ATLAS Aquatic Species Habitat Model

Summary Report

February 2022

James spinymussel (*Pleurobema collina*)

Model Purpose

This report summarizes data inputs, methods, and final products of a species habitat model produced for the NCDOT ATLAS project. The models are intended to be used in project management tools to:

1. flag areas of higher versus lower risk of "May Affect" biological conclusions within a species range to improve project planning and management, and
2. add a landscape-scale perspective to improve biologists' field planning and site assessment

Species Summary

This freshwater mussel is found in the upper James and Dan River basins. The species has declined rapidly during the past two decades and now exists only in small, headwater tributaries of the upper James River basin in Virginia and West Virginia. In 2000, it was discovered in the Dan River basin in North Carolina and Virginia. Suitable habitat for this species includes free-flowing streams with a variety of flow regimes. The James spinymussel is found in a variety of substrates that are free from silt. The primary reason for its decline is habitat loss and modification. Threats to this species include siltation, invasion of the non-native Asiatic clam (Corbicula fluminea), impoundment of waterways, water pollution, stream channelization, sewage discharge, agricultural runoff including pesticides and fertilizers, poor logging and road/bridge construction practices, and discharge of chlorine.

Habitat Data

All species' models use occurrence data from the Natural Heritage Program (NHP Tier 2 data) and the NC Wildlife Resource Commission (NCPAWS data). Species occurrence observations identified as historic (extirpated) or older than 2000 are excluded from our models, as are data with low spatial accuracy (e.g., NHP Accuracy "4-Low" or "5-Very Low"). We retained NHP "3-Medium" accuracy observations and labelled these as Low precision and retained NHP 2-High and 1-Very High accuracy observations as High precision data. The occurrence data are presence-only and are not a randomized or representative sample from the species' range. Our models score any catchment with a current, moderate to high precision occurrence as "Potential Habitat".

In addition to direct species observations, we use expert judgments of habitat conditions (Potential Habitat versus Non-Habitat) gathered through desktop and field rapid site assessments. Species experts provided feedback at two stages of model development. An early draft model was reviewed in ArcGIS Online. Completed via a desktop application, experts could scan and zoom to provide feedback on model performance based on their regional knowledge and any data layers they chose to add and view. Experts entered desktop assessments as high precision points (e.g., “this location is habitat or non-habitat”) and/or low precision polygons (e.g., "this region generally provides good habitat or non-habitat). After further development, an improved model received field validation. Field validation consisted of a rapid visual site assessment to judge a stream segment as Potential Habitat versus Non-Habitat based on evidence of stream condition and, in some cases, observation of associated species. All field validation reviews were entered as high-precision line features denoting the stream section observed. In both reviews, time, knowledge, and/or access limitations prevented a true randomized design, but experts attempted to maximize the distribution of feedback spatially and across performance categories (True and False Positives, True and False Negatives) and to target areas with high uncertainty (high variance among models). Judgments vary in their spatial precision, including both polygons (low precision) and points and lines (high precision). Experts’ judgments identify some non-habitat locations, but these are in the minority.

To ensure a balanced sample, additional catchments were randomly drawn as needed from the background (No Data catchments) to serve as pseudoabsence. Pseudoabsence catchments are unsurveyed catchments modeled treated as absence sites for the purposes of defining absence (versus presence) habitat characteristics.

Available Observation Records

OBSERVATION	N
Current	103

Available Expert Review Records

REVIEWER	SOURCE	N
Non-Habitat	AGOLReview	43
Non-Habitat	Field Review	13
Potential Habitat	AGOLReview	35
Potential Habitat	Field Review	36

Observation Set Used in Models

Often multiple data records fell within a single catchment. For the purposes of modeling, to avoid pseudo-replication, we removed duplicates; each catchment was scored as potential habitat or non-habitat. Where a single catchment contained records of both potential habitat and non-habitat (OBSERVATION SET = “Conflict”), it was scored as habitat. Based on these decisions, the final model predicts the probability of any habitat within the catchment rather than the abundance of habitat within the catchment.

OBSERVATION SET	CLASSIFICATION	N
No Data	No Data	1729
Non-Habitat	Non-Habitat	32
Pseudoabsence	Non-Habitat	103
Conflict	Potential Habitat	6
Potential Habitat	Potential Habitat	129

With the available observation data (including pseudoabsence) we were able to classify 270 of the 1999 catchments within the species expert delineated model range.

Environment Data

The environmental data attributes included variables drawn from the US EPA StreamCat and the NHD Plus V2 data. From the available StreamCat data, we extracted 87 raw data variables plus 12 indices. The raw data variables primarily described local catchment characteristics across a spectrum of land cover, climatic, physiographic, hydrologic, chemical, geological, and disturbance metrics. A few of these variables provided data at finer (riparian only) and coarser (all upstream watershed area) scales. The 12 indices (six for the local catchment and six for the entire upstream watershed area) were calculated values summarizing multiple raw variables to represent the quality of aquatic habitat conditions. The environmental data that proved useful to distinguish habitat catchments from non-habitat (and pseudoabsence) catchments are listed in Appendix A.

Model Method

We used Random Forest models to predict the probability of habitat at the scale of USGS National Hydrography Data (NHD Plus V2) catchments and their associated stream segments (median area: 272 acres; 5th and 95th percentile area: 4 acres and 1541 acres). Random Forest models generate predictions through repeated construction of decision-tree style models. At multiple points during model construction and assessment, the random forest process draws a random subset of habitat and non-habitat data and selects a random subset of the environmental variables by which to compare them. This randomization is beneficial to reduce overfitting of the models. The model procedure tracks (1) how frequently sites are predicted to be habitat vs non-habitat, (2) which variables contributed most to accurate classification of habitat vs non-habitat sites, and (3) overall statistics about model performance. We ran the models in R primarily using functions from the randomForest (Liaw and Weiner 2002) and rfUtilities (Evans and Murphy 2018) packages.

We ran multiple random forest models, each initiated with a different suite of environmental (predictor) and habitat/non-habitat (response) data. Models meeting minimum performance criteria (see the Statistics section) were averaged to produce a single predicted probability of habitat per catchment. These model suites included six core models, plus two additional models if associate species had been defined. “Full Set” refers to models initiated with the 87 raw environmental variables from StreamCat plus the NHD Plus environmental variables, while the “Index Set” refers to models initiated with

StreamCat's 12 calculated condition indices plus NHD Plus environmental variables. For the response variable, we evaluated models run with all species observations and expert judgments ("All Habitat Data"), run with only the verified target species observations ("No Reviewer Habitat Data"), run with only the high precision observations (excluding large generalized polygons from reviewer and observation sets) ("No Low Precision Data"), and run with all habitat data for the target species plus the current, high precision observations of associate species ("Plus Associates" for species with expert identified associates).

- Full Set, All Habitat Data
- Index Set, All Habitat Data
- Full Set, No Reviewer Habitat Data
- Index Set, No Reviewer Habitat Data
- Full Set, No Low Precision Habitat Data
- Index Set, No Low Precision Habitat Data
- Full Set, Plus Associate Species Data
- Index Set, Plus Associate Species Data

The inclusion of reviewer data, low precision data, and/or associate species data did not always improve the models. Model statistical scores could be high, indicating good fit to available data, but fail to match experts' expectations based on knowledge of regional habitat variability. Also, low precision comments (large polygons) often included mixed habitat, introducing more noise than signal to the training data. Throughout our analysis, we attempted to balanced consideration of both statistical and expert assessment.

Statistics

We focused on four statistics to evaluate model performance:

- *Sensitivity* (SENS): The proportion of known habitat correctly identified as Potential Habitat (true positive rate).
- *Specificity* (SPEC): The proportion of known non-habitat (including pseudoabsence) correctly identified as Non-Habitat (true negative rate).
- *Area Under the Curve* (AUC): A summary of overall model performance based on both sensitivity and specificity.
- *Cohen's Kappa* (KAPPA): A summary of overall model performance based on performance relative to a random classification.

Individual models were accepted and carried forward if they achieved values of 0.6 or greater for sensitivity, specificity, and area under the curve. For this species, the models accepted and averaged to produce the final map and statistics were: FullSet, FullSetNoRev, FullSetWithAssoc, IndicesNoLow, IndicesWithAssoc.

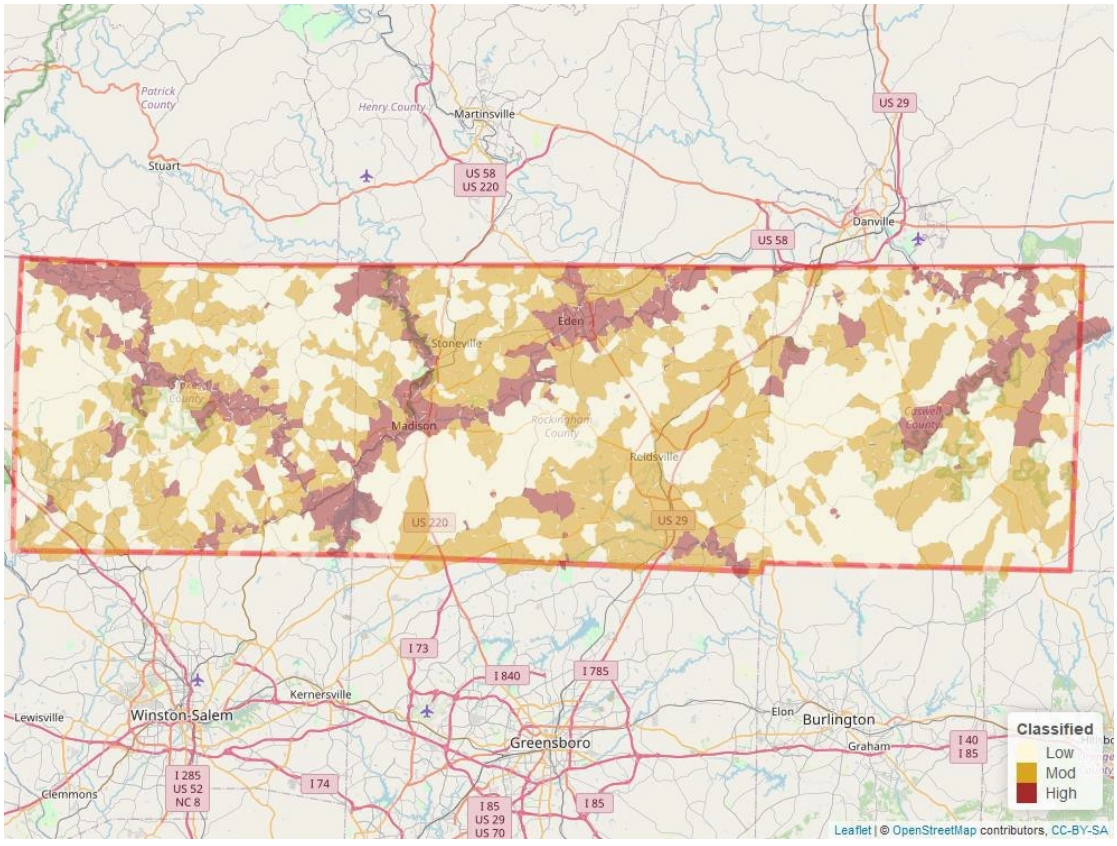
Performance of averaged model predictions

Statistic	Mean	SD
AUC	0.81	0.08
KAPPA	0.41	0.22
SENS	0.68	0.06
SPEC	0.95	0.11

All statistical values should be interpreted with caution as each has unique limitations. The most informative statistic depends on the intended application and the nature of the underlying data.

Model Classification

Prior to delivery to ATLAS, each catchment was classified as Low, Moderate, or High probability of habitat based on the model average predicted probability of habitat.



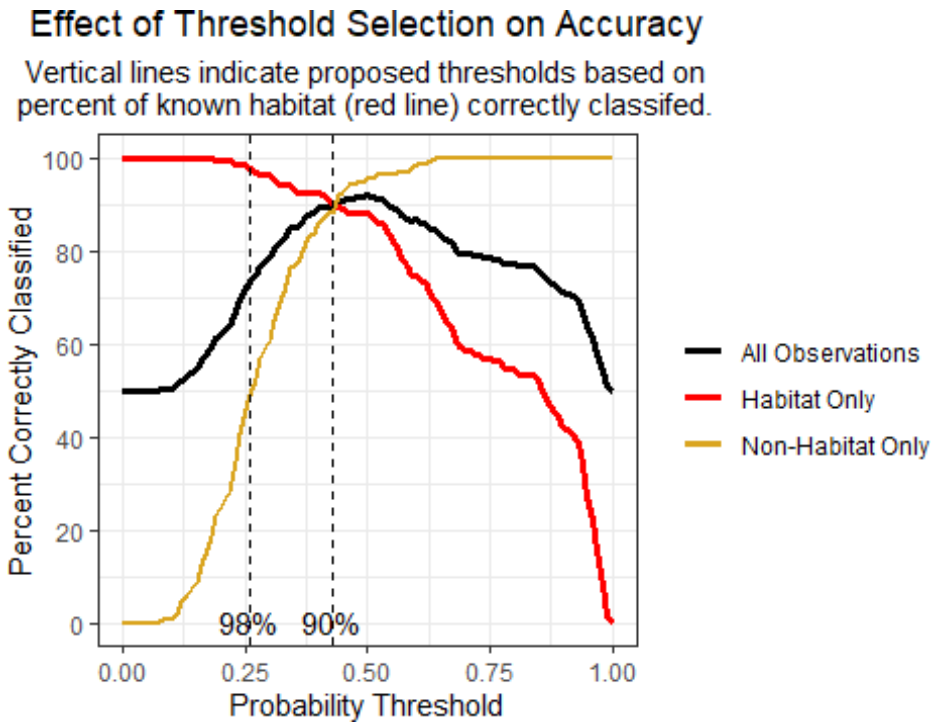
Map of Pleurobema collina range with catchments classified as Low, Moderate, and High probability of habitat.

We set the category thresholds based on the distribution of predicted values for known habitat. The High-Moderate threshold is set at the level where 90% of the observed potential habitat (species presence and reviewer judgments) falls within the High category (Presence Percent Correctly Classified). We set the Moderate-Low threshold at the level

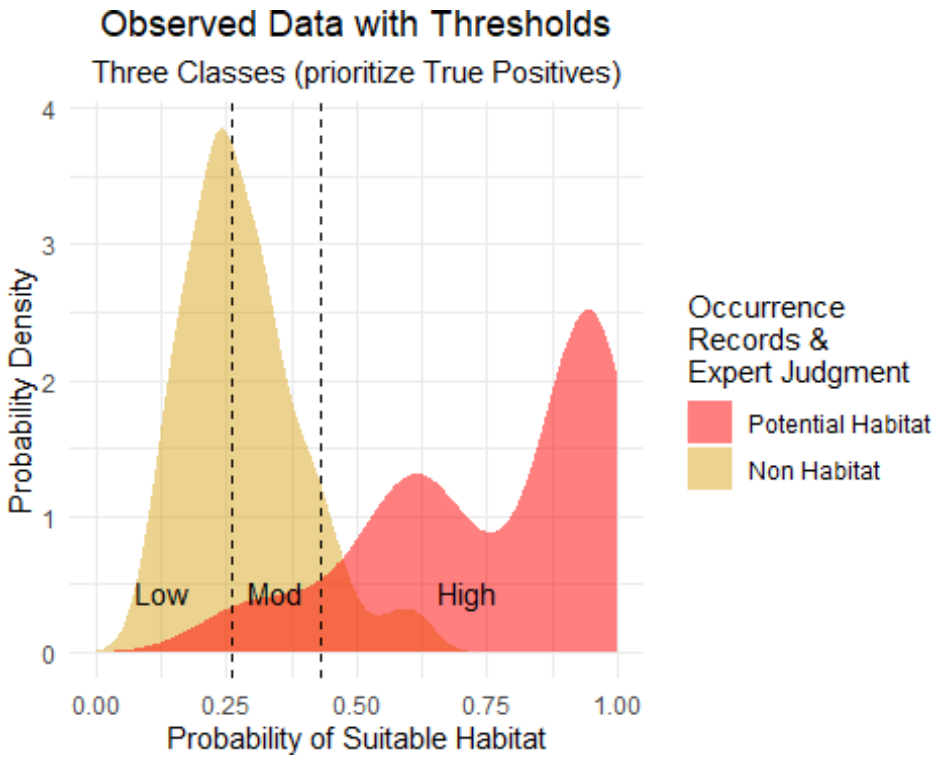
where 8% of the observed potential habitat (species presence and reviewer judgments) falls within the Moderate category and 2% within the Low category. The final thresholds for this species are 0.26 and 0.43 for the Low-Moderate and Moderate-High thresholds, respectively.

Threshold Selection and Interpretation

Lower thresholds result in more of the range labelled as High probability of habitat and greater misclassification of known non-habitat locations. Higher thresholds result in more of the range labelled as Low probability of habitat and greater misclassification of known habitat locations.



Given the larger spatial unit of ecological models and ecological characteristics of aquatic species, the landscape scale environmental attributes of potential habitat varied greatly among sites and could closely resembled non-habitat sites. **It is important to remember that potential habitat (and non-habitat) can occur at any classification level within a catchment and must be verified by a qualified biologist.**



Appendix A: Variable Importance

Each random forest within the ensemble initiated with different environmental variables (following input selection, removal following multicollinearity tests, and removal following model selection tests) and different subsets of observational data. The model process identifies the environmental variables most useful to accurately and precisely classify potential habitat versus non-habitat. The following table lists the environmental variables found useful to accurately distinguish between known habitat and non-habitat (including pseudoabsence) catchments and indicates how many of the models selected each variable. All references to land cover are from the 2016 National Land Cover Data (NLCD), unless otherwise noted.

VARIABLE	N	DESCRIPTION
ORDER	6	Modified Strahler Stream Order
FLOWCFS	5	Flow from gage adjustment (cfs)
DamNrmStorWs	4	Volume all reservoirs (NORM_STORA in NID) per unit area of catchment (cubic meters/square km)
MWST_2014	4	Predicted mean winter stream temperature (Jan-Feb) for year 2015
PctConif2016Cat	4	% of catchment area classified as evergreen forest
PctDevelWs	4	Sum of StreamCat watershed percentages forHigh Density and Moderate Density Urban Development
PctOw2016Cat	4	% of catchment area classified as open water

VARIABLE	N	DESCRIPTION
ELEVCM	3	Minimum elevation (smoothed) in centimeters
HEADW	3	Feature is (1) or is not (0) a headwater stream
PctImp2011Cat	3	Mean imperviousness of anthropogenic surfaces (NLCD 2011) within catchment
SCat	3	Mean % of lithological sulfur (S) content in surface or near surface geology within catchment
SLOPE	3	Slope of flowline (meters/meters) based on smoothed elevations
WetIndexCat	3	Mean Composite Topographic Index (CTI)[Wetness Index] within catchment
AgKffactCat	2	Mean soil erodibility (Kf) factor (unitless) of soils within catchment on agricultural land
CaOCat	2	Mean % of lithological calcium oxide (CaO) content in surface or near surface geology within catchment
CBNFCat	2	Mean rate of biological nitrogen fixation from the cultivation of crops in kg N/ha/yr, within catchment
Fe2O3Cat	2	Mean % of lithological ferric oxide (Fe2O3) content in surface or near surface geology within catchment
ICI	2	Index of catchment integrity
ManureCat	2	Mean rate of manure application to agricultural land from confined animal feeding operations in kg N/ha/yr, within catchment
MgOCat	2	Mean % of lithological magnesium oxide (MgO) content in surface or near surface geology within catchment
Na2OCat	2	Mean % of lithological sodium oxide (Na2O) content in surface or near surface geology within catchment
NCat	2	Mean % of lithological nitrogen (N) content in surface or near surface geology within catchment
PctForestWs	2	Sum of StreamCat watershed percentages for Coniferous, Deciduous, Mixed Forest, and Woody Wetland
PctForestWsRp100	2	Sum of StreamCat watershed riparian percentages for Coniferous, Deciduous, Mixed Forest, and Woody Wetland
PctHay2016Cat	2	% of catchment area classified as hay land use
PctMxFst2016Cat	2	% of catchment area classified as mixed deciduous/evergreen forest
PctUrbLo2016Cat	2	% of catchment area classified as developed, low-intensity land use
PctWdWet2016Cat	2	% of catchment area classified as woody wetland
Pestic97Cat	2	Mean pesticide use (kg/km2) in 1997 within catchment

VARIABLE	N	DESCRIPTION
PopDen2010Cat	2	Mean populating density (people/square km) within catchment
ElevCat	1	Mean catchment elevation (m)
FertCat	1	Mean rate of synthetic nitrogen fertilizer application to agricultural land in kg N/ha/yr, within the catchment
HUDen2010Cat	1	Mean housing unit density (housing units/square km) within catchment
HydrlCondCat	1	Mean lithological hydraulic conductivity (micrometers per second) content in surface or near surface geology within catchment
PctCrop2016Cat	1	% of catchment area classified as crop
PctGrs2016Cat	1	% of catchment area classified as grassland/herbaceous
PctUrbMd2016Cat	1	% of catchment area classified as developed, medium-intensity land use
PctUrbOp2016Cat	1	% of catchment area classified as developed, open space
RdCrsCat	1	Density of roads-stream intersections (2010 Census Tiger Lines-NHD stream lines) within catchment (crossings/square km)
Sinuosity	1	Reach length (m) divided by the same reach's straight line length (m, from the beginning node of the reach to the end node of the reach).