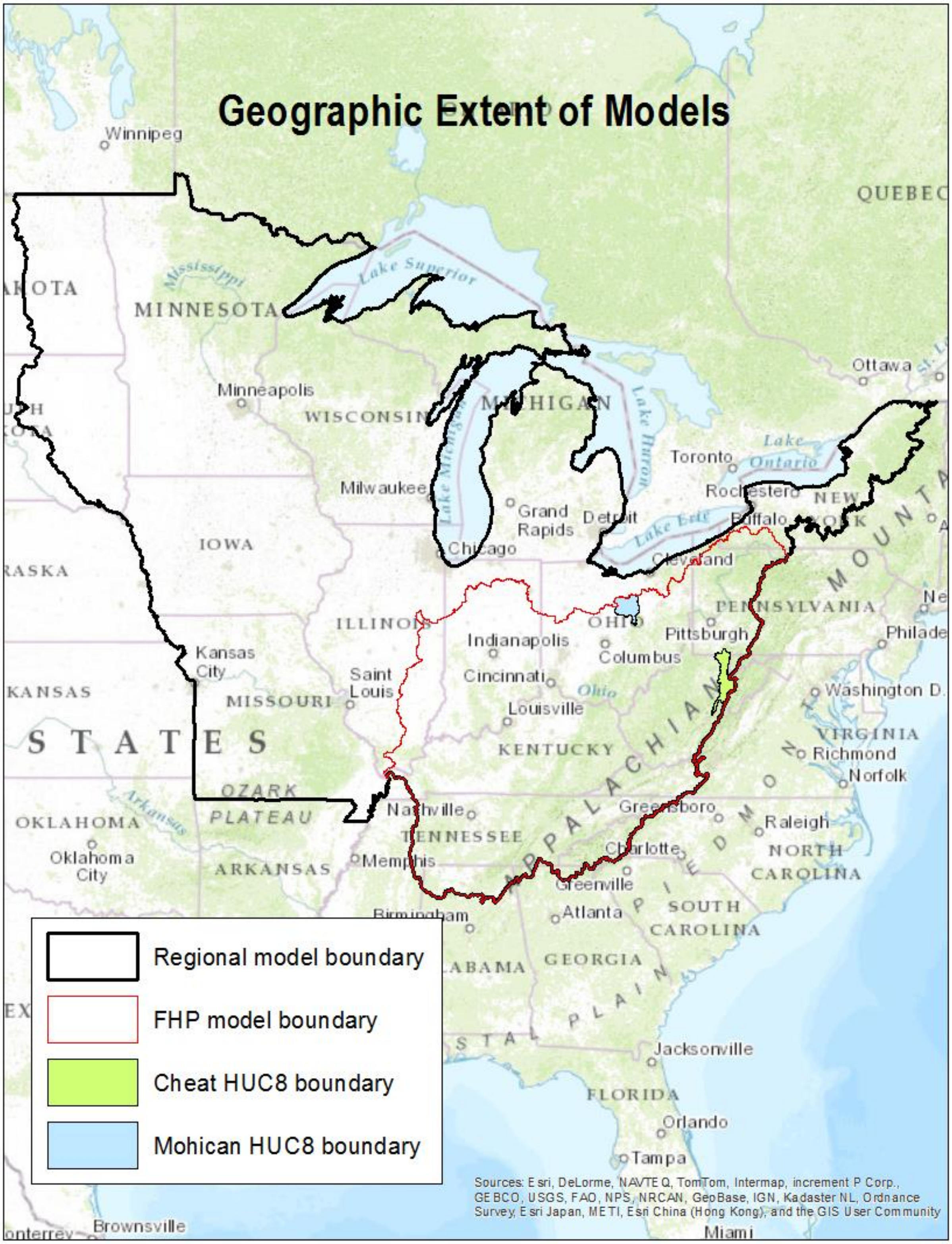
Recent modeling efforts at the regional and FHP scale have indicated that smaller-scale models are likely necessary to pinpoint localized stressors. From discussions with experienced modelers and fishery professionals, HUC8 watersheds were agreed upon as the most appropriate scale.

This report summarizes a case study that demonstrates the effect of scale on the assignment of stressors from predictive BRT models. Specifically, we modeled the same response at three different scales and for two separate HUC8 watersheds. Modeling at different scales demonstrates the change in dominant predictors at differing spatial scales. The analysis of two separate HUC8 watersheds indicates the change in dominant predictors between separate geographies at the same spatial scale.

2. METHODOLOGY

We built models using available data for each geographic extent (Figure 1). The regional-scale model included a very broad geographic area that covered 18 states across the Midwest. The FHP-scale model included the entire Ohio River Basin and was nested within the regional model. This model encompassed a smaller geographic extent than the regional model, yet still covered portions of 14 states. There were two separate HUC8-scale models. The first was the Cheat HUC8 watershed, which is a relatively high-elevation watershed in north-central West Virginia. The second was the Mohican HUC8 watershed in north-central Ohio.

Figure 1: Geographic extent of model boundaries



Geographic Extent of Models

Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

The HUC8 watersheds were chosen because both had similar response variable structures. Each of these HUC8 watersheds had a high percentage of presences in the available sample data, but were perceived to be dissimilar in other regards (land cover, elevation, stream network pattern, stressors). These HUC8 watersheds were chosen to demonstrate that the same response can be structured by different landscape variables in different geographic locations, because individual stressors or combinations of stressors vary spatially.

Each model had the same response variable (coldwater guild as defined by the United States Fish and Wildlife Service (USFWS) regional assessment) and the same suite of predictor variables. Correlation analysis was done for each model to remove redundant variables. Once redundant variables were removed, a preliminary model was created. We used the results of this preliminary model to eliminate any additional predictor variables that were of very low relative influence (relative influence < 1.0). This methodology of removing irrelevant predictors was used during previous modeling applications and was found to improve interpretability. Further, it only negligibly reduced cross-validated performance.

After removing redundant and irrelevant predictors, the final model was created for each geographic extent. Cross-validated correlation statistics were examined to analyze model strength. The BRT model also produced an output table of relative influences of each predictor variables. The ranking and relative influences of predictors were then compared among spatial scales and between the two HUC8-scale models.

Results for each model were extrapolated to all catchments. These mapped responses were compared among models and the spatial patterns of predictions were compared to known sample locations.

3. RESULTS

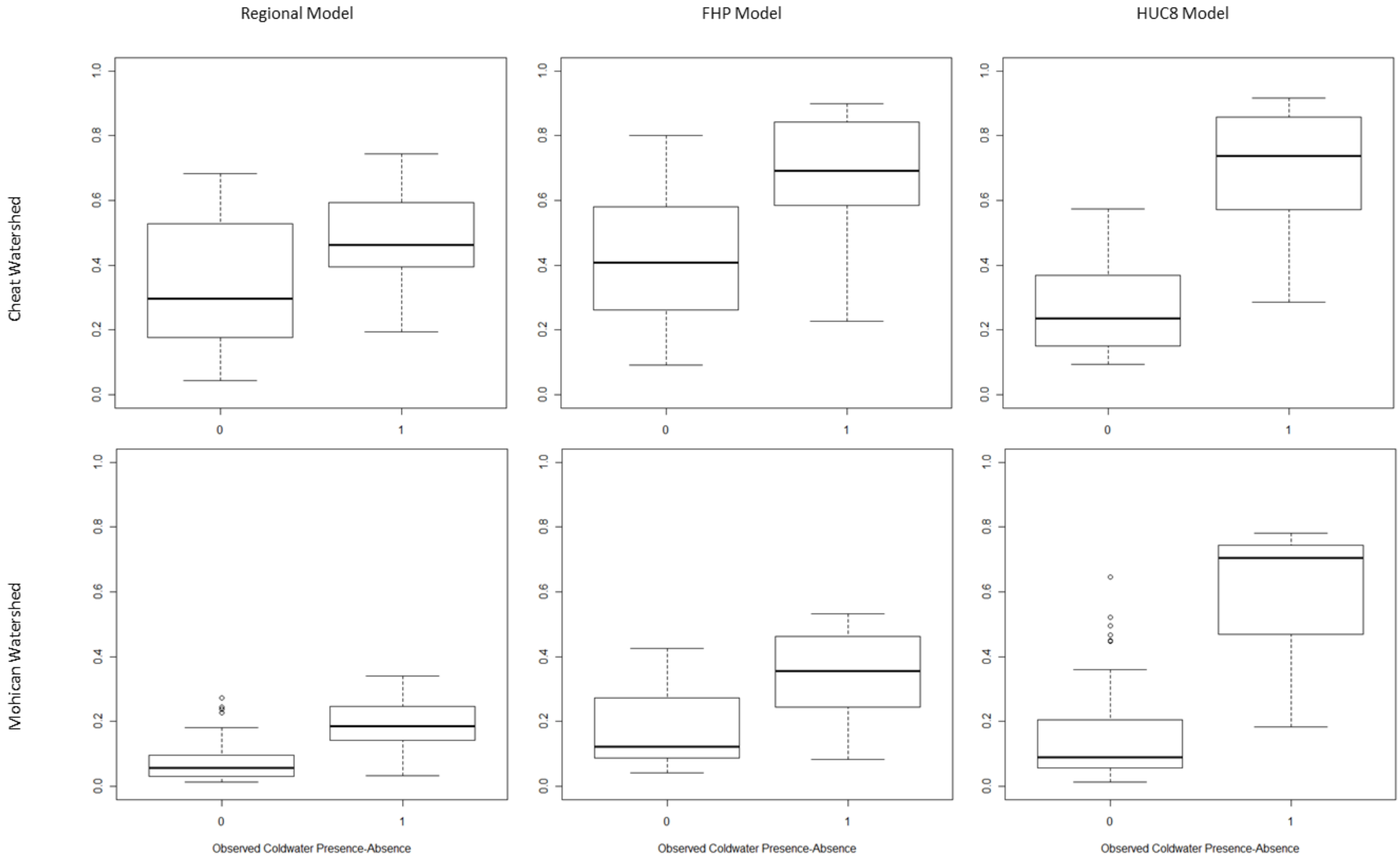
Cross-validation statistics indicated that predictive capacity differed amongst models (Table 1). Cross-validated correlation and receiver operating characteristic ROC scores were lowest in the Cheat HUC8 model, likely from low sample size (N = 51), but none of the cross-validated statistics fell outside of the acceptable range .

Table 1: Cross-validated statistics for each model

Statistic	Regional	ORB	Cheat	Mohican
Sample size of response data	18,908	6,048	51	70
Number of predictor variables in final model	10	14	8	6
CV correlation	0.541	0.496	0.455	0.545
CV ROC	0.868	0.852	0.753	0.85
Number of trees in final model	6,200	6,150	950	5,500

While the cross-validated statistics for the regional and ORB models are useful for analyzing the model strength for those entire study areas, they are less useful when examining the model strength for a particular subset of the overall area (the HUC8 watersheds in our study), as model strength can vary spatially. In order to illustrate the ability of each model to correctly predict presences and absences for our focal HUC8 watersheds, we created boxplots for each model and for both HUC8 watersheds. The boxplots illustrate the predicted versus observed conditions of samples only within the indicated HUC8 watershed and indicate that predictions for our focal watersheds tend to be more accurate as spatial scale is reduced.

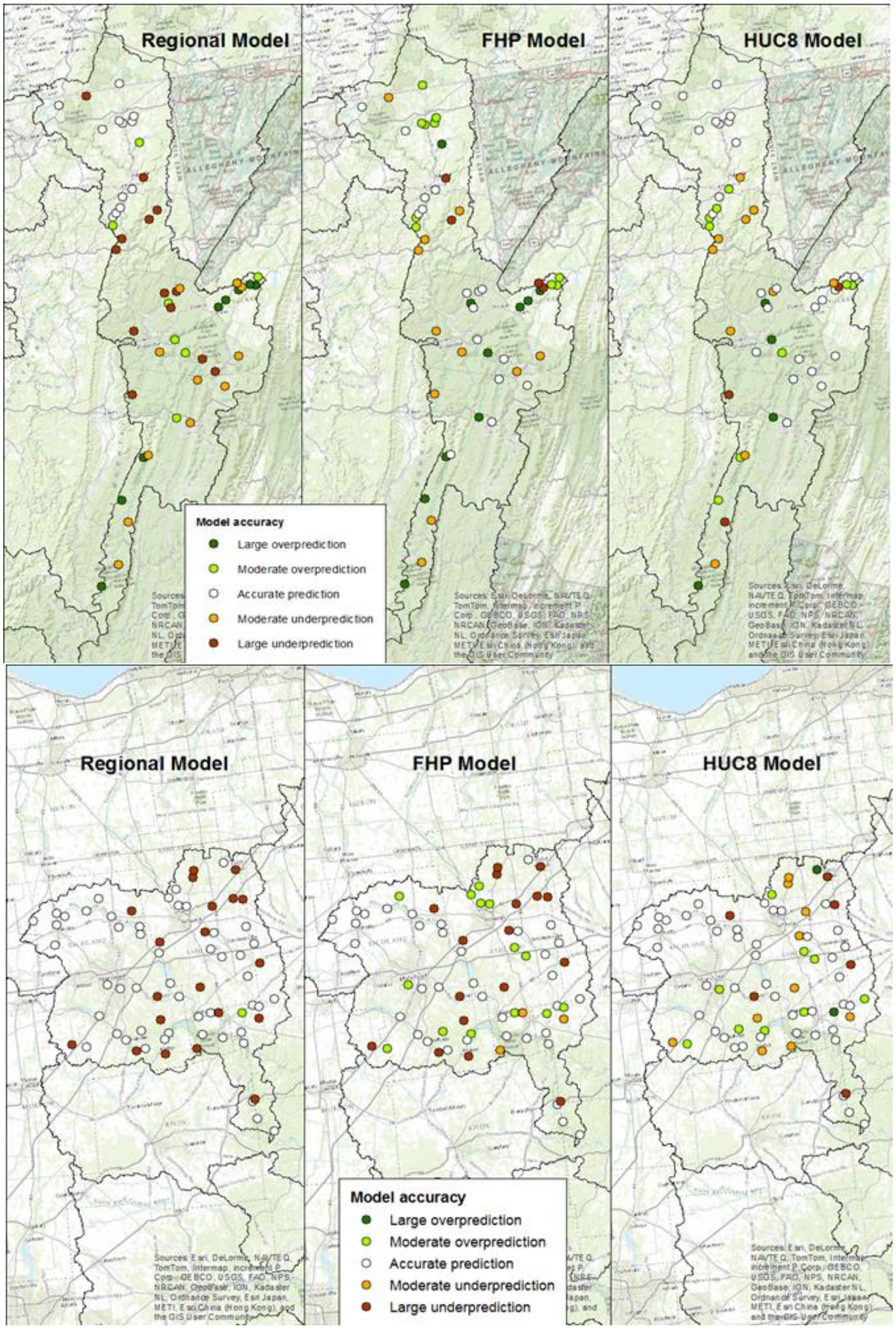
Figure 2: Boxplots of predicted versus observed conditions



Note: The x-axis of each plot indicates actual presences (1) and absences (0). The y-axis of each plot indicated the predicted probability of presence.

We also mapped the accuracy of each model for sample sites within the focal HUC8 watersheds. These maps (Figure 3) indicate that as extent of the model is reduced, the intensity of over/underpredictions is likewise reduced. In addition, the number of sites predicted accurately is increased.

Figure 3: Model accuracy by model scale for focal watersheds.


Regional Model
FHP Model
HUC8 Model
Model accuracy
Large overprediction
Moderate overprediction
Accurate prediction
Moderate underprediction
Large underprediction
ALLEGHENY MOUNTAINS
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community
Regional Model
FHP Model
HUC8 Model
Model accuracy
Large overprediction
Moderate overprediction
Accurate prediction
Moderate underprediction
Large underprediction
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

The predictor variables that most heavily structured the response were different at each geographic scale. Mean annual air temperature, mean annual precipitation, and network mean baseflow were the only predictors that were in the top five most-influential variables for more than two of the four models. The most influential predictor variable for each of the four models was different. Table 2 summarizes the predictor variables for each model and their relative influence.

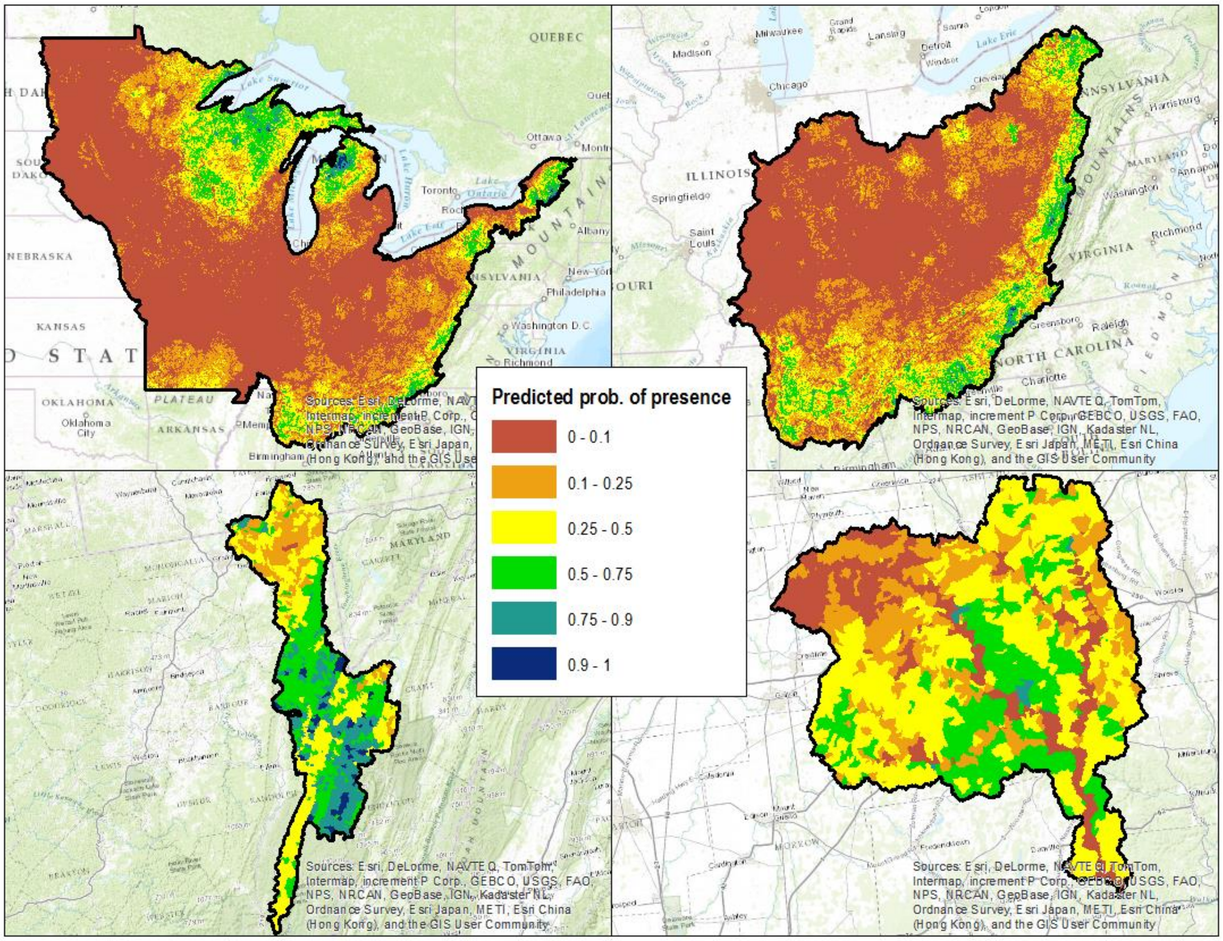
Table 2: Relative influence of predictor variables for each model

Variable	Variable description	Regional	ORB	Cheat	Mohican
AG_PC	Network percent agriculture	4.2	3.4		5.6
AREASQKMC	Network drainage area	4.5	5.3	**9.1**	**57.2**
BAR_PC	Network percent barren land			7.6	
BFI_MEANC	Network mean baseflow	**23.5**	**7.8**		**7.5**
BR1PC	Network percent carbonate bedrock geology		1.1	**27.5**	
BR6PC	Network percent sandstone bedrock geology			**23.7**	
BR7PC	Network percent shale bedrock geology		1.8		
DAMSC_den	Network dam density		1.3		
FOR_P	Catchment percent forest			5.4	**14.9**
IMPSURF_MC	Network mean impervious surface cover	6.8	**6.8**		
MINELEVRAW	Elevation of catchment pourpoint	**10.0**	**29.6**	5.3	
PRECIP	Mean annual precipitation	**9.0**	**13.6**	**13.1**	
ROADCR_den	Catchment road/stream crossing density		1.8		
ROADCRC_den	Network road/stream crossing density		1.3		
ROADLENC_den	Network road density			**8.2**	
SLOPE	Slope of catchment flowline	**17.0**			
TEMP	Mean annual air temperature	**22.1**	**18.6**		**8.4**
WATER_GWC	Network groundwater use	1.5			
WATER_SWC	Network surfacewater use	1.3	5.7		**6.5**
WET_PC	Network percent wetland		1.8		

Note: Variables are in alphabetical order by variable code. Underlined values indicate the variable was one of the top five most influential for that model. Blue highlighted values indicate the most important variable for each model.

The model results were extrapolated to the appropriate geographies (Figure 4

Figure 4: Extrapolated model results



Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

In order to more fully examine the differences between model predictions, we mapped the extrapolated results for the three model scales for the individual HUC8 watersheds. These HUC8-scale maps show the differences in predictions that occured from models built at differing spatial scales. Figure 5 shows the difference for the Cheat watershed, and Figure 6 shows the same comparison for the Mohican watershed. Of note, especially with the Mohican HUC8 watershed, is that presences are generally underpredicted by the larger-scale models and seem to be more accurately predicted by the HUC8-scale model when examining the known sample points and predicted conditions.

Figure 5: Model results comparison for Cheat HUC8

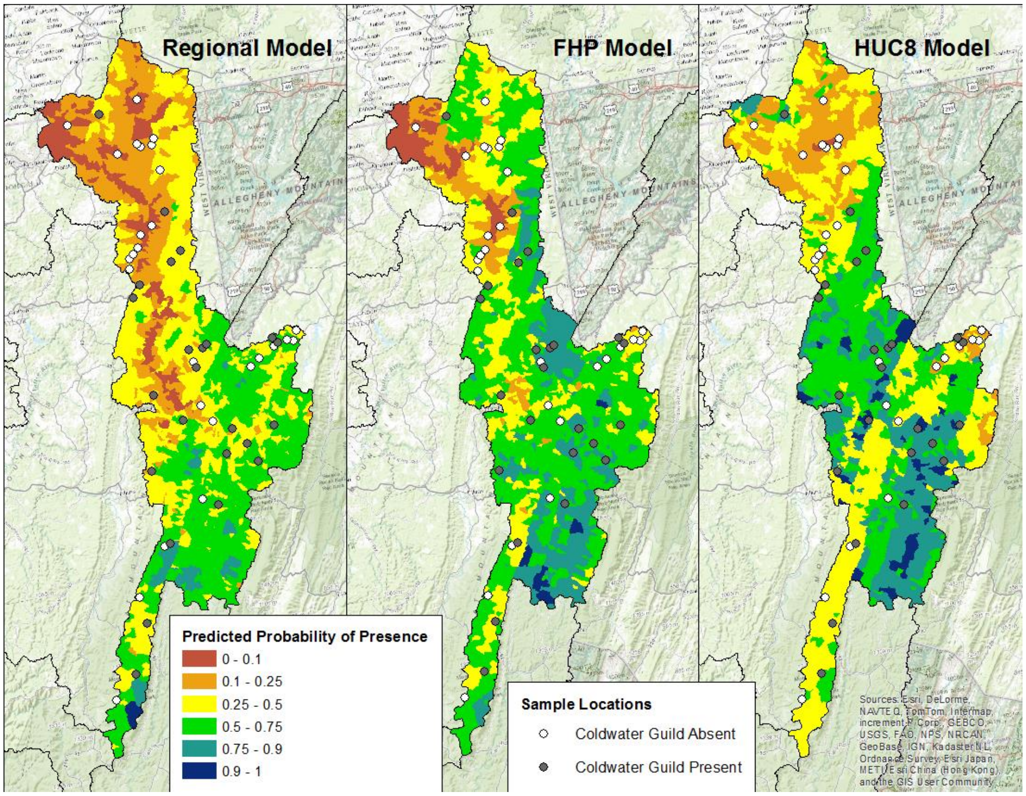
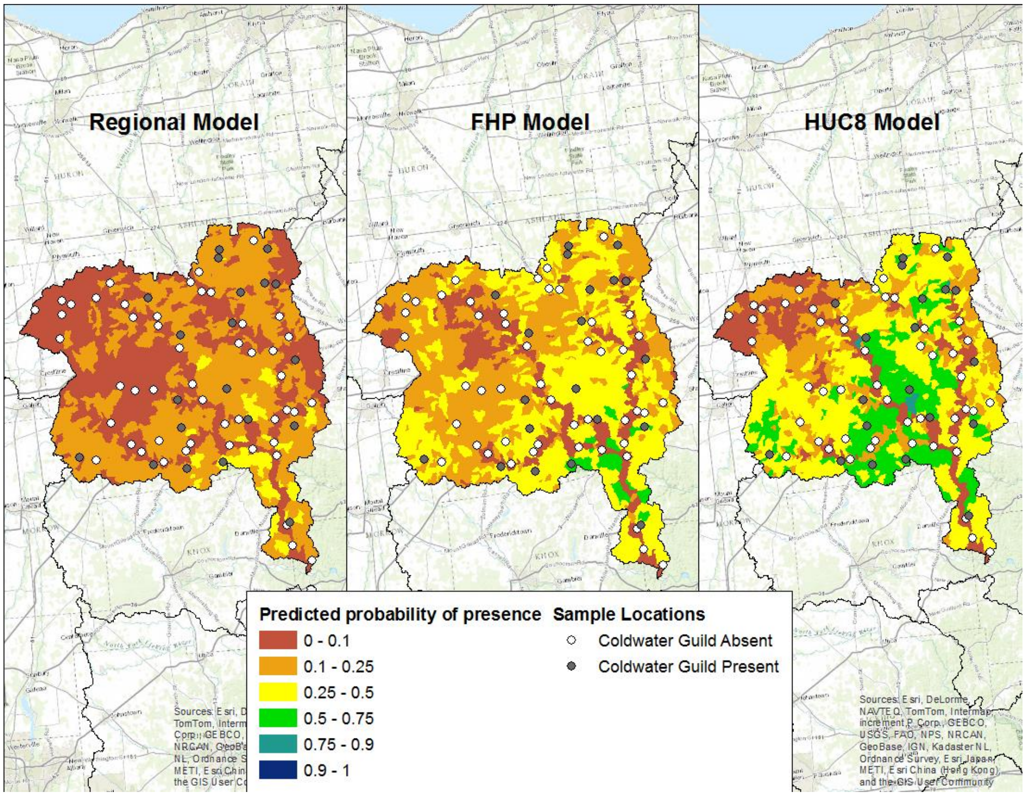

Regional Model
FHP Model
HUC8 Model
ALLEGHENY MOUNTAINS
Predicted Probability of Presence
0 - 0.1
0.1 - 0.25
0.25 - 0.5
0.5 - 0.75
0.75 - 0.9
0.9 - 1
Sample Locations
○ Coldwater Guild Absent
● Coldwater Guild Present
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

Figure 6: Model results comparison for Mohican HUC8


Regional Model
FHP Model
HUC8 Model
Predicted probability of presence
0 - 0.1
0.1 - 0.25
0.25 - 0.5
0.5 - 0.75
0.75 - 0.9
0.9 - 1
Sample Locations
○ Coldwater Guild Absent
● Coldwater Guild Present
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

4. CONCLUSION/DISCUSSION

We have shown that smaller-scale models can be created for smaller geographic areas, given ample response data. The smallest sample size we utilized for an individual model included 51 sites. This amount of data proved to be sufficient for creating a model with a cross-validated ROC and correlation statistic within an acceptable range, but was slightly lower than the models built with more data. This suggests that approximately 50 sample sites could be nearing the minimum required to build adequately robust models for fish habitat assessments.

By analyzing the predictive performance of each model upon only the known sites within each HUC8 watershed of interest, we were able to show that presences and absences were more accurately predicted by more local-scale models. This was shown in the boxplot graph in Figure 2, where predictions more closely matched observations as the scale of the model was reduced.

As evidenced by the results above, predictive models can be heavily influenced by the scale at which they are built. The variables of influence can change drastically depending on the precise geography modeled, even when the response variable remains constant.

The larger-scale models (regional and FHP) seem to be driven by more broad-scale factors that influence stream temperature, as expected. Within the HUC8 scale, the predictor variables of influence were quite different for the two watersheds that were modeled.

For the Cheat HUC8, bedrock geology and precipitation rates were the main driving factors. Within the Cheat watershed, historic and current mining have had large impacts upon stream water quality, and it is likely the bedrock geologies in the model are indicating influences from mining. Acid precipitation is another contributor to water quality issues in the Cheat watershed, and bedrock geology and precipitation rates are likely capturing the acid rain issue.

For the Mohican HUC8, nearly 60% of the model’s relative influence came solely from drainage area, and the next highest predictor value of importance was local forest cover. This indicates that this watershed’s coldwater habitats are structured much more by stream shading than by any other factors (local forest cover and small streams allow for ample stream shading). This is different from the broad climatological patterns and elevation structuring the coldwater guild in the regional and FHP model, and also from the geologic/mining factors prevalent and important in the Cheat drainage.

While not illustrated here because of time constraints, we anticipate changes in the functional relationships between variables and responses at differing scales as well. Since our calculations of stress and natural quality come directly from the functional relationships within the BRT models, we can expect dominant stressors and the relationship of those stressors to change between spatial scales, which will allow for more localized stressors to be indicated from modeling as opposed to broad regional variables/stressors of influence.

These results indicate that choosing the appropriate geographic area to model is critical, and that predictor variables of influence will change dependent upon scale. Smaller-scale models allow for a more “fine-tuned” set of predictors that are most influential within the watershed modeled. Conversely, limited data at finer scales could cause model strength to suffer. Fisheries professionals should couple local knowledge and professional judgment to determine the appropriate scale of analysis based upon landscape characteristics and data availability, but from our results we feel confident that in most situations, given enough sample data, the HUC8 scale seems to be an appropriate scale at which to build models for more precise assignment of stressors.