COMMONWEALTH of VIRGINIA

Predicting Locations for Restoring Biologically-diverse Longleaf Pine Communities in Virginia

Prepared for:
Florida Natural Areas Inventory
1018 Thomasville Road, Suite 200-C
Tallahassee, FL 32303

Virginia Department of Conservation and Recreation
Division of Natural Heritage
Natural Heritage Technical Report 21-16
August 2021

DCR
Department of Conservation & Recreation
CONSERVING VIRGINIA'S NATURAL AND RECREATIONAL RESOURCES

Predicting Locations for Restoring Biologically-diverse Longleaf Pine Communities in Virginia

Submitted to:

Florida Natural Areas Inventory
1018 Thomasville Road, Suite 200-C
Tallahassee, FL 32303

Prepared by:

David N. Bucklin, J. Christopher Ludwig, Joseph T. Weber,
Kirsten R. Hazler, and Danielle N. Kulas
Virginia Department of Conservation and Recreation – Division of Natural Heritage
600 East Main Street, 16th Floor
Richmond, VA 23219

August 2021

Suggested citation:
Bucklin, D.N., J. C. Ludwig, J.T. Weber, K.R. Hazler, and D.N. Kulas. 2021. Predicting Locations for Restoring Biologically-diverse Longleaf Pine Communities in Virginia. Natural Heritage Technical Report 21-16. Richmond, Virginia: Department of Conservation and Recreation, Division of Natural Heritage.

Table of Contents

Introduction

As conceptualized by Norman Myers (1988, 1990) a global biodiversity hotspot is a biogeographic region with significant levels of biodiversity that is severely threatened by human development and other anthropogenic changes to the land. The concept was further refined by Mittermeier et al. (2000) and Myers et al. (2000) and since then, 36 biodiversity hotspots have been recognized, including the most recent - the North American Coastal Plain (NACP). More than 1,500 endemic vascular plants have been identified in the region which has experienced greater than 70% habitat loss. This 280-million acre hotspot includes almost the entire 95-million acre range of longleaf pine (*Pinus palustris*). This reinforces an important aspect of longleaf and the broader suite of natural communities associated with the tree – they have a remarkable, unique and important biota.

Though this region was recognized as a global biodiversity hotspot in just the last few years, the significance of this region and the demise of the keystone longleaf pine and its associated biodiversity have been recognized for decades. It has inspired conservation agencies and organizations to protect and restore large forested blocks to ensure the significant biodiversity of the longleaf pine system endures throughout the southeastern U.S.

Since about 1980, conservation partners in Virginia have worked to locate, protect and manage blocks of landscape supporting the remnant patches of southeast Virginia's former longleaf forests and/or their associated biodiversity. While Virginia's estimated original 1-million acre longleaf forest has been reduced to fewer than 200 mature native trees (Virginia Department of Forestry, 2014), substantial biodiversity can still be found, and has been protected in the region where the groundcover and native plants of the original longleaf pine forest persist.

Over the past 20 years, Federal and State natural resource agencies as well as The Nature Conservancy have protected 35,000 acres of land in southeast Virginia including 20,000 acres dedicated to management of longleaf pine forest biodiversity. This expanding network of conserved lands is where longleaf and its associated biodiversity are making an inspiring return to Virginia's coastal plain.

One such area is the 3,750+-acre South Quay Sandhills Natural Area Preserve (NAP), and it is here that the majority of native Virginia genotype longleaf occurs. While a few mature longleaf pines are found in other protected areas, many thousands of acres have been protected with the intent of restoring the longleaf pine savanna ecosystem. These areas have often been targeted for protection because the ground cover or other significant biodiversity associated with longleaf pine systems remains, even though the longleaf is gone.

While great strides have been made in recovery of longleaf forest habitat, we must redouble our efforts - across the entire range of longleaf - to expand protection of stands with longleaf and/or remnant biodiversity and to scale up the pace and scope of our habitat management work. The

urgency of this work is high, particularly in the face of much uncertainty about how climate change, future development and land-use will further disrupt the biota of the landscape. Ironically, combating climate change with the advancement of solar power development is currently the most significant emerging threat to restoration of longleaf ecosystems in Virginia. The Virginia Clean Economy Act, signed into law in 2020, establishes that 16,100 megawatts of solar and onshore wind is "in the public interest." Many of the sites desirable for longleaf restoration are also desirable for solar development.

Recognizing resources for land protection will continue to be limited, it is imperative we focus our land acquisition work on the most biologically important tracts. To that end, ecologists with the Virginia Department of Conservation & Recreation's Natural Heritage Program (VANHP) developed a habitat model for longleaf pine in Virginia, using established methods developed for modeling of rare, threatened and endangered species in Virginia. The longleaf pine habitat model combines geographic information on longleaf and longleaf-associated biodiversity with environmental variables, to score the landscape with a suitability score for restoration of the longleaf pine and its associated biodiversity.

This work is intended to support restoration of longleaf pine communities and provide information for the Southeast Longleaf Ecosystems Occurrences Geodatabase (LEO GDB), a project underway to develop a comprehensive map database of documented longleaf pine locations and ecological conditions across the range. The Florida Natural Areas Inventory (FNAI) is working in partnership with the Longleaf Alliance to build the LEO GDB with funding from the Natural Resources Conservation Service (NRCS) via the U.S. Endowment for Forestry and Communities, and in close conjunction with America's Longleaf Restoration Initiative - Longleaf Partnership Council, and other partners. This range-wide effort is modeled after the Florida Longleaf Pine Geodatabase, created by the Florida Forest Service and FNAI, which houses data for almost 2 million acres of existing longleaf pine in Florida. The LEO GDB will enable partners to track longleaf acres and condition, and will be useful in conservation and cost-share planning efforts at local and regional scales.

Methods

Study area and Element Occurrence (EO) selection

In Virginia, three rivers, the Nottoway, the Meherrin, and the Blackwater, form the majority of the Chowan River drainage, and they are believed to hold almost the entirety of the northernmost range of longleaf pine (Frost 1995, 1998; Ware et al. 1993). To define the area of interest and modeling extent, we selected the four 8-digit hydrologic units in the Chowan River drainage in Virginia from the Watershed Boundary Dataset (U.S. Geological Survey, 2018). We buffered the drainage boundary by two miles and clipped the result to the Virginia border, to define the final project area (Figure 1).

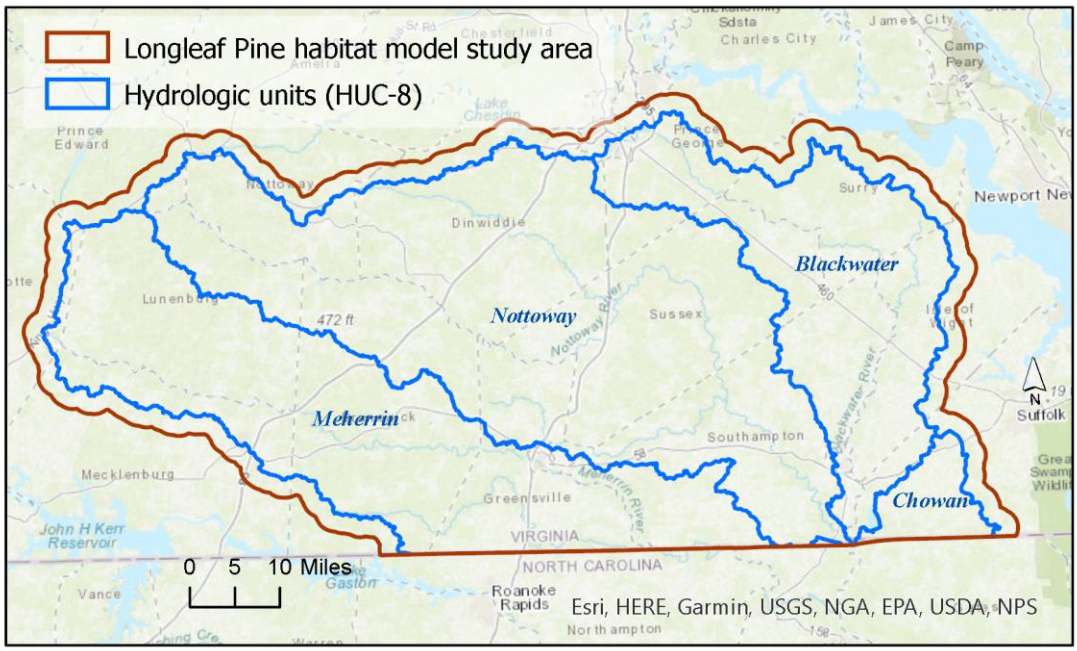


Figure 1. Study area for the longleaf pine habitat model in Virginia, encompassing a 2-mile buffer around the Chowan River drainage.

A full list of Element Occurrences (EOs) was generated for the Chowan drainage in Virginia. Element Occurrences are geographically-delineated occurrences of rare plants, rare animals, and significant natural communities, the elements of biodiversity that are collectively known as Natural Heritage Resources (NHR). An EO delineates the area of land and/or water where an NHR was observed, and represents the habitat of the observed population. EOs are ranked by rarity at state and global levels, and by viability.

The list of EOs was generated by performing a spatial join between the Virginia statewide Element Occurrences layer and the four project area HUC-8 drainages (Figure 1). The resulting EOs were then exported to an excel spreadsheet. Occurrences with a 'Last Observed' date prior to 1970 were removed from the dataset. These records were removed because their mapped locational accuracy was questionable and/or the habitat may have been markedly altered in the last 50 years.

From the full list of 858 EOs within this watershed, each was evaluated to determine if it was known to occur with longleaf pine in all or part of its range. Examples include longleaf pine itself, pitcher plants (*Sarracenia flava* and *S. purpurea*), Red-cockaded Woodpeckers (*Picoides borealis*), and remnants of bog vegetation. Using this criteria, a total of 408 EOs (from 112 unique NHR) were identified for consideration as habitat model training data.

Training data

An EO can be made up of one or more polygons, called Procedural Features (PF). To develop a dataset to train the model, we selected the PFs associated with the EOs previously selected. We excluded 113 PFs which were either of low accuracy, or associated with EOs for introduced populations. A total of 776 PFs (from 345 EOs) remained after exclusions.

VANHP maintains an ArcGIS toolbox[1] for processing occurrence polygons (e.g., PFs) to use as habitat model training data polygons. The primary steps in the workflow include: adding and calculating a set of attributes for the polygons, excising areas of overlap from polygons with lower accuracy and/or earlier observation dates, and assigning a group identifier to each polygon. Using this workflow for longleaf pine PFs, we developed a training polygon dataset including 665 polygons, from 320 groups. The original EO identifier was retained to use as the grouping identifier. This final training data polygon dataset was reduced to include only those attributes used in the habitat modeling process; key fields are listed in Table 1.

Table 1. Training data polygon dataset attributes used in the habitat modeling process.

Attribute	**Attribute description**
GROUP_ID	Unique identifier for polygon group. Inherited from the Element Occurrence ID.
RA	Representation Accuracy of the polygon, reflecting confidence in spatial accuracy. Values range from 1 (very low) to 5 (very high).
OBSDATE	Observation date associated with the species occurrence.

Habitat model

VANHP has established standard Species Habitat Modeling methods, which have been used to develop models and Predicted Suitable Habitat maps for 179 rare and threatened or endangered species through August 2021[2]. VANHP also maintains a GitHub repository[3] containing scripts used to execute habitat models and develop a standard metadata document, developed in collaboration with partners at several other state Natural Heritage programs. This report provides a brief overview of key steps in the modeling process. All modeling procedures are run in the R statistical software environment (R Core Team, 2021).

[1] https://github.com/VANatHeritage/SDM-PresencePreProc
[2] https://www.dcr.virginia.gov/natural-heritage/sdm
[3] https://github.com/VANatHeritage/Virginia_SDM

The first step in the modeling process was generating point sample locations within the training polygons ('presences'), and in all other parts of the study area ('pseudo-absences'). The samples were attributed with values from a set of raster environmental variables describing climate, topography, geology, hydrography, and land cover, at a resolution (raster cell size) of 30-meters. VANHP maintains a standard set of 80+ environmental variables for use in habitat modeling of terrestrial NHR. From this initial set, we excluded those for impervious surface percentage, and land cover variables which represent distance to a certain land cover types, which were not expected to be useful for modeling of longleaf pine habitat. This resulted in a total of 66 variables entering into the model. For land cover variables, values assigned to samples were from the time period closest to the training polygon's observation date (OBSDATE attribute).

Samples were used in a three-step model-building procedure, using the random forest machine-learning algorithm, implemented with the R package *randomForest* (Liaw and Wiener, 2002). Random forest builds a large number of individual classification and regression models ('trees') from random subsets of the input samples and environmental variables, and the final model is an ensemble of the individual trees. In all models, a sampling scheme is implemented to select a higher number of points originating from polygon groups with higher representation accuracy values (RA; a measure of confidence in the spatial location).

In the first step of the model-building procedure, we created an initial model with 1000 trees, and used the 'mean decrease in accuracy' metric calculated by *randomForest* to rank variables. Variables which were not top-ranked within their correlated variable groups were removed, after which the top 75th-percentile of remaining variables were retained for use in subsequent steps (n=44). In the second step, we carried out a 'leave-one-out' cross-validation procedure across groups (GROUP_ID attribute). Here, a random forest model with 1000 trees was built with points from all but one group, and then tested to determine if that model could predict the points in the excluded group as suitable habitat. A set of statistics was calculated for each iteration (n=320) of the cross-validation, which were used to evaluate the model's performance and calculate a set of thresholds. In the third and final step, a 'full' model with 2000 trees was built using all samples. The full model was used to develop model predictions and evaluate variable importance and contribution to the model.

Model products, review and post-processing

The full model was used to develop a prediction raster, where each 30-meter cell for the raster in the study area was assigned a probability value between 0 and 1, with higher values indicating higher probability of suitable habitat for longleaf pine. A 'threshold' version of the prediction raster was also created, in which cells were classified using the set of seven calculated threshold values; a cell is assigned the value of the highest threshold value which it exceeds.

The prediction raster and threshold raster were then posted on ArcGIS Online for review, where a biologist selected two thresholds best representing suitable lands for restoration of biologically-diverse longleaf pine communities. For these thresholds, we generated a raster dataset, where

cells above the higher threshold were classified as "primary restoration areas" (i.e. the best candidates for restoration), and cells between the two thresholds were classified as "secondary restoration areas". We then used the National Land Cover Database 2019 (NLCD; Yang et al., 2018) to remove any areas classified as developed in 2019. Given legacy impacts of agriculture on ecological systems (Foster et al., 2003), and specifically on soil and understory plant composition in longleaf pine woodlands (Brudvig et al., 2013), we removed any lands that were in agricultural usage (pasture or cultivated crops) for any NLCD period from 2001-2019. Additionally, we used the National Hydrography Dataset High Resolution (NHD; U.S. Geological Survey, 2018) *NHDArea* and *NHDWaterbody* feature classes to remove areas of open water, marsh, or swamp. The remaining areas were converted to polygons and assigned a set of attributes, including unique identifiers and acreages for fragments (contiguous polygons) and complexes (groups of fragments within 50-meters of one another). In a final step, we merged polygons in the same complex and class, resulting in the final dataset *Predicted Locations for Restoring Biologically Diverse Longleaf Pine Communities* (Table 2).

Table 2. Attributes of the polygon layer, *Predicted Locations for Restoring Biologically Diverse Longleaf Pine Communities*.

Attribute	**Attribute description**
complex_ID	Unique ID for a given complex (a group of fragments separated by less than 50-meters)
complex_acres	Acreage of the entire complex
class	Restoration class: Primary restoration areas are those with model prediction values above the higher threshold. Secondary areas are those with model prediction values at or above the lower threshold only.
model_info	Information about the habitat model and threshold used
acres	Acreage of the polygon (i.e., acreage of the class in the complex)
complex_area_flag	Indicates if the complex's total area is ≥25 acres (1) or <25 acres (0).

Results

Habitat model

The longleaf pine habitat model metadata is included as an appendix to this report. It includes cross-validation results, a table of threshold values used to classify the continuous prediction raster, variables used in the final model, variable importance rankings, and variable response curves. Based on the cross-validation procedure, model performance was high, as indicated by various metrics including TSS (0.82), AUC (0.98), Kappa (0.82), and overall accuracy (0.91).

Ranked by mean decrease in accuracy in the full model, the top three variables were Shrub cover 100-cell mean (reflecting landscape-scale shrub coverage), normalized dispersion of precipitation, and June precipitation (Appendix 1, Figure 2). The partial dependence plots for these variables (Appendix 1, Figure 3) show the relationship between these variables and the model prediction.

Seven thresholds were calculated during the cross-validation procedure (Appendix 1, Table 3). They encompassed a wide range of values, from 0.278 for Minimum Training Presence, to 0.776 for Minimum Training Presence by Group. The threshold version of the prediction raster is shown in Figure 2.

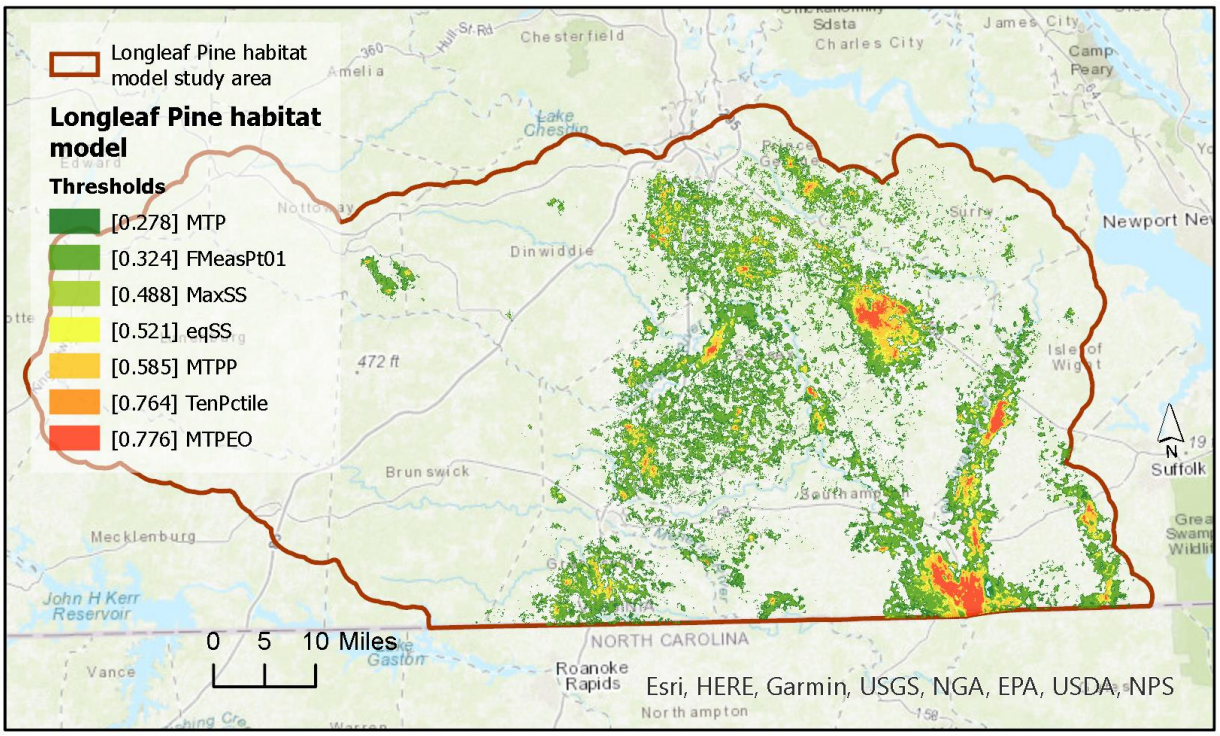


Figure 2. Threshold version of the longleaf pine habitat model prediction raster. See the model metadata (Appendix 1, Table 3) for threshold descriptions.

Model review and threshold selection

During review, two thresholds were selected to delineate areas with high probability for restoration of longleaf pine communities. The higher of the thresholds is Minimum Training Presence by Polygon (MTPP; 0.585), which is the maximum probability value where each training polygon has at least one point sample classified as suitable. The lower threshold selected is Maximum of Sensitivity plus Specificity (MaxSS; 0.488), which is the probability at which the sum of sensitivity (proportion of correctly classified presence training points) and specificity (proportion of correctly classified pseudo-absence training points) is maximized.

In the final dataset, areas at or above the higher threshold (Primary restoration areas) include 37,913 acres. Of the Element Occurrences used for the model, 326 of 345 (94.5%) are covered at least partly by Primary restoration areas, including all Longleaf Pine EOs (both for the species and communities including longleaf). Areas between the lower and upper threshold (Secondary restoration areas) include 31,056 acres, and cover two additional EOs. Note that most of the EOs which did not intersect the final polygon layer are in areas removed through exclusion of developed, recent agricultural, and swamp areas.

The attribute 'complex_area_flag' in the final dataset indicates polygons which are part of complexes with at least 25 acres in size (1) or less than 25 acres (0), as we consider 25 acres a recommended minimum area for restoration of longleaf pine communities. Figure 3 displays a map of the final dataset, displaying only complexes at least 25 acres in size.

Discussion

We developed a habitat model for longleaf pine in Virginia, using occurrences of longleaf pine, as well as for a large suite of associated species known to occur with longleaf pine in all or part of its range. The model can be interpreted to predict not only suitable habitat for extant longleaf pine, but more generally the probability of suitable restoration areas for longleaf pine communities including associated rare species.

In the final output dataset, we provide two classification levels to distinguish between locations with higher and lower potential for restoration. Primary restoration areas delineate lands where there is the highest likelihood that longleaf could be successfully restored along with a potential for capturing additional elements of biodiversity including remnant understory vegetation, declining species which utilize longleaf pine systems, and restorable natural communities. These areas encompass over 90% of the Element Occurrences selected for use in the model, as well as other "natural lands" likely to have ground-cover native to the longleaf system. Secondary Restoration areas delineate additional lands where longleaf could be successfully restored, and may have some potential to capture additional elements of biodiversity associated with longleaf. These areas may highlight important buffer areas around Primary restoration areas, and/or corridors connecting multiple Primary restoration areas.

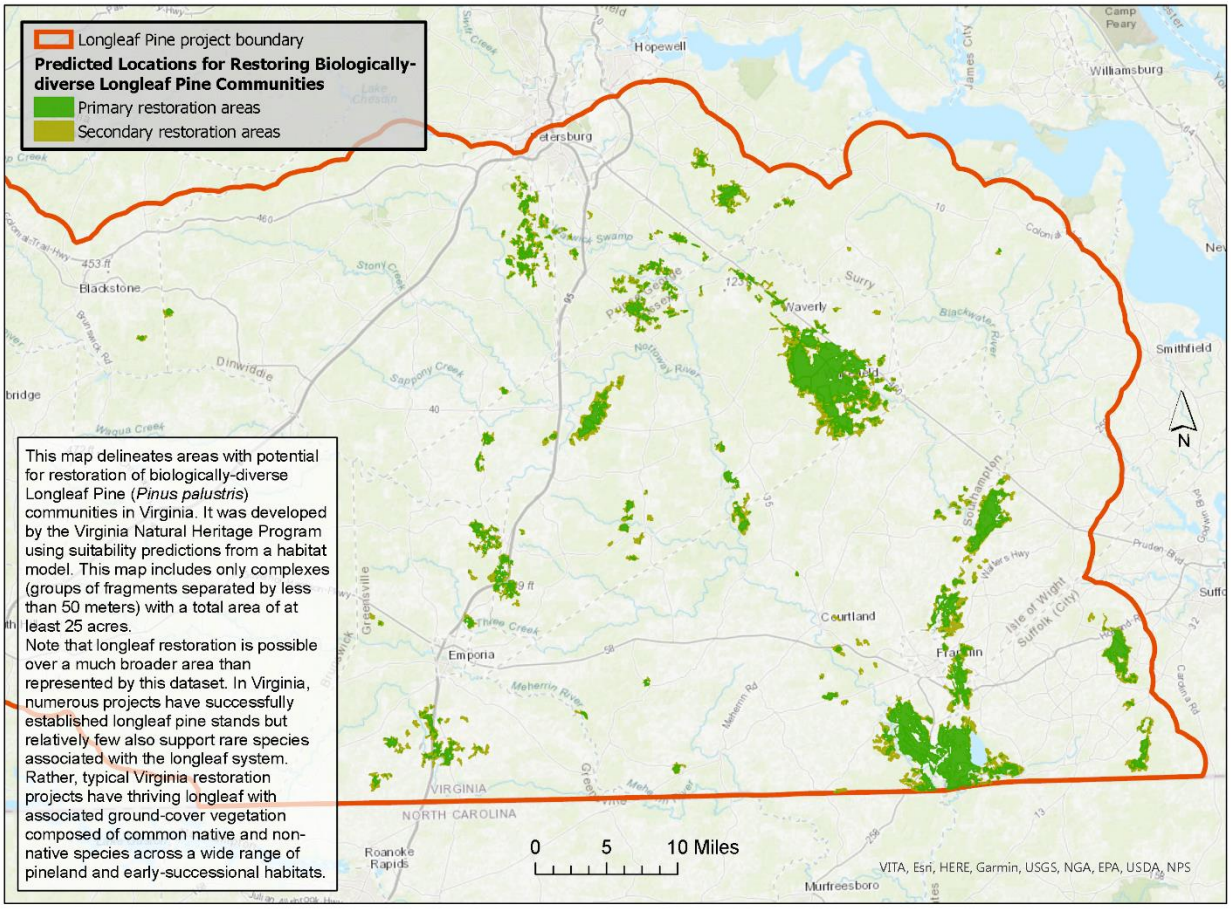


Figure 3. *Predicted Locations for Restoring Biologically Diverse Longleaf Pine Communities*, displaying only those complexes at least 25 acres in size.

Note that longleaf restoration is possible over a much broader area than represented by this dataset. In Virginia, numerous projects have successfully established longleaf pine stands, but relatively few also support rare species associated with the longleaf system. Rather, typical Virginia restoration projects have thriving longleaf with associated ground-cover vegetation composed of common native and non-native species across a wide range of pineland and early-successional habitats.

References

Brudvig, L., E. Grman, C. W. Habeck, J. L. Orrock, and J. A. Ledvina. 2013. Strong Legacy of Agricultural Land Use on Soils and Understory Plant Communities in Longleaf Pine Woodlands. *Forest Ecology and Management* 310 (December): 944–55.

Esri. 2020. ArcGIS Pro (Version 2.7) (Software). Esri Inc. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview.

Foster, D., F. Swanson, J. Aber, I. Burke, D. Brokaw, D. Tilman, and A. Knapp. 2003. The Importance of Land-Use Legacies to Ecology and Conservation. *BioScience* 53 (1): 77–88.

Frost, C.C. 1995. Presettlement fire regimes in southeastern marshes, peatlands, and swamps. Pp. 39–60 in S.I. Cerulean and R.T. Engstrom (eds.). *Proceedings of the Tall Timbers fire ecology conference*, No. 19. Tallahassee, FL: Tall Timbers Research Station.

Frost, C.C. 1998. Presettlement fire frequency regimes of the United States: a first approximation. Pp. 70–81 in T.L. Pruden and R.I. Engstrom (eds.). *Proceedings of the 20th Tall Timbers fire ecology conference*. Tallahassee, FL: Tall Timbers Research Station.

Liaw, A. and Wiener, M. 2002. Classification and Regression by randomForest. *R News* 2(3), 18-22.

Mittermeier, R.A., N. Myers, and C. G. Mittermeier. 2000. Hotspots: Earth's Biologically Richest and Most Endangered Terrestrial Ecoregions, Conservation International. ISBN 978-968-6397-58-1.

Myers, N. 1988. Threatened biotas: "Hot spots" in tropical forests. *The Environmentalist* 8: 187-208.

Myers, N. 1990. The biodiversity challenge: Expanded hot-spots analysis. *The Environmentalist* 10: 243-256.

Myers, N., R. A. Mittermeier., C. G. Mittermeier, G. A. B. da Fonseca, and J. Kent. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403: 853-858.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

U.S. Geological Survey. 2019. National Hydrography Dataset (ver. USGS National Hydrography Dataset Plus High Resolution (NHDPlus HR) for 4-digit Hydrologic Unit - 0301 (published 20180503)), accessed July 10, 2019 at URL https://www.usgs.gov/core-science-systems/ngp/national-hydrography/access-national-hydrography-products.

Virginia Department of Forestry. 2014. From the Brink! The Effort to Restore Virginia's Native Longleaf Pine, 2014 Status Report. Charlottesville, VA: Virginia Department of Forestry. 21pp.

Ware, S., C. Frost, and P.D. Doerr. 1993. Southern mixed hardwood forest: the former longleaf pine forest. Pp. 447–493 in W.H. Martin, S.G. Boyce, and A.C. Echternacht (eds.). Biodiversity of the southeastern United States: lowland terrestrial communities. New York: John Wiley & Sons Inc.

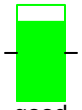
Yang, Limin, Suming Jin, Patrick Danielson, Collin Homer, Leila Gass, Stacie M. Bender, Adam Case, et al. 2018. "A New Generation of the United States National Land Cover Database: Requirements, Research Priorities, Design, and Implementation Strategies." *ISPRS Journal of Photogrammetry and Remote Sensing* 146 (December): 108–23.

Appendix 1. Longleaf Pine habitat model metadata

(Begins on following page)

Pinus palustris

Species Habitat Model (SHM) assessment metrics and metadata
Common name: Longleaf pine
NatureServe Grank/Srank: G5-Secure / S1-Critically Imperiled
Code: pinupalu (EGT_ID: 152746)
Date: 10 Jul 2021



good
TSS=0.82
validation success

The following metadata describes a Species Habitat Model (SHM) for a species tracked by the Virginia Natural Heritage Program. This SHM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1,2] in the R statistical environment [3,4]. We validated the model by jackknifing (also called leave-one-out [5,6,7]) by polygon group for a total of 320 groups. The statistics in Table 2 report the mean and variance of validation statistics for these jackknifing runs.

Table 1. Input statistics. Presence points are points placed in polygon-based location information or point-based observations. Groups describe groupings of points based on polygon data or spatial grouping of observations. Background points are placed throughout model area excluding known species locations. In cases of fewer than 5 groups, cross-validation is performed by-polygon.

Name	Number
Presence points	9751
Polygons(Groups)	665(320)
Background points	10000

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve [8,9,6].

Name	Mean	SD	SEM
Overall Accuracy	0.91	0.14	0.01
Specificity	0.91	0.12	0.01
Sensitivity	0.92	0.27	0.01
TSS	0.82	0.29	0.02
Kappa	0.82	0.29	0.02
AUC	0.98	0.08	0.00

Validation runs used 41 environmental variables, the most important of 66 variables (top 75 percent). Each tree was built with 2 variables tried at each split (mtry) and 1000 trees built. The final model was built using 2000 trees, all presence and background points, with an mtry of 2, and 41 environmental variables.

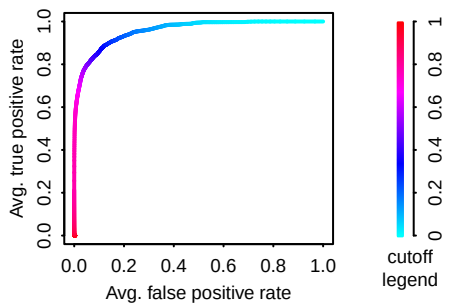


Figure 1. ROC plot for all 320 validation runs, averaged along cutoffs.

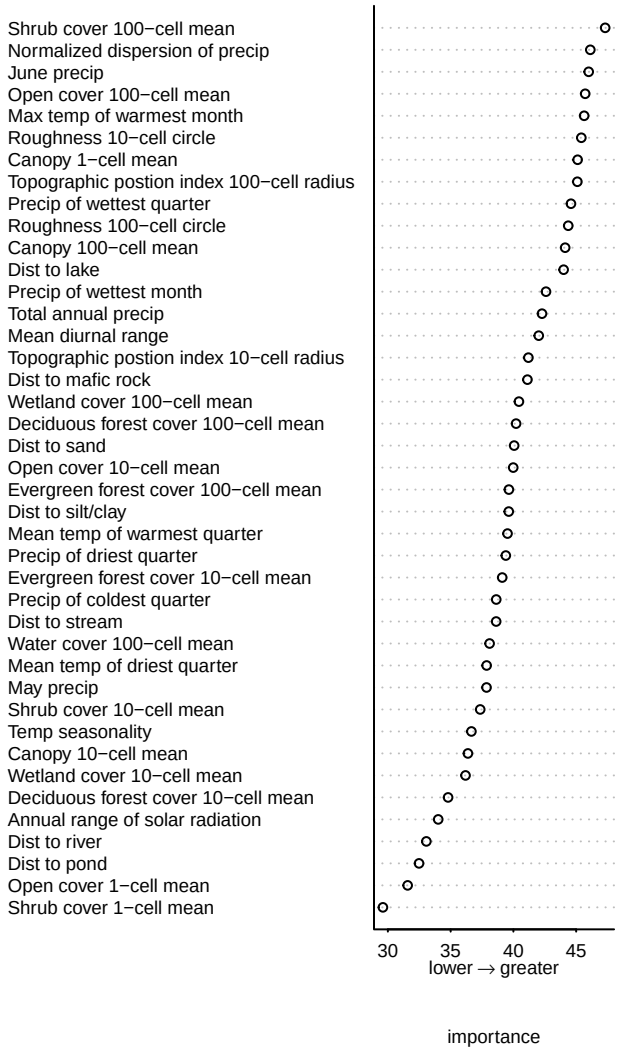


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Importance values (mean decrease in accuracy) are extracted from the randomForest [3] function. See Appendix 1 for variable descriptions.

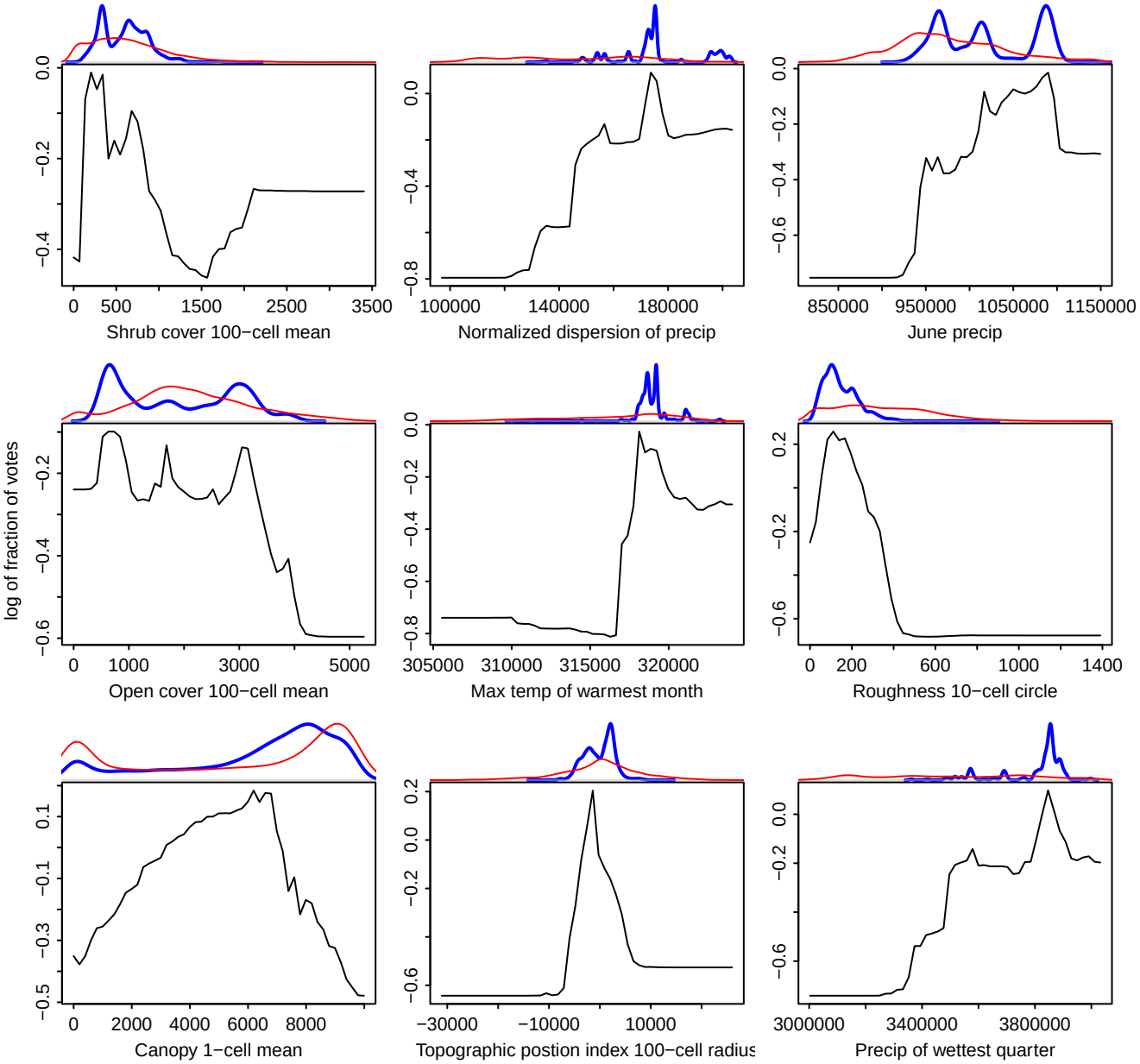


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed[3]. The x-axis covers the range of values for the variable assessed; the y-axis represents the effect between the variable and model response. Peaks in the black line indicate where this variable had the strongest influence on predicting appropriate habitat. Decreasing lines from left to right show a negative relationship overall; increasing lines, positive. The distribution of each category (thin red = Background points, thick blue = Presence points) is depicted at the top margin. See Appendix 1 for variable descriptions.

Species habitat model outputs display the probability (0-1) of a location (i.e. stream reach or raster cell) having similar environmental conditions in comparison to known presence locations. No model will ever depict sites where a targeted element will occur with certainty, it can *only* depict locations it interprets as appropriate habitat for the targeted element. The delineation of suitable habitats is made by the selection of a threshold value, where locations with values above the threshold are designated as likely suitable habitat, and those with values below the threshold may be unsuitable. Threshold values are often statistically calculated. SHMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an SHM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower threshold depicting more area may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less area, a higher threshold may be more appropriate. Different thresholds for this model (full model) are described in Table 3.

Table 3. Thresholds[11,12] calculated from the final model. The Value column reports the threshold; Groups indicates the percentage (number in brackets) of groups within which at least one point was predicted as suitable habitat; Pts indicates the percentage of PR points predicted having suitable habitat. Total numbers of groups and presence points used in the final model are reported in Table 1.

Threshold	Value	Groups	Pts	Description
Equal sensitivity and specificity	0.521	100(320)	98.8	The probability at which the absolute value of the difference between sensitivity and specificity is minimized.
Maximum of sensitivity plus specificity	0.488	100(320)	99.4	The probability at which the sum of sensitivity and specificity is maximized.
Minimum Training Presence	0.278	100(320)	100	The highest probability value at which 100% of input presence points remain classified as suitable habitat.
Minimum Training Presence by Polygon	0.585	100(320)	97.3	The highest probability value at which 100% of input polygons have at least one presence point classified as suitable habitat.
Minimum Training Presence by Group	0.776	100(320)	88.7	The highest probability value at which 100% of input groups have at least one presence point classified as suitable habitat.
Tenth percentile of training presence	0.764	100(320)	90	The probability at which 90% of the input presence points are classified as suitable habitat.
F-measure with alpha set to 0.01	0.324	100(320)	100	The probability value at which the harmonic mean of precision and recall, with strong weighting towards recall, is maximized.

Model Evaluation and Intended Use

All SHMs are sensitive to data inputs and methodological choices. Table 4 presents scoring of modeling factors based on the model evaluation rubric presented in Sofaer et al. 2019[13].

Table 4. Model evaluation results based on Sofaer et al. 2019. Scores can be attributed as ideal, acceptable, or interpret with caution.

Category	Metric	Score	Notes
Species Data	Presence data quality	Ideal	Heritage Network data are vetted for accuracy and weighted for spatial representation.
	Absence/Background Data	Acceptable	Background points randomly placed throughout study area excluding species locations.
	Evaluation Data	Acceptable	Models are validated by jackknifing (i.e. leave-one-out).
Environmental Predictors	Ecological and predictive relevance	Acceptable	Selection of predictor variables were based on previous modeling experience by the Natural Heritage Network. Time constraints of this project prevented making species specific selections.
	Spatial and temporal alignment	Acceptable	Reasonable attempts to align predictor and presence data were made.
Modeling Process	Algorithm choice	Acceptable	Random Forest is highly rated classification model that is well documented as suitable for modeling rare species.
	Sensitivity	Acceptable	Settings for Random Forest were adjusted to best model the species; however, different models/parameters were not tested within one model run.
	Statistical rigor	Acceptable	Collinearity of predictors recognized and addressed; presence points grouped to minimize sample bias and minimize spatial autocorrelation boost during validation; other assumptions recognized and considered.
	Performance	Acceptable	Model TSS $\geq$ 0.6. Mapped model output is evaluated for ecological plausibility by expert review.
	Model review	Interpet with Caution	Model was not reviewed by regional, taxonomic experts.
Model Products	Mapped products	Acceptable	Single calculated threshold selected for the final model.
	Interpretation support products	Ideal	All standards met.
	Reproducibility	Ideal	All standards met.
	Iterative	Interpet with Caution	Model not revised.

Model Comments

This model was developed to identify potential suitable sites for longleaf pine restoration. In addition to known longleaf locations, the model training data also includes indicator species for longleaf pine communities. The standard variables (impsur, dnw) were excluded from this model.

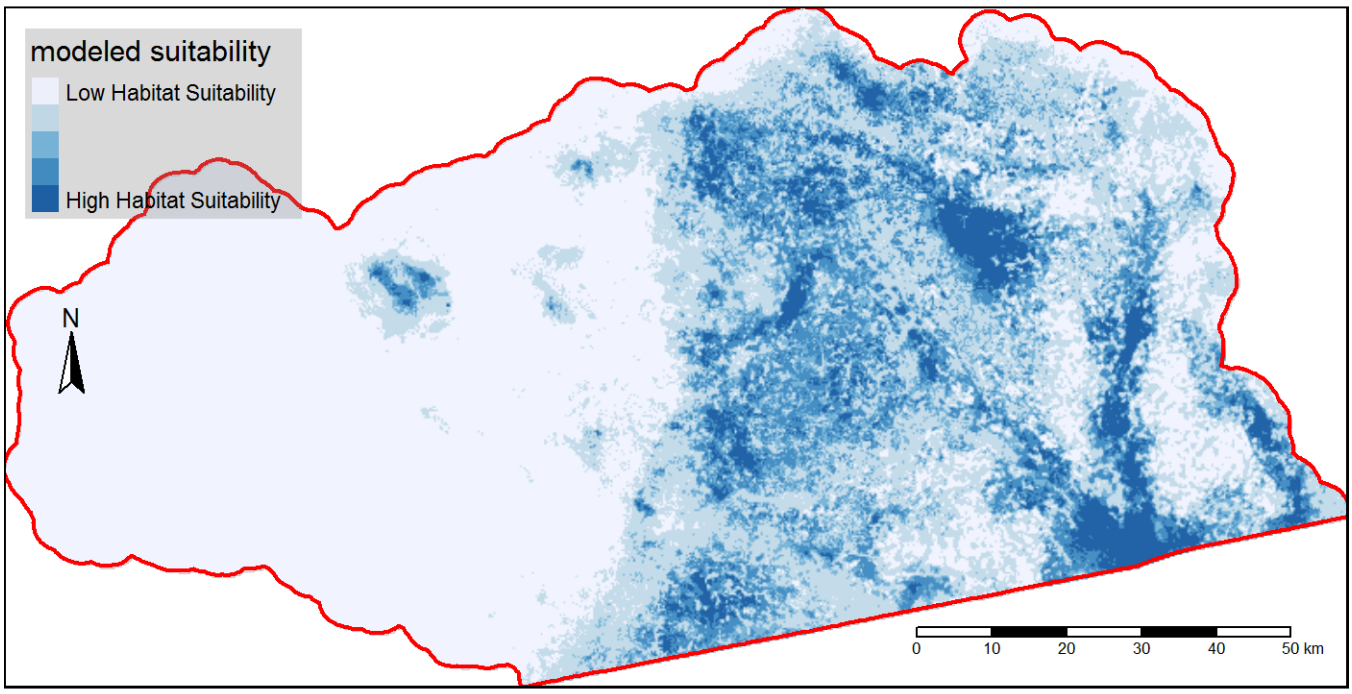


Figure 4. A generalized view of the model predictions throughout the modeled area. State boundaries are depicted as a thin gray line. The modeled area is outlined in red. Basemap: Esri World Topographic Map (©2021 Esri).

This habitat model would not have been possible without data sharing among organizations. Other data sets and sources may have been evaluated, but this final model includes data from these sources:

- Virginia Natural Heritage Program

This model was built using a methodology developed through collaboration among the Florida Natural Areas Inventory, the New York Natural Heritage Program, the Pennsylvania Natural Heritage Program, and the Virginia Natural Heritage Program, all member programs of the NatureServe Network. It is one of a suite of species habitat models developed using the same methods, scripts, and environmental data sets. Our goal was to be consistent and transparent in our methodology, validation, and output.

Please cite this document and its associated SHM as:
Virginia Natural Heritage Program. 2021. Species habitat model for Longleaf pine (*Pinus palustris*). Created on 10 Jul 2021. Virginia Department of Conservation and Recreation - Division of Natural Heritage, Richmond, VA.

References

[1] Breiman, L. 2001. Random forests. Machine Learning 45:5-32.
[2] Iverson, L. R., A. M. Prasad, and A. Liaw. 2004. New machine learning tools for predictive vegetation mapping after climate change: Bagging and Random Forest perform better than Regression Tree Analysis. Landscape ecology of trees and forests.Proceedings of the twelfth annual IALE (UK) conference, Cirencester, UK, 21-24 June 2004 317-320.
[3] Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18-22. Version 4.6-14.
[4] R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. R version 4.1.0 (2021-05-18).
[5] Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38-49.
[6] Fielding, A. H. 2002. What are the appropriate characteristics of an accuracy measure? Pages 271-280 in Predicting Species Occurrences, issues of accuracy and scale. J. M. Scott, P. J. Helglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, F. B. Samson, eds. Island Press, Washington.
[7] Pearson, R.G. 2007. Species Distribution Modeling for Conservation Educators and Practitioners. Synthesis. American Museum of Natural History. Available at http://ncep.amnh.org.
[8] Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43:1223-1232.
[9] Vaughan, I. P. and S. J. Ormerod. 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology 42:720-730.
[10] Sing, T., O. Sander, N. Beerenwinkel, T. Lengauer. 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21(20):3940-3941.
[11] Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28:385–393.
[12] Liu, C., G. Newell, and M. White. 2015. On the selection of thresholds for predicting species occurrence with presence-only data. Ecology and Evolution 6:337–348.
[13] Sofaer, H. R., C. S. Jarnevich, I. S. Pearse, R. L. Smyth, S. Auer, G. L. Cook, T. C. Edwards, Jr., G. F. Guala, T. G. Howard, J. T. Morisette, and H. Hamilton. (In press). The development and delivery of species distribution models to inform decision-making. BioScience.

Appendix 1. Descriptions for environmental variables included in model.

Variable Name	Variable Description
Annual range of solar radiation	Difference between Summer and Winter solstice total insolation derived from direct and diffuse, but not reflected, radiation [radsum-sol - radwinsol]
Canopy 1-cell mean	mean percent canopy cover in 1-cell radius (30 meter cells)
Canopy 10-cell mean	mean percent canopy cover in 10-cell radius (30 meter cells)
Canopy 100-cell mean	mean percent canopy cover in 100-cell radius (30 meter cells)
Deciduous forest cover 10-cell mean	mean deciduous forest cover within 10-cell radius
Deciduous forest cover 100-cell mean	mean deciduous forest cover within 100-cell radius
Dist to lake	Euclidean distance to nearest lake/pond/resevoir ¿ 1 ha
Dist to mafic rock	Euclidean distance to mafic bedrock
Dist to pond	Euclidean distance to nearest lake/pond/resevoir ¡= 1 ha
Dist to river	Euclidean distance to nearest stream/river
Dist to sand	Euclidean distance to sand
Dist to silt/clay	Euclidean distance to silt/clay
Dist to stream	Euclidean distance to nearest stream (features represented by lines only)
Evergreen forest cover 10-cell mean	mean evergreen forest cover within 10-cell radius
Evergreen forest cover 100-cell mean	mean evergreen forest cover within 100-cell radius
June precip	June precipiation
Max temp of warmest month	maximum temperature of warmest month
May precip	May precipitation
Mean diurnal range	(mean of monthly (max temp - min temp))
Mean temp of driest quarter	mean temperature of driest quarter
Mean temp of warmest quarter	mean temperature of warmest quarter
Normalized dispersion of precip	normalized dispersion (CV) of precipitation
Open cover 1-cell mean	mean open cover within 1-cell radius
Open cover 10-cell mean	mean open cover within 10-cell radius
Open cover 100-cell mean	mean open cover within 100 cell radius
Precip of coldest quarter	precipitation of coldest quarter
Precip of driest quarter	precipitation of driest quarter
Precip of wettest month	precipitation of wettest month
Precip of wettest quarter	precipitation of wettest quarter
Roughness 10-cell circle	The standard deviation of elevation values within a circular neighborhood with a radius of 10 cells.
Roughness 100-cell circle	The standard deviation of elevation values within a circular neighborhood with a radius of 100 cells.
Shrub cover 1-cell mean	mean shrub cover within 1-cell radius
Shrub cover 10-cell mean	mean shrub cover within 10-cell radius
Shrub cover 100-cell mean	mean shrub cover within 100 cell radius
Temp seasonality	(STD * 100)
Topographic postion index 10-cell radius	Topographic position index using elevation values within a circular neighborhood with a radius of 10 cells.
Topographic postion index 100-cell radius	Topographic position index using elevation values within a circular neighborhood with a radius of 100 cells.
Total annual precip	total annual precipitation
Water cover 100-cell mean	mean open water cover within 100 cell radius
Wetland cover 10-cell mean	mean wetland cover within 10-cell radius
Wetland cover 100-cell mean	mean wetland cover within 100 cell radius

Appendix 2. Model details for reproducibility

- All R Scripts are available at github
- The repository branch:head used for this run was: `terrestrial:985b81d5`
- Validation metrics requiring a threshold use MTP (minimum training presence)
- R version: R version 4.1.0 (2021-05-18)
- Random seed for full randomForest model: 709174600
- randomForest mtry: 2
- randomForest ntrees: 2000