Random Forest Species Model Documentation

Developed for the NCDOT ATLAS Project

Swamp pink *(Helonias bullata)*

- **Lead modeler**: Ashton Drew, KDV Decision Analysis (ashton.drew@kdv-decisions.com) 919-886-2811
- **Model version**: Version 1 (2019-09-10)
- **Delivered Products**:
  - *Potential Habitat, Version 1*: A 3-level reclassification of the model prediction for use by ATLAS (shapefile)
  - *Model summary documentation*: This document.
  - *Probability of Potential Habitat, Version 1*: Appendix 1. The 30-m resolution probability raster map produced by the model (tif).
  - *Model R code and associated documentation*: Appendix 2. This collection of files includes model code (R), the lookup tables used to define environmental data layers and post-processing masks (Excel), figures and R data products used to assess model performance (various), and a README document explaining the contents.
  - *Desktop Review Results, Draft model*: Appendix 3. Reviewer comments (shapefile) and a summary of review interpretation and recommendations for model improvement (pdf).
  - *Field Assessment Results, Version 1 model*: Appendix 4. Reviewer comments (shapefile) and a summary of field observations and recommendations for model improvement and application (pdf).

The Swamp pink model is a Random Forest (machine-learning) model. As such, it returns the **probability of potential habitat**, based on the core assumption that current presence locations are representative of potential habitat within the state of North Carolina. For the purposes of ATLAS applications, this model is reclassified to a 3-level map product distinguishing 30-m raster grid cells with predicted low, moderate, and high probability of potential habitat.

Species Description

Swamp pink occurs in clonal clumps in a variety of groundwater- influenced wetland habitats including southern Appalachian bogs and swamps, Atlantic white cedar swamps, swampy forests bordering meandering small streams, boggy meadows, headwater wetlands, and spring seepage areas. The perennial herb requires a constantly saturated, but not flooded, water supply. The plant often grows on hummocks formed by trees, shrubs, and sphagnum moss, and exhibits varying degrees of shade tolerance. Swamp pink

occurs in acidic soils that contain a very thin layer of decomposed organic matter over a dark silt loam and a subsoil of sand, loam, and gravel.

Data Resources

Species Data

We gathered presence data from multiple sources, listed below, and rasterized these to a 30-m scale to match our environmental data. Any grid cell intersecting known occurrence points or polygons was attributed as "presence". No true absence data were available, so the remaining grid cells (areas without known occurrence) were attributed as "pseudoabsence".

1. **US Fish and Wildlife Service (USFWS) Range Data**: The model extent was defined based on USFWS current range data, applied through agreements between NCDOT and USFWS.
2. **NC Natural Heritage Program (NCNHP) Element Occurrence (EO) Data**: Observations evaluated for use in the model included all plant species records where STATUS=Current and ACCURACY=1-Very High, 2-High, and 3-Medium as of the most recent Tier 2 data release. Some, but not all, models included the medium accuracy data.
3. **NC Department of Transportation (NCDOT) Field Pre-Validation Survey Data**: Field surveys conducted to verify current EO status and improve the accuracy of EO records for several species added new data for some species.
4. **NCDOT Past NRTR Project Data**: Data gathered from past project files provided 6 years of presence/absence polygons and up to 2 years of habitat/non-habitat polygons within NRTR study areas.
5. **Expert Reviewer AGOL Desktop Review Data**: Species experts completed a structured, spatially explicit review of a draft version of this model (see below). Experts' potential habitat/non-habitat judgments served as additional input for some models.

Within the USFWS range there are 5,496,102 30-m raster grid cells. From the intersection of these grid cells with the available occurrence data, we obtained:

- **Presence**: 861 cells attributed as high precision, current observations of Swamp pink. All presence locations were used to train the model, because the random forest model process includes randomized out-of-bag testing as part of development.
- **Mediums**: 29 cells attributed as moderate precision, but current observations. The use of these presence cells to train the model depended on how much noise versus signal they added. Medium cells were excluded from this version of the model.
- **Historic**: 19 cells attributed as historic (extirpated) observations. These observations were not used to train the model but were referenced during model review.
- **Associates**: 0 cells attributed as current, high precision observations of associated species, but without record of the Swamp pink. If present, these observations were not used to train the model but were referenced during model review.

- **Target Taxa Group**: 10,204 cells attributed as current, high precision observations of other plant species where the Swamp pink was not documented as present. Target taxa group cells were handled as a special class of pseudoabsence data.
- **Pseudoabsence**: No true absence data were available for this project, so random draws from the remaining grid cells served as pseudoabsence data. The number of points drawn for each model run was equal to the total number of presence points.

Environmental Data

We had access to 70 environmental data layers across 6 thematic areas: Spectral, Landform, Land Cover and Vegetation, Geology and Soils, Disturbance, Climate. All data are in NAD 1983 State Plane North Carolina FIPS 3200 (US feet), 30-m spatial resolution, with state-wide extent. Appendix 2 includes further documentation of environmental data layers.

We initiated this model with a subset of 15 variables based on variable importance in earlier drafts (the reviewed draft initiated with all 52 layers available at that time), previously untested layers (new or updated data layers), reviewer feedback, and interpretation of patterns in earlier versions. The initial subset of variables presented to the model was further refined by testing for multicollinearity and performing model selection.

Table 1. Environmental variable set provided to the random forest model.

Group	Variable
Climate	Mean Annual Precipitation
Disturbance	Burn Area Density
Geology and Soils	Drainage Class
Geology and Soils	Hydric Classification Presence
Geology and Soils	Predominant Lithology
Geology and Soils	Soil Percent Clay
Geology and Soils	Soil Percent Organics
Land Cover and Vegetation	Distance to Streams
Land Cover and Vegetation	GAP Land Cover (Merged)
Land Cover and Vegetation	NLCD Imperviousness
Land Cover and Vegetation	NLCD Land Cover
Land Cover and Vegetation	Segment NDVI
Landform	Elevation
Landform	Slope
Landform	Geomorphon: Valley

Model Approach and Output

Random forest models generate predictions through repeated construction of decision-tree style models. At multiple points during model construction and assessment, the random forest draws a random subset of presence and pseudoabsence data, as well as random subset of available environmental data. The model procedure tracks (1) how frequently sites are predicted to be presence vs absence, (2) which variables contribute most to accurate classification of presence vs absence sites, and (3) overall statistics about model performance. We ran the model in R using the randomForest (Liaw and Weiner 2002) and rfUtilities (Evans and Murphy 2018) packages. This document summarizes key aspects of the model specification and outputs, as well as results of a desktop review and field assessment. The model code and further details are available in Appendix 2.

The model predicts the probability of potential habitat for the species, given the assumption that the available presence data are representative of suitable habitat within the state.

The predicted probability of potential habitat (0 to 1) reflects the frequency with which a raster grid cell was classified as potential habitat versus non-habitat through all the permutations of random forests (see Appendix 1).

We created the 3-level (low, moderate, and high probability of potential habitat) representation of the model prediction by setting probability thresholds at 0.25 (low/moderate) and 0.74 (moderate/high). These thresholds were set through discussion with the expert biologists in reference to their observations of model strengths and weaknesses during the field assessments. The selection of a threshold is a judgement based on acceptable risk and desired level of precaution for a given application of the model. As a threshold is dropped, more area with decreasing similarity to known presence locations will be categorized as the higher level class (e.g., lowering the moderate/high threshold labels more habitat as high probability). Model documentation (Appendix 3) provides a table of all possible thresholds, in 0.01 percent increments, with associated percent correctly classified, sensitivity, and specificity accuracy statistics.

Figure 1. The model uses data from, and makes predictions for, the species' county range area (red) as designated by USFWS IPaC as of 2019-09-10.

Figure 2. The probability of potential habitat as predicted by the random forest model (raster map data available in Appendix 1).

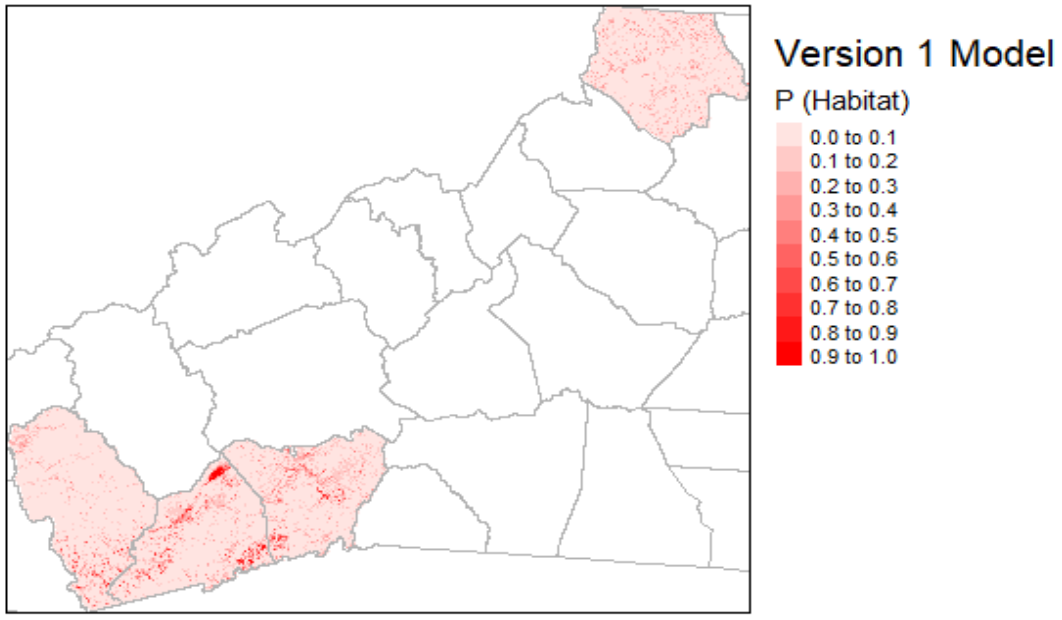
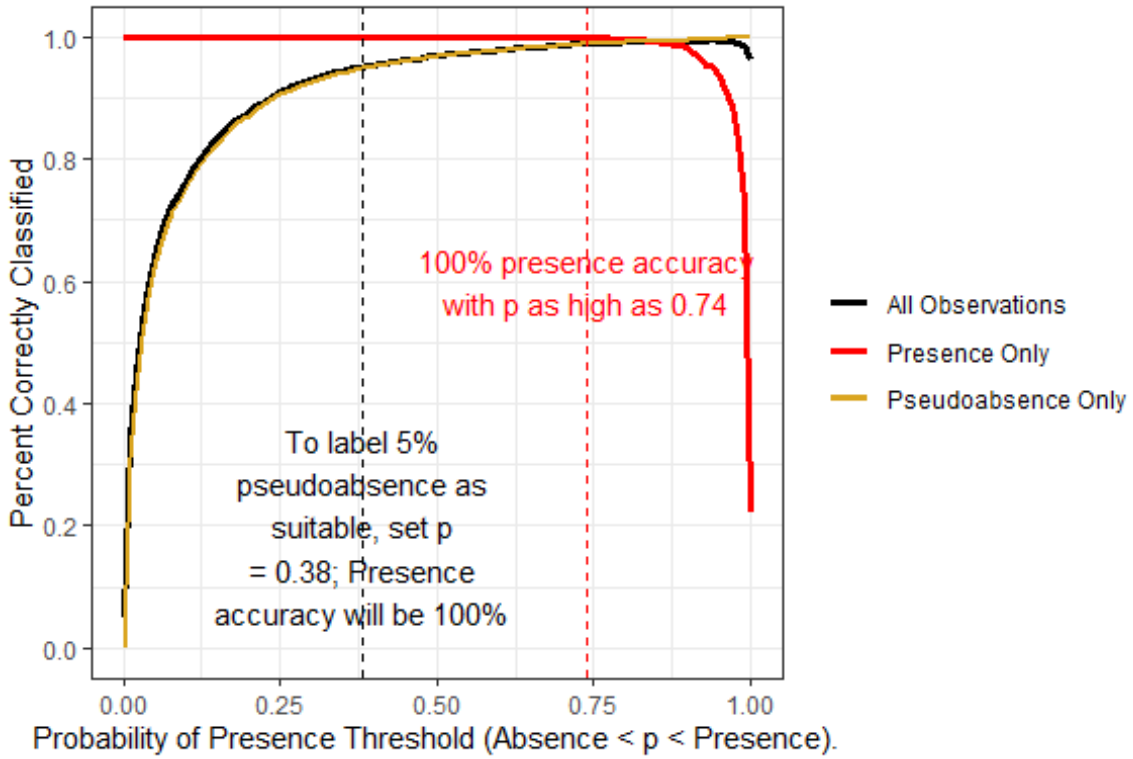


Table 2. Variable importance scores for the final set of environmental variables used in the random forest model (after testing for multicollinearity and parsimony). Mean decrease in accuracy refers to how much more poorly the model performs if the variable is excluded. Mean decrease in Gini refers to how the Gini coefficient, a measure of homogeneity within groups after a random forest split, is affected by the removal of the variable.

Variable	Mean Decrease in Accuracy	Mean Decrease in Gini
Elevation	36	159
Slope	36	99
Mean Annual Precipitation	32	62
Burn Area Density	27	137
Soil Percent Clay	25	30
Predominant Lithology	25	118
Hydric Classification Presence	23	74
Segment NDVI	22	33
GAP Land Cover (Merged)	21	37
NLCD Land Cover	18	29
Drainage Class	18	59
Distance to Streams	17	21
Soil Percent Organics	17	30
Geomorphon: Valley	14	14
NLCD Imperviousness	10	6

Figure 3. Threshold plot showing the percent correctly classified at all possible probability thresholds. At lower thresholds, more of the landscape is classified as potential habitat; all known presence (red line) are correctly classified, but many pseudoabsence (gold line) are also called presence. At very high threshold values, most pseudoabsence are classified as absence, but some known presence sites are misclassified as absence.



Draft Model Review and Improvements

We conducted draft model reviews and analysis of reviewer feedback between September 2018 and July 2019 (Appendix 3). The experts providing feedback on a draft model, via an ArcGIS Online (AGOL) portal, were:

- Jame Amoroso, a Conservation Information Specialist for the North Carolina Natural Heritage Program. She has been with NCNHP since 1994, starting as Program Botanist. Past and current work has included publishing the NCNHP Rare Plant List and maintaining conservation data for federally-protected species. Jame received her Masters of Science degree in Botany from the University of Florida with the thesis A Floristic Study of Cedar Key Scrub State Reserve, Levy County, Florida.
- Lesley Starke, the Plant Conservation Program Manager with NC Department of Agriculture and Consumer Services. She has worked with North Carolina's imperiled plant species through her work at the Plant Conservation Program since 2010 and has a strong background in remote sensing and species distribution modeling.
- Suzanne Mason, a data manager for the North Carolina Natural Heritage Program. She has been with the NCNHP since 2005 and specializes in maintaining conservation data

for federally-protected species. Suzanne previously studied the genetic diversity of Schweinitz's sunflower for her Master of Science thesis.

The AGOL review requested each reviewer to individually examine the model at approximately 20 flagged locations chosen by the modeler plus a minimum of 20 additional locations of the reviewer's choice. For this review, we presented a binary representation of the continuous probability prediction, where "potential habitat" represented a proposed threshold for "moderate to high probability of potential habitat" and "non-habitat" represented grid cells with lower probability. We requested comments address both modeled non-habitat and potential habitat, with at least 5 sites where they disagreed and 5 where they agreed with the classification within each category. At each location, the reviewer (1) indicated if the modeled classification (potential habitat or non-habitat) matched their own best professional judgment given their experience, the aerial imagery, and any additional information they chose to consult, and (2) commented on how they reached their conclusion. Multiple responses at flagged locations gave insight into reviewer consensus while their dispersed comments ensured breadth of spatial coverage. Reviewer comments (Appendix 3) informed model improvements and supplemented available occurrence data.

The experts provided a moderately complete review. Responses were unbalanced, providing more information where the model performed well than where the model performed poorly. In addition, multiple points placed outside the model area (e.g. outside USFWS range) provided us no opportunity to learn about or improve model performance and reduce perceived model accuracy. Based on their feedback and our own review of model performance, we made the following changes:

- Update NLCD landcover and associated derived layers to the new 2016 data to better reflect current land use and condition.
- Create and add new layer Distance to Agriculture to reduce probabilities within agricultural fields (too disturbed).

Based on reviewer feedback, we also applied a mask to the final model. Under any of the following conditions, model predictions were overruled and converted to 0 probability of potential habitat: Open water, High density urban development, Tidally influenced area, Tidally influenced and saline area, Impervious surface, Interior fields and ditches.

Figure 4. Binary classification of the draft model as presented to reviewers

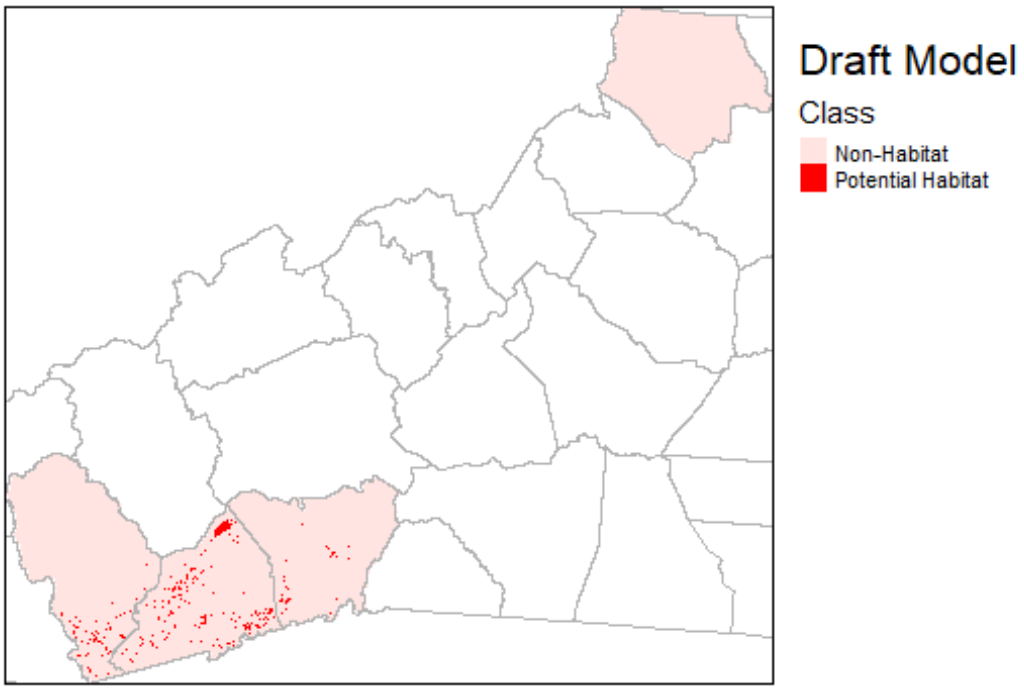


Figure 5. Locations and class of AGOL desktop reviewer comments.

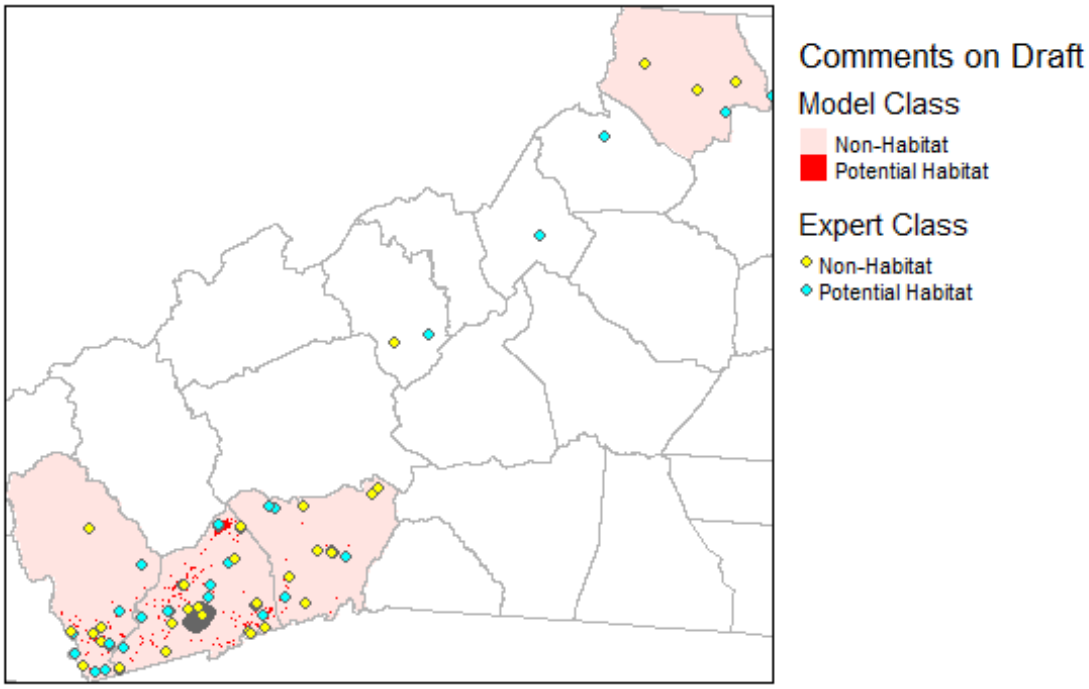
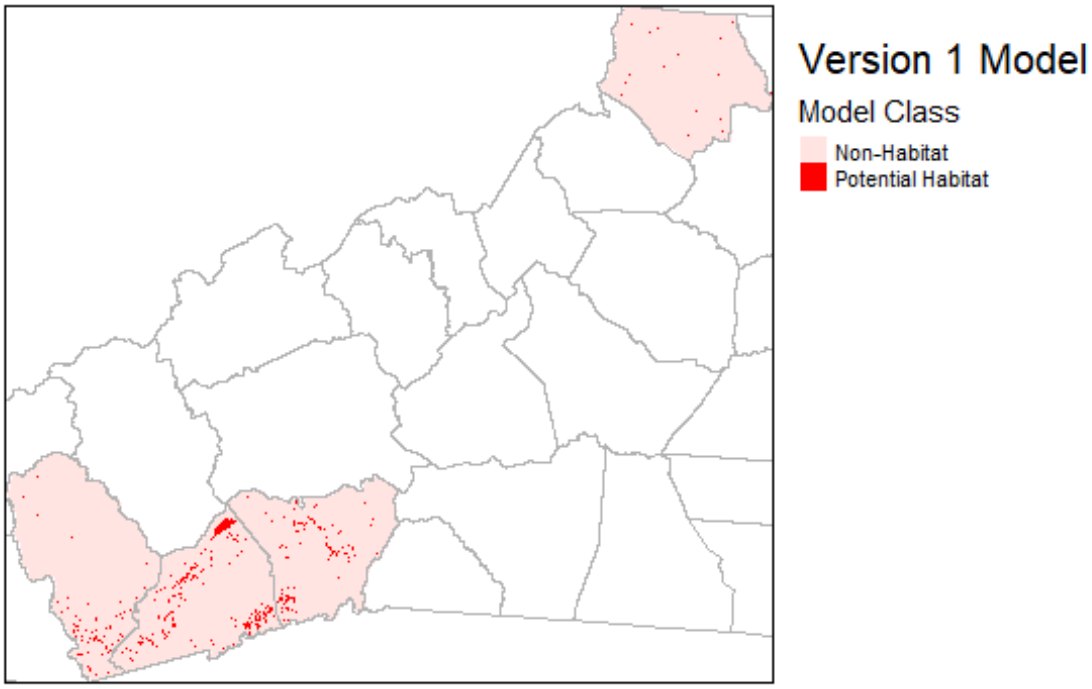


Figure 6. Binary classification of Version 1 of model. We assessed changes in model performance by comparing the accuracy statistics (based AGOL Review comments) of the Draft versus Version 1 model. We then used the binary classification to design a stratified sample for field assessment of the Version 1 model.



Accuracy Improvements: Draft to Version 1

We assessed model improvement from the draft to Version 1 by calculating the accuracy statistics of each based on the desktop review point and polygon location judgments. This was an assessment of the binary classification, not the probability prediction of the model; accuracy scores are dependent upon both the underlying model and the selected threshold(s).

Figure 7. Accuracy summary of Draft (left) and Version 1 binary (right) models.

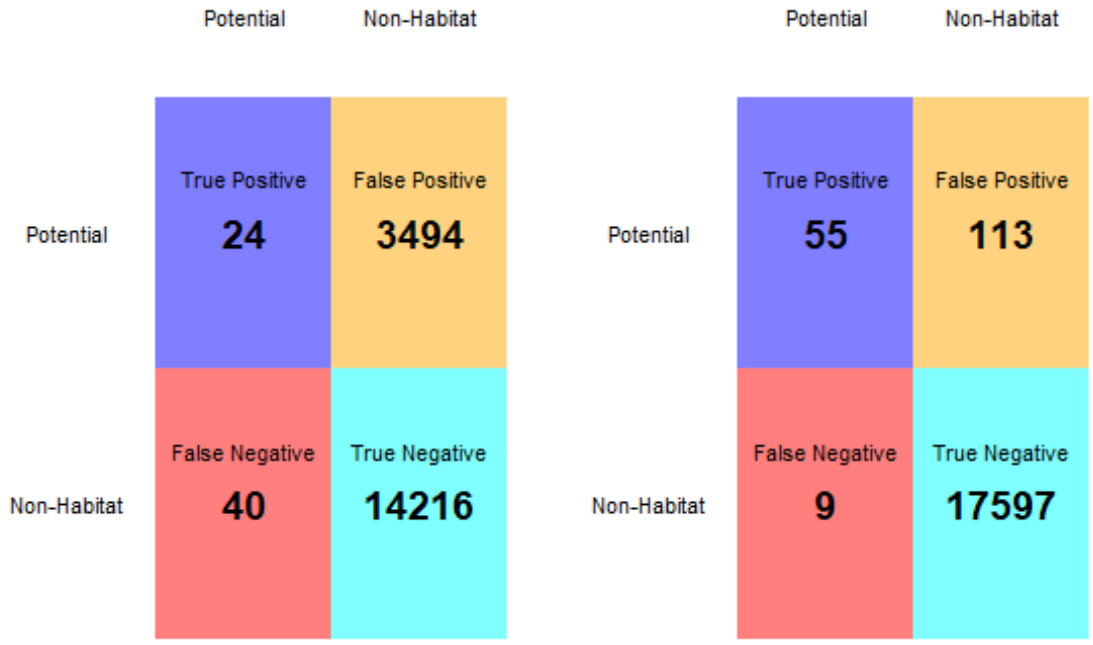


Table 3. Desktop review accuracy statistics based on the counts in the summary table.

Statistic	Draft	Version 1
Percent Correctly Classified	80.1	99.3
Sensitivity	0.4	0.9
Specificity	0.8	1.0

- *Percent Correctly Classified*: Sum of all True Positives and True Negatives divided by total number of review points.
- *Sensitivity*: Sum of all True Positives divided by sum of all points modeled as potential habitat. Lower numbers indicate bias towards calling everything habitat to avoid missing a single habitat location, but it means that any given site predicted to be potential habitat has a high likelihood of being a false prediction. A high sensitivity model is usually most useful where predicting non-habitat.
- *Specificity*: Sum of all True Negatives divided by sum of all points modeled as non-habitat. Lower numbers indicate bias towards calling everything non-habitat, even at the risk of missing one or two potential habitat sites. A high specificity model is usually most useful where predicting potential habitat.

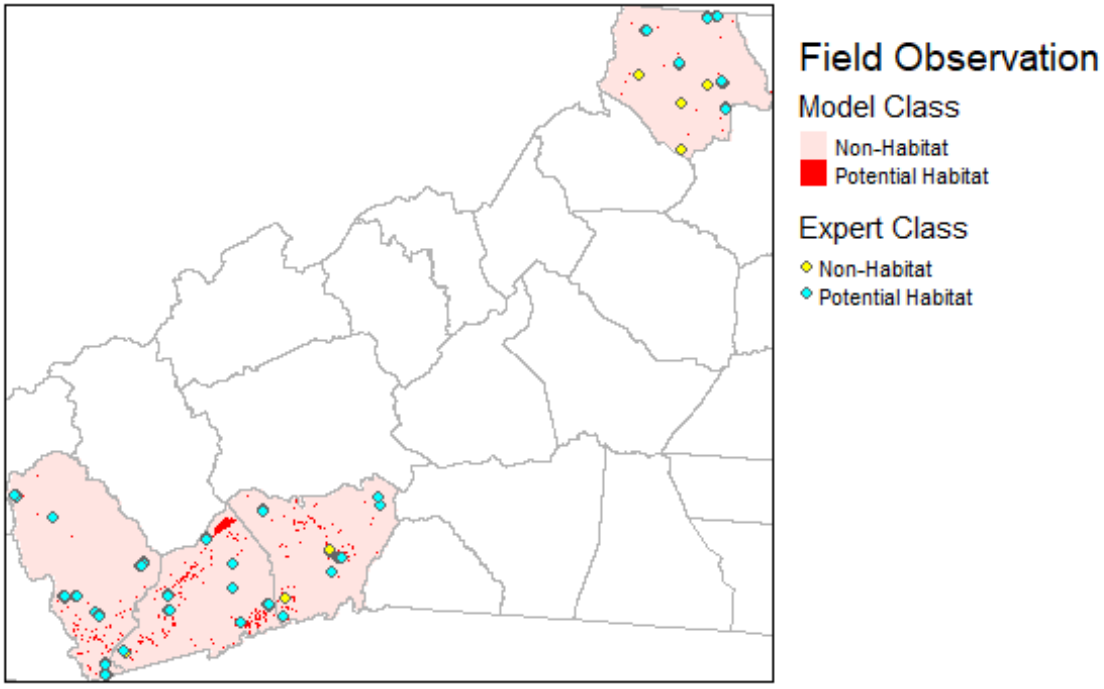
Version 1 Model Field Assessment

From November 2019 through February 2020, biologists conducted rapid field assessments of a binary classification (Potential Habitat and Non-Habitat) of the Version 1 continuous model predictions . A stratified sample of points were generated on “accessible lands” (generally public lands and right-of-ways) and biologists aimed to survey at least 10

points per county within the range. At each visited point, biologists characterized the site as "Potential Habitat" or "Non-Habitat", based upon their best professional judgment of the visible vegetation community and environmental characteristics. They also mapped the area as a polygon and provided site descriptions and photos to support their conclusion. If a single site, based on the scale of a 30-m grid cell, included both Potential Habitat and Non-Habitat (e.g., differing habitat on either side of a road), two polygon entries were logged. The experts providing feedback on the binary classification of this Version 1 model were:

- Eric Black, a Senior Environmental Project Manager with Scenic Consulting Group with experience in federally protected plant and animal surveys in both the private and public environmental sectors. He served as western plant coordinator for the ATLAS project.
- Pam Ferral, an environmental scientist and Certified Wildlife Biologist® with Stantec with more than 25 years of experience in conservation and consulting. She spent most of her career as a biologist with the U.S. Forest Service and conservation ecologist with The Nature Conservancy, specializing in the management and monitoring of endangered species in coastal plain ecosystems. She now conducts rare plant and animal surveys for a variety of private sector clients.

Figure 8. Locations and class of field assessment comments.



Field Assessment Accuracy Statistics

In addition to providing data to support an accuracy assessment, the biologists summarized their overall impressions of model performance. Biologists were asked to consider three specific questions: (1) What, if any, patterns did they note regarding False Positives?, (2) What, if any, patterns did they notice regarding False Negatives?, and (3)

Considering all the counties within the range, did any regions of the model seem to do particularly well (or poorly)? We used their feedback to discuss the final classification of the model and set thresholds to define three bins and to provide recommendations to ATLAS for model applications and improvements.

Figure 9. Accuracy summary based on field assessment. The units in the confusion matrix are polygons drawn by the biologists.

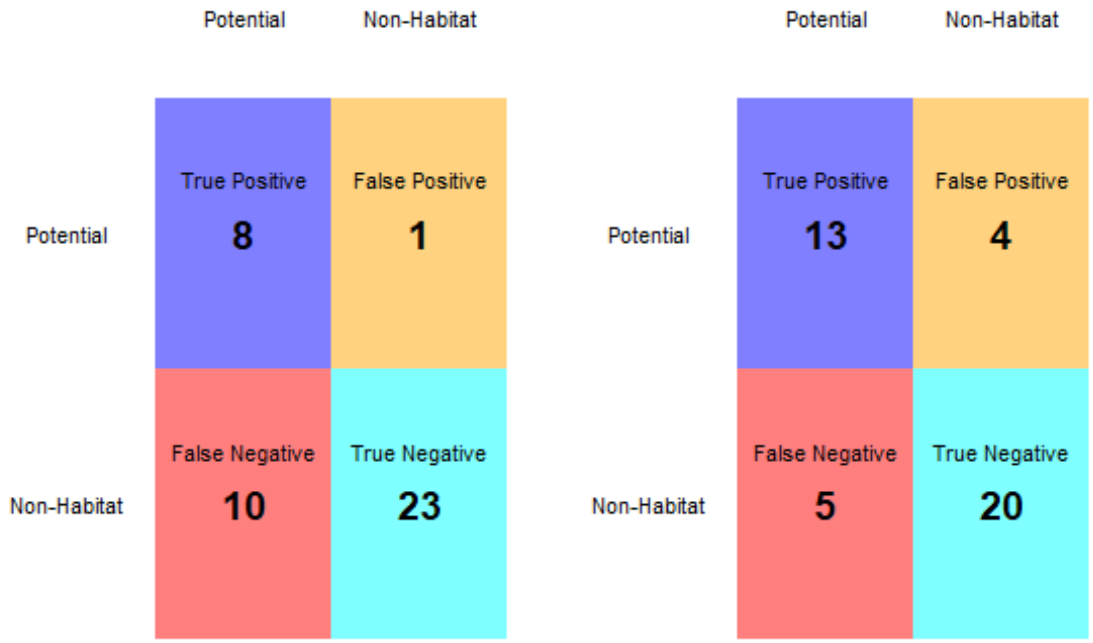


Table 4. Field-based accuracy statistics based on the counts in the summary table.

Version	PCC	Sensitivity	Specificity
Draft	0.738	0.444	0.958
Version 1	0.786	0.722	0.833

Field-based Comments on Model Performance

The swamp pink model generally predicts known habitat (e.g. stream associated wetlands, spring seepages, wet meadows, swamp forest, bogs, stream sides) and excludes unlikely habitat (e.g. upland forests, pastures, parking lots, regularly maintained properties, urban areas, ridgelines, hilltops, hill sides etc.). All observations were based solely on visual cues in the field. In areas where predicted vs non-predicted habitat looked the same (e.g. There was no visual distinction between sites), I assumed that non-visual factors such as soil type, pH, soil moisture, temperature may have affected habitat prediction. In summary:

- False positives were generally in areas where flood plain soil hydrology had been altered (e.g. drain or fill, incised stream, man-made levees) or steep gradient high velocity streams.
- False negatives were typically associated with microhabitats (e.g. small drainage associated seeps) or floodplain depressions.

- The model itself seemed to work well throughout all counties but tended to over predict in agriculturally impacted floodplains in Ashe, Henderson, Transylvania Counties.

Final Three-Level Classification of Version 1 Model for ATLAS

With the gathered data and in direct consultation with the field biologists, we assessed model performance and adjusted the thresholds to create a three-level version of the model for delivery to ATLAS. The three levels are: Low, Moderate, and High Probability of Potential Habitat (based on similarity of environmental conditions to those found at known occurrence locations). These levels represent the fact that given limited knowledge of species biology, continuously changing environments, and potential for gaps and error in both species and environment data, a model prediction dependent on remotely-sensed data can never predict species occurrence or habitat with absolute accuracy and precision. Thus, "Low" probability habitat represents regions and sites where biologists would be very surprised to find this species and its habitat (occurrence here should be extremely rare). In "High" probability habitat, biologists expect to frequently encounter areas that look like potential habitat based on visible environmental and vegetation community characteristics. The thresholds for this species are: Low-Moderate (0.25) and Moderate-High (0.74).

Table 5. Distribution of field assessed survey points across the Low, Moderate, and High classification of the predicted probability of Potential Habitat. The minimum, median, and maximum associated probability values are also shown. Note: The models never predict 0 probability of Potential Habitat; a 0 only occurs where model predictions are overwritten by an expert mask (e.g., open water, >85% impervious surface, etc.).

Predicted Class	Field Review	N	Min	Median	Max
Low	Non-Habitat	18	0.005	0.056	0.240
Low	Potential Habitat	2	0.157	0.182	0.207
Moderate	Non-Habitat	6	0.273	0.406	0.671
Moderate	Potential Habitat	11	0.323	0.423	0.672
High	Non-Habitat	0	NA	NA	NA
High	Potential Habitat	5	0.788	0.840	0.958

Figure 10. Probability density of sites identified as "Potential Habitat" and "Non-Habitat" during field assessment of Version 1 model. The field data have been resampled by class (n=200 each class from raster grid cells within field drawn polygons, with replacement), to ensure balanced sample within each probability class. NHP data are shown below the x-axis.

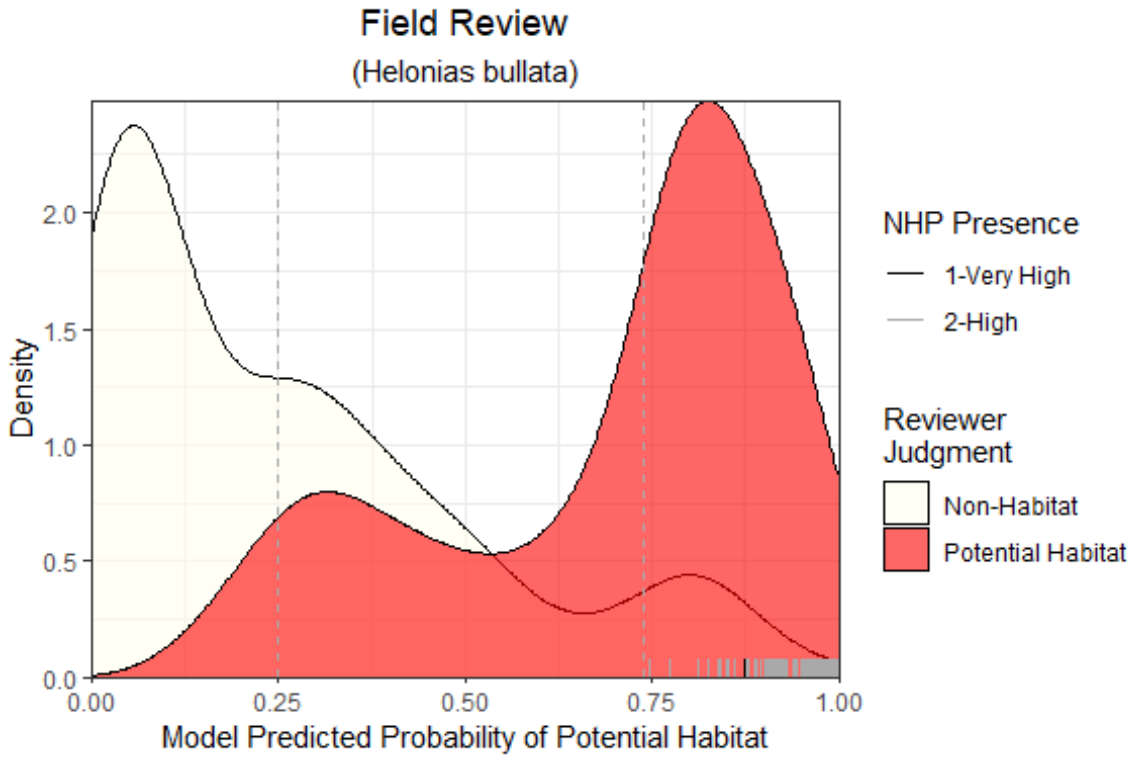
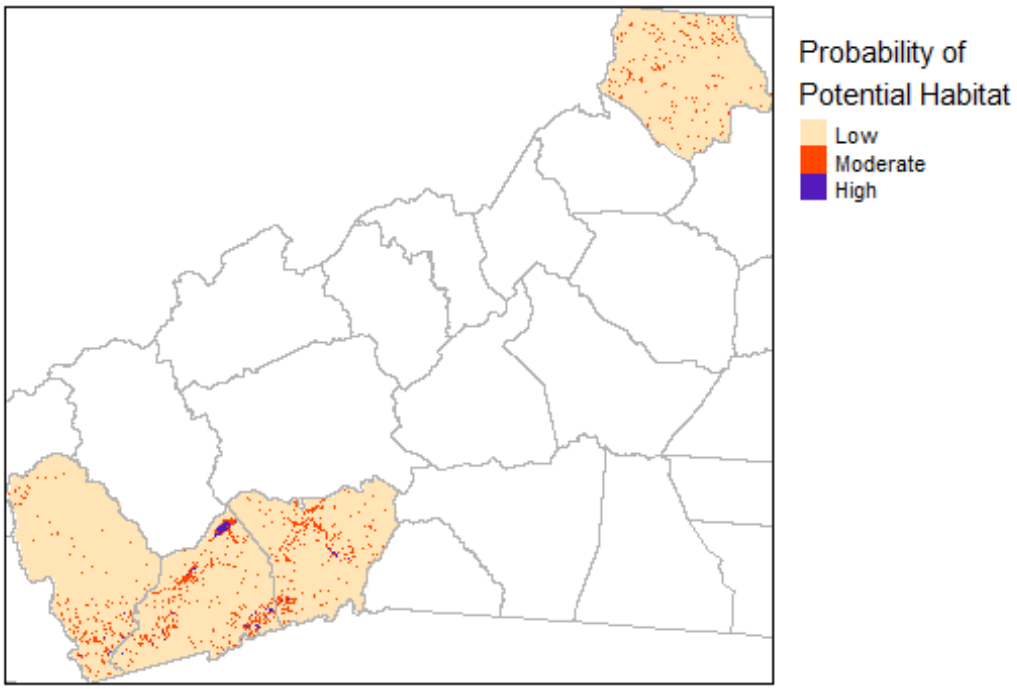


Figure 11. Final three-level map product approved based on all available observational data and discussions with field biologists.



Model Recommendations

- The model would benefit from better information to characterize riparian and stream channel characteristics. Data in development by ATLAS for aquatic species models (e.g., stream gradient, upstream disturbance) and stream modeling (e.g., floodplain limit and flood frequency) might be useful layers to add to the model to address reviewer comments.
- Re-run model so Macon County included in output - a new range addition.

References

- Liaw A, Wiener M (2002) Classification and Regression by randomForest. R News 2(3) 18-22.
- Evans JS, Murphy MA (2018) rfUtilities. R package version 2.1-4, https://cran.r-project.org/package=rfUtilities.

Model R Code

Appendix 2 provides (1) the R files used for final data prep, models, model assessment, and report generation of this model, (2) associated reference data (environment layers lookup, species information lookup, and post-model mask lookup), and (3) a guide to these resources. General details of this model:

- R version 3.6.1 and R Studio version 1.2.1578
- Consultant internal model reference code: Helonias_bullata_NoMedNoAssocWithRev
- Core modeling packages: randomForest 4.6-14 and rfUtilities 2.1-4
- Training data:
  - Use "mediums" as presence to train model: FALSE
  - Use "associates" as presence to train model: FALSE
  - Use "target taxa" as pseudoabsence to train model: TRUE
  - Use reviewer judgments as habitat and non-habitat to train model: TRUE
  - Number of grid cells provided as presence: 918
- Environmental data:
  - Number of environmental variables provided: 15
  - Test for multicollinearity and remove indicated variables?: YES
  - Use model selection procedures to improve model parsimony?: YES
  - Number of environmental variables in final model: 15
- Model parameters:
  - Number of trees: 501
  - Number of variables tested per split: 3
  - Model averaging: YES, 10 iterations, each with unique balanced sample of pseudoabsence
- Model evaluation:
  - Cross-validation performed: YES (out-of-bag method)
  - Significance testing performed: YES
    - Random OOB error: 0.501
    - Model OOB error: 0.013
- Final out-of-bag accuracy statistics (draft values in parentheses):
  - Percent Correctly Classified: 92.3 (90.7)
  - Area Under the Curve: 0.79 (0.9)
  - Kappa: 0.69 (0.7)
  - Sensitivity: 1 (0.9)
  - Specificity: 0.58 (1)