# AUTOMATED IDENTIFICATION OF WETLANDS USING GIS IN NORTH CAROLINA



August 2021

By Susan Gale, Environmental Program Consultant

North Carolina Dept. of Environmental Quality, Div. of Water Resources

Completed as partial fulfillment of US EPA Multipurpose Grant AA-01D03020

# Contents

Introduction
Study objectives
Methods
General approach6
Model training and testing data7
Overview of variables used as model inputs8
Modeling methods10
Overlay model methods10
MaxEnt model methods11
Wetland Identification Method (WIM) model methods12
Software13
Geoprocessing13
Classification accuracy assessment methods13
Results and Discussion
MaxEnt models results
ESPL Watland Identification Model (WIM) results
References

# Tables

Table 1 Summary of model variables	9
Table 2 Descriptions of the different overlay models tested	
Table 3. Contingency table example	14
Table 4 Variable contributions to MaxEnt model complete model (run75)	
Table 5 Variable contributions to MaxEnt model run 76 (minimal model)	23
Table 6 Accuracy statistics for the NOP targeted study area for MaxEnt and overlay models	26
Table 7 Comparison of Producer's Accuracy and User's Accuracy for NWI (baseline) and NCDOT	wetland model
	29

# Figures

Figure 3 Map of NC showing Level III ecoregions and NOP targeted study area	6
Figure 4. Location of field-delineated data by source agency	7
Figure 5. Relative density of field-delineated features	8
Figure 6 Odds ratios for overlay models by ecoregion and statewide	15
Figure 7 Producer's Accuracy for wetlands classification (PA <sub>WL</sub> ) for overlay models by Level III ecoregion	16
Figure 8 User's Accuracy for wetlands classification (UA <sub>WL</sub> ) for overlay models by Level III ecoregion	16
Figure 9 Odds ratio for overlay models by groundtruthed feature size	17
Figure 10 Producer's Accuracy for wetlands classification (PA <sub>WL</sub> ) for overlay models by field-delineated wet	land
size class	18
Figure 11 User's Accuracy for wetlands classification (UA <sub>WL</sub> ) for overlay models by field-delineated wetland	l size
class	18
Figure 12 An extreme example of banding due to model overfitting in spatial model results (left) and	
precipitation contours for the same area (right) in the northern Piedmont of NC	19
Figure 13 Location of field-verified training and testing points within NOP targeted study area	20
Figure 14 Spatial output of MaxEnt model run 75 (complete model)	21
Figure 15 Spatial output of MaxEnt model run 76 (minimal model)	21
Figure 16 Example of model outputs, showing comparison of NWI coverage versus model predictions by be	oth
MaxEnt models	22
Figure 17 Producer's Accuracy and User's Accuracy for nwi and for various thresholds (percentiles) for Max	ĸEnt
complete model (run 75) and minimal model (run 76)	24
Figure 18 Odds ratio for nwi and for various thresholds (percentiles) for MaxEnt run 75 (complete model) a	and
run 76 (minimal model)	25
Figure 19 Producer's Accuracy and User's Accuracy of nwi, Maxent run 75 (complete model), and MaxEnt r	un
76 (minimal model) by wetland feature size class in NOP focus area	25

# Introduction

### Background

Maps of the location and extent of wetlands are used for many purposes, including: natural resource research and management; environmental impact assessments for transportation planning and other development activities; hydrological and climatological modeling; water quality and watershed assessments; identification of conservation or ecological restoration opportunities, including compensatory mitigation; training and/or verification data for other remote sensing methods for land classification; and outreach/education to the general public. The most commonly used wetlands map in North Carolina (NC) is the National Wetlands Inventory (NWI) and this source has been incorporated into other mapping products, such as the US Geological Survey (USGS) 1:24,000 scale topographic maps and current US Topos, National Resources Conservation Service (NRCS) county soil survey maps, National Land Cover Database (NLCD), and NC Division of Coastal Management (NC DCM) maps developed for the eastern portion of NC (Davis et al. 2019, Sutter 1999, Yang et al. 2018). However, the NWI was not designed to support identification of features commonly considered as wetlands, which are characterized by shallow inundation and saturation sufficient to support hydrophytic-dominated vegetation communities and development of hydric soils.

Development of the National Wetland Inventory (NWI) was begun by the US Fish and Wildlife Service in the 1970's using the data sources, technologies, and wetland classification system available at that time. Mapping in NC was primarily completed using 1980's-era aerial imagery and there have been few updates since its initial mapping. In addition to the age of the data, the utility of the NWI is also limited by a relatively large minimum mapping size (0.5 ac.), relatively coarse resolution of the source imagery, and inclusion of many features (such as deepwater and lotic systems) that do not meet the common definition of "wetland". However, it is the only widely available map of wetland locations and extent within NC and for much of the US, and so is widely used for purposes for which it was not designed. The accuracy of the NWI was recently assessed for NC by NCDWR (NCDWR 2021) and found to be of inconsistent and often poor accuracy, particularly for smaller wetlands (<1.0 ac.). This was particularly problematic in the central and western portions of the state, where wetlands tend to be much smaller than 1.0 ac. in size, and so there were high rates of omission in these areas.

Several past and current initiatives across the US have been taken to facilitate updating the NWI (Figure 1). Many of these projects, particularly the recently completed statewide mapping in Minnesota, have been well-documented (e.g., Ducks Unlimited 2008, 2013, 2016, and 2018; Saint Mary's University of Minnesota 2015, 2018). Updates were completed in accordance with the existing wetland mapping standards for NWI (FGDC 2009), which require manual review of aerial photography and ancillary data sets (such as topography and soils), manual digitization of likely



Figure 1 Recent (blue) and active (hatched) NWI Update projects (US Fish and Wildlife Service 2021)

wetland features, and assignment of a wetland type using the Cowardin classification system (Cowardin 1979). This approach is time- and resource-intensive, though far less so than field mapping. For example, the statewide update for Minnesota took 10 years and cost just over \$7 million (Figure 2) (Kloiber and Macleod 2019). NWI updates still rely on manual photointerpretation and heads-up digitizing within GIS (Herb Bergquist, US FWS, pers. comm., May 19, 2020), but recent projects like those in Minnesota use some level of automated classification and modeling to identify areas with a high likelihood for being a wetland based on soils, topography, hydrology, and other factors. This information complements and helps streamline the manual photointerpretation process, though often requires up-front investment in terms of data acquisition.



Figure 2 Reported schedule and costs for MN statewide update of NWI (from Kloiber 2019)

Other projects have investigated the potential for remote identification of wetland maps outside of the NWI framework. Some are relatively simple and intended to replicate the types of processes currently done on an *ad hoc* basis by natural resource professionals, such as simple overlay analyses in GIS (Martin et al. 2012). Others have taken advantage of the variety of high-resolution data and the relatively cheap and fast computing abilities currently available and applied "big data" modeling approaches, such as machine learning. This latter approach is particularly prevalent in the field of remote sensing (i.e., spectral analysis of aerial and satellite imagery) and often focuses on identification of key spectral signatures for individual vegetation species or communities found in wetlands. Other researchers have focused on using non-spectral spatial data as modeling inputs, such as terrain and surface hydrology characteristics derived from Digital Elevation Models (DEM). These terrain-derived methods are often less resource-intensive, since they focus on creating more generalized models to capture a variety of wetland types based on geomorphologic and hydrologic signatures.

One example of research using non-spectral data for predicting wetland locations was completed within the Great Smoky Mountains National Park in western NC. This research study used the machine learning approach MaxEnt along with a combination of DEM-derivatives, vegetation, and climate variables to develop maps predicting the probability of wetland presence within the park boundaries (Pfennigwerth et al. 2019). Researchers in neighboring Virginia applied another machine learning method (Random Forests) to identify wetlands using primarily terrain derivatives (O'Neil 2018), and their methods have since been integrated into ESRI ArcGIS Pro software as part of the Arc Hydro toolbox (ESRI 2020). The Random Forests modeling approach has also been used by researchers working for the NC Department of Transportation (NCDOT) to identify potential wetlands in transportation project corridors, which is anticipated to help streamline their environmental impact assessments and result in significant savings in time and costs.

These examples have demonstrated the potential for predicting location and extent of wetlands using readily available spatial data in the types of landscapes typically found in NC without the limitations inherent in the required methods associated with updating the NWI. A model predicting location and extent of wetlands in this way could potentially result in a wetland mapping product that is more relevant to the majority of the current users of the NWI. The project described herein was initiated to build on the prior research on remote

identification of wetlands in and near NC and determine the feasibility of this type of approach to support a broader, statewide wetland mapping effort in the future. Updating or creating publicly available wetland maps for NC statewide would be a significant effort, though incorporating automated classification methods would likely provide significant efficiencies as compared to the largely manual process required by the NWI.

## Study objectives

This project examined several different spatially explicit, automated modeling approaches for remote identification of wetlands in NC. Accuracy of the modeling results were then compared to currently available wetland maps (NWI and field-verified wetlands polygons) to identify modeling methods that showed potential for accurately identifying the location and extent of wetlands within NC. It is anticipated that these types of automated modeling methods could result in maps that provide a better representation of wetland types of interest to most researchers, scientists, regulators, and the regulated community than the NWI and be more cost-effective and quicker to produce than the methods required by NWI.

# Methods

# Project area

Overlay models were completed for all of NC but results were also examined by Level III ecoregions (Griffith 2002) (Figure 3) to assess any potential spatial variability. For the machine-learning methods (MaxEnt and WIM) the processing time required to develop models for large areas was much greater, so a smaller, targeted study area was selected based on availability of field delineated data to be used for model testing and training. This targeted study area was comprised of the largest contiguous portion of the Northern Outer Piedmont (NOP) Level IV ecoregion in NC, which begins south of the Raleigh metropolitan area and stretches north to the Virginia state line, east to a major geological fall line where the Piedmont transitions to the Inner Coastal Plain, and west to the NOP boundary with the Triassic Basins and Slate Belt Level IV ecoregions. Much of the southern portion of the targeted study area contains a range of developed land use, including intense development in and around the area of Raleigh. The northern portion of the targeted study area is far less developed, and includes significant forested areas, as well as two major river systems (the Neuse and Tar) that have been impounded to create two major water supply reservoirs (John H. Kerr Reservoir and Falls Lake). The varied conditions within the targeted study area support development of a variety of types and sizes of wetlands.



Figure 3 Map of NC showing Level III ecoregions and NOP targeted study area

# General approach

Three modeling methods were identified that have previously been used for remote identification of wetlands in NC or a neighboring state:

- An unweighted overlay approach used to identify isolated wetlands in the Dougherty Plain region of Georgia (Martin et al. 2012);
- The machine-learning method MaxEnt used to identify wetlands in the Great Smoky Mountains National Park in far western NC (Pfennigwerth et al. 2019);
- A geoprocessing workflow (Wetland Identification Method, or WIM) (ESRI 2020) recently added to ESRI ArcGIS Pro software that is based on the machine learning approach Random Forests and had previously been used to identify wetlands in five different study areas in Virginia (O'Neil et al. 2018).

Additional research papers were reviewed to determine spatial data sources used for modeling wetland locations (including Gritzner 2006, McCauley and Jenkins 2005, RTI International 2011, Sutter 1999, and USEPA

2016). The most commonly used non-spectral data sources were selected for use in this project and included the existing wetland maps (NWI), soils, terrain derivatives (i.e., derived from DEMs, including surface hydrology), vegetation communities, and climate.

Modeling inputs were prepared and models applied in ESRI ArcGIS Pro for the overlay and WIM models. For MaxEnt, values for all inputs for each raster cell were exported to text files for modeling in the MaxEnt software. All model results and wetland presence/absence from field-verified conditions for each raster cell were exported to a text file for further analysis, including calculation of accuracy metrics.

## Model training and testing data

Generally, both presence and absence data (e.g., known locations of wetlands and non-wetlands) are required to assess the accuracy of results from land classification modeling. This type of data is used for model training and assessing the final accuracy of the model results (testing). Many research projects use existing spatial data (including NWI) for their training and testing data. However, the gold standard is field delineations of wetland and non-wetland areas due to their reliance on on-the-ground indicators (such as hydrophytic plants and hydric soils) for definitive wetland identification. This project used a statewide data set of field delineations that had been compiled previously and used to complete a statewide accuracy of NWI (NCDWR 2021). The results from that prior study were used as a baseline for comparison to determine if one or more modeling methods used in this study showed an improvement over NWI.

Spatial data representing field-delineations of wetland and non-wetland areas across NC were obtained from several different state and federal government agencies, including the NC Department of Transportation (NCDOT), NC Division of Water Resources (NCDWR), NC Division of Mitigation Services (NCDMS), and the National Park Service Great Smokey Mountains National Park (NPSGRSM). These data represented just over 103,000 acres of field-verified conditions from across the state, collected between 2001 and 2019 (Figure 4, Figure 5). Details on the preparation and characteristics of this data set were described in a previous report (NCDWR 2021), but the data set consisted of a statewide raster that included the wetland status for each cell (wetland, non-wetland, or null/no data). While all available data were used for model training, only the NCDOT data were used for accuracy assessments due to sources of significant bias found in the other data sources (see NCDWR 2021 for more details).



Figure 4. Location of field-delineated data by source agency



*Figure 5. Relative density of field-delineated features* 

### Overview of variables used as model inputs

A summary of the 22 variables used as model inputs, and which models they were used in, is provided in Table 1. Sources for original spatial data included:

- Soils: NC pre-staged statewide soils geodatabase; USDA NRCS Geospatial Data Gateway, <u>https://datagateway.nrcs.usda.gov/</u>
- Digital Elevation Model (DEM): Statewide 20 ft. seamless DEM; NC State University Libraries, GIS data collection, <a href="http://geodata.lib.ncsu.edu/NCElev/">http://geodata.lib.ncsu.edu/NCElev/</a>
- Vegetation: Statewide 2011 raster of GAP/LANDFIRE National Terrestrial Ecosystems; USGS Gap Analysis program, <u>https://www.usgs.gov/core-science-systems/science-analytics-and-synthesis/gap/</u>
- Climate: NC statewide rasters; PRISM Climate Group, Oregon State University, <u>https://prism.oregonstate.edu/normals/</u>
- NWI: NC pre-staged statewide NWI geodatabase; US Fish and Wildlife Service, <u>https://www.fws.gov/wetlands/Data/Data-Download.html</u>

#### Table 1 Summary of model variables

Short name	Variable	Description	Source and calculation method		MaxEnt	WIM
SOILS		•		•	•	
hydricpa	Hydric soil presence/absence	Hydric if >25% of soil mapping unit (MU) area is comprised of hydric components.	NRCS soils attribute hydclprs	x		
hydric	% hydric	Percent of soil MU area made up of hydric components	NRCS soils, attribute hydclprs		x	
aws25	Available water supply 0-25cm	Total potential water available for plants in top 25cm of soil	NRCS soils attribute aws025wta		x	
aws50	Available water supply 0-50cm	Total potential water available for plants in top 50cm of soil	NRCS soils attribute aws050wta		x	
drain	Soil drainage class	Relative level of soil drainage for MU	NRCS soils attribute drclasswet		x	
TERRAIN DE	RIVATIVES (DEM-deriv	ed)				
elev	Elevation	Elevation above sea level	Raw DEM		х	
sinkdep	Sink depth	Difference in elevation between raw DEM and filled DEM	Sink geoprocessing (GP) tool, then raster math on filled and raw DEM		x	
issink	Sink presence/absence	Located in a hydrologic sink	True if sinkdep > 0	x		
slope	Slope	Steepness of surface	Slope GP tool on filled DEM		х	
asp	Aspect	Topographic aspect	Aspect GP tool on raw DEM		х	
curv	Total curvature	Combined plan and profile curvature	Curvatures GP tool on raw DEM		x	
plan	Plan curvature	Curvature perpendicular to slope (potential for convergent/divergent flow)	Curvatures GP tool on raw DEM		x	
prof	Profile curvature	Curvature parallel to slope (potential for downslope flow)	Curvatures GP tool		x	
tpi	Topographic position index (TPI)	Difference between cell elevation and the average of the elevations from its 8-cell neighborhood	Neighborhood analysis of raw DEM using annular 3x3 peighborhood		x	
twi	Topographic wetness index (TWI)	Ratio of specific catchment area to slope	TWI GP tool/WIM workflow using DEM		x	x
dtw	Depth to water	Modeled depth to groundwater based on distance from surface water	WIM GP tool, DEM and NHD			x
CLIMATE						
precip	Average annual rainfall	30-year average, average total precipitation/year	PRISM		x	
tempmin	Minimum annual temperature	30-year average of annual minimum temperature	PRISM		x	
tempmax	Maximum annual temperature	30-year average of annual maximum temperature	PRISM		x	

Short name	Variable	Description Source and calculation method		Overlay	MaxEnt	WIM
MISCELLAN	EOUS					
vegclass	Vegetation class	General class of vegetation type	GAP/LANDFIRE attribute		v	
			macro		^	
vegcomm	Vegetation	Vegetation community name	GAP/LANDFIRE attribute	×		
	community type		ecosys		X	
nwi	NWI wetland	Presence/absence of wetland feature in NWI	USFWS NWI (modified)	x		

# Modeling methods

# Overlay model methods

The primary inputs selected for use in the simple spatial overlay approach were hydric soils, topographic depressions, and NWI, as they are commonly used in combination with aerial imagery to manually identify wetlands using GIS (e.g. RTI International 2011). These same variables were also used in a similar study in Georgia that focused on remote identification of isolated wetlands (Martin et al. 2012); the approach in the Georgia study was used as a model for this project.

The source data for hydric soils was the Natural Resources Conservation Service (NRCS) statewide geodatabase of soils mapping units, which was downloaded from the NRCS website in June 2020. The original vector data (polygons) were first converted to raster format and the raster joined to several of the attribute tables provided in the original geodatabase. The values for each raster cell for 21 attributes associated with hydric soils as well as wetland presence/absence from the field delineations were exported to a text file for review. Data reviews found that a threshold of >25% for the attribute *hydclprs* from the table *muaggatt* was the strongest indicator for hydric soils statewide. The original soils raster was then reclassified so that cells with *hydclprs* > 25% were assigned a value of 1 (hydric) and all others assigned a value of 0 (non-hydric) and was used to represent hydric soils in the overlay models.

Topographical depressions in the Georgia study were manually digitized from digital versions of historic USGS topographic maps. This was not a feasible approach for our study due to the statewide scale of our analyses. Instead, hydrological sinks were used as a surrogate for identification of topographic depressions, since sinks can be quickly and automatically identified on a large area using a DEM. A combination of the Fill and Raster Math geoprocessing tools were used to identify hydrologic sinks (i.e., groups of cells with elevation lower than all surrounding cells). From this, a raster of hydrologic sink presence was made with 1 indicating a cell was part of a hydrologic sink and 0 indicating it was not in a hydrologic sink, and this raster was used to represent topographic depressions in the overlay models.

The NWI data used was a modified version prepared for a previous study examining NWI accuracy for the state. Preparation of an NWI raster was detailed in a prior report (NCDWR 2021), but essentially consisted of removing features from the original vector (polygon) data that are not commonly considered "wetlands", such as deepwater, open water, and lotic systems. The NWI raster was then reclassed, with NWI wetland features assigned a value of 1 (wetland present) and all other raster cells assigned a value of 0 (no wetland present). Seven models were prepared using all combination of the hydric soils (hyd), hydrologic sinks (snk), and NWI wetland presence (nwi) (Table 2). The inputs for each model were the binary rasters indicating presence (1) or absence (0) of each variable. For each model, the input raster(s) were summed using the Raster Math geoprocessing tool. All inputs were equally weighted. Results for each raster cell were extracted to a text file for analysis. Accuracy statistics were calculated for the entire state as well as for each of the Level III ecoregions. Additional analyses were performed for the Northern Outer Piedmont (NOP) focus area to allow comparisons to the results from the machine learning modeling.

The report from the Georgia study (Martin et al. 2012) did not indicate how the results from models with more than one input were interpreted as positively predicting a wetland, i.e., wetland predicted if at least one model variable was positive ("any") vs. wetland predicted only if all model variables were positive ("all"). Both the "any" and "all" approach were analyzed. The "any" approach resulted in slightly higher accuracy metrics and are the results described in this report.

Model	NWI	Hydric soils	Sink	Wetland indicator
nwi	Х			NWI wetland feature present
hyd		Х		Soil mapping unit >25% hydric soil components by area
snk			Х	Hydrologic sink present
nwi + hyd	Х	Х		NWI wetland <u>or</u> hydric soil presence
nwi + snk	Х		Х	NWI wetland or hydrologic sink presence
hyd + snk		Х	Х	Hydric soil <u>or</u> hydrologic sink presence
nwi + hyd + snk	Х	Х	Х	NWI wetland <u>or</u> hydric soil <u>or</u> hydrologic sink presence

#### Table 2 Descriptions of the different overlay models tested

### MaxEnt model methods

Maximum entropy (MaxEnt) is a machine learning method often used in data science and artificial intelligence that automates complex statistical model building. While initially used in fields such as thermodynamics, linguistics, image analysis, and other complex systems, it has been widely adopted by ecologists in recent years. The MaxEnt open-source software used in this project was specifically developed for ecological modeling within a spatially-aware environment.

One very significant advantage of the MaxEnt approach is that it only requires presence data for model training, whereas almost all other modeling approaches require absence data as well. Absence data can be especially problematic for modeling species distributions since true absences are often difficult or impossible to identify, so this unique feature of MaxEnt has made this approach especially popular for habitat suitability modeling.

The inputs for MaxEnt software include: 1) a text file of the coordinates of confirmed presence records (wetlands, in this case); 2) rasters (in ASCII format) of the environmental variables to be used to produce predictive models; and 3) rasters (in ASCII format) of environmental variables for the geographic area where the model will be applied (which can be the same as #2). A subset of the presence records can be held back by the software to use to test the final model, or a separate file of presence records can be provided for testing. The software iteratively reviews all combinations of linear-, quadratic-, product-, and threshold/hinge ("hockey stick")-type models for the input variables and selects the combination that provides the greatest separation of characteristics between known presence locations and background locations (i.e., locations where presence has not been confirmed). It then applies the models in the specified geographic area and assigns a value

representing the relative likelihood of presence to each raster cell. The geographic area can be the extent of the rasters representing the environmental variables used for model development, or a separate set of rasters may be provided for "projection" of the model to a new geographic area. The final output of each modeling run is a raster with cell values representing the relative likelihood of occurrence for that cell. There are several different options for scaling of this likelihood, but the cloglog output was selected for use in this project, which provides likelihood as a decimal value between 0-1. While this scale is similar to output from other modeling methods (e.g. logistic regression), it is important to note that the MaxEnt cloglog output is a relative, not absolute, likelihood. In other words, a MaxEnt cloglog value of 0.65 should not be interpreted as a "65% probability of occurrence" as it would in logistic regression. Instead, 0.65 should be used as a relative value for comparison to predictions for other raster cells in the model output.

In addition to the raster of modeled predictions, the software also provides a large amount of diagnostic information, such as area under curve (AUC), response curves for individual environmental variables, and contribution/importance of individual variables to the final model. These results can be used to guide setting up additional modeling trials with different combinations of variables to try to optimize model performance. For example, overall quality of model trials can be quickly compared based on AUC, which ranges from 0-1, with higher values indicating better model performance. Results from jackknifing and response curves for individual variables show which ones provide unique information to the model or do not seem to contribute much value to the model.

One shortcoming of MaxEnt is that the output of the software does not allow full assessment of model performance using the types of accuracy metrics traditionally used for land classification mapping applications. While the software can calculate model sensitivity (i.e., errors of omission) using a subset of the presence data withheld for testing, specificity (errors of commission) is estimated statistically. Since the extent of over- and under-prediction can be much more critical in land classification projects than when predicting species distributions, and confirmed absence records were available for this project, traditional accuracy statistics (Producer's Accuracy, User's Accuracy, and odds ratio) were calculated by comparing model results to field-delineated data. This allowed direct comparisons of accuracy to other modeling methods.

Above is a very brief overview of the MaxEnt modeling method and software, but there are many references available that provide more detail (e.g., Elith et al. 2011, Merow et al. 2013, Phillips 2017). MaxEnt has a wide range of support documentation and journal publications available to assist new users with implementing the approach and configuring the software.

For the MaxEnt modeling portion of this project, the geographic area was limited to the NOP targeted study area (Figure 3). Rasters for a total of 18 environmental variables (Table 1) were created that included various attributes of soils, topography/terrain, climate, and vegetation. All rasters were exported to ASCII and reformatted to meet the MaxEnt software requirements. The raster created from the field-delineated data (i.e., model training and testing data) was used to create a text file of coordinates for presence records.

### Wetland Identification Method (WIM) model methods

The Wetland Identification Model (WIM) was recently integrated into the ArcHydro geoprocessing toolbox addin for ESRI ArcGIS Pro. It is based on research by O'Neil et al. (2018) and uses another machine learning approach (Random Forests) to identify potential wetlands using Topographic Wetness Index (TWI), surface curvatures, and a depth-to-water index (ESRI 2020). The method was previously applied in several test locations in Virginia, with results showing significant increases in accuracy of wetland identification as compared to NWI. The WIM includes a set of Python script tools that call other existing ArcGIS geoprocessing tools (such as the Segmentation and Classification and Random Trees toolsets) as well as machine learning functions from the Python scikit-learn library.

All of the rasters representing model variables were created by analysis of the statewide DEM. Calculation of the depth-to-water index required spatial data representing surface water (streams, lakes, etc.) in addition to the DEM. The source for this was the USGS National Hydrography Dataset (NHD). The Random Trees modeling method requires both presence and absence data, which were both available in our training and testing data set (i.e., field delineations of wetlands and non-wetlands).

### <u>Software</u>

Spatial data management, review/editing, and geoprocessing were completed using ESRI ArcGIS Pro and terrain and hydrology tools included in the ArcHydro toolbox add-in. Geoprocessing tools available for use were limited due to the available license level (Standard, with Network Analyst and Spatial Analyst extensions). One modeling method (the Wetland Identification Method in ArcHydro) required the installation of an additional Python library (scikit-learn). Jupyter notebooks were used within ArcGIS Pro projects to document any Python code created for the project and to allow for consistent re-running of workflows. Raster data were generally exported to text files and imported into SAS JMP for additional data reviews and statistical analyses, including calculation of accuracy metrics. Microsoft Excel was used for preparing graphs. For MaxEnt modeling, spatial data were exported from ESRI ArcGIS Pro and analyzed using MaxEnt v3.4.3 software (Phillips et al. 2020).

### Geoprocessing

The spatial data used to prepare inputs for modeling were obtained as a mix of vector and raster types. All data were converted to raster to allow modeling and accuracy assessments at the scale of the individual raster pixel. The use of consistent geoprocessing settings throughout the project ensured proper alignment of all spatial data and reduced risks of unexpected or erroneous results. Settings were selected to match those used by the DEM used in this project since resampling and re-projecting the DEMs would have had a much more significant effect than on other data sources, with a high probability of introducing artefacts (such as sinks/pits) that would be propagated from the DEM derivatives. The projection used throughout the project was NC State Plane Feet. Raster resolution was set to 20 feet. All rasters that were created were snapped to a consistent reference raster to ensure perfect alignment of all rasters.

# Classification accuracy assessment methods

The NWI and field-delineated rasters were overlaid in ESRI ArcGIS Pro. The classifications for individual pixels from both data sets (along with ancillary characteristics from the field-delineated data) were exported to a text file for analysis. Any pixel that had a null value in one or both data sources was excluded from analysis.

The first step in the analysis process was to create a contingency table (Table 3). For each data record (i.e., raster pixel), the NWI classification was compared to the field-delineated classification and assigned a wetland classification accuracy of true positive (*A* in Table 3), false positive (*B*), false negative (*C*), or true negative (*D*). The total number of records in each category was used to build a contingency table, which is used as the basis for calculating many common accuracy measures (Congalton 1991, Fielding and Bell 1997).

*Table 3. Contingency table example* 

		Field-delineated classification			
		Wetland	Non-wetland	TOTAL	
NWI classification	Wetland	А	В	A + B	
	Non-wetland	С	D	C + D	
	TOTAL	A + C	B + D	A + B + C + D	

For this project, accuracy metrics included overall accuracy (OA), error of commission (EC), error of omission (EO), Producer's Accuracy (PA), and User's Accuracy (UA), and were calculated using the following equations:

$$\begin{aligned} Overall\ accuracy\ (OA)\ (\%) &= \frac{A+D}{A+B+C+D} \times 100 \\ Error\ of\ omission_{wetland}\ (EO_{WL})\ (\%) &= \frac{C}{A+C} \times 100 \\ Error\ of\ commission_{wetland}\ (EC_{WL})\ (\%) &= \frac{B}{A+B} \times 100 \\ Producer's\ accuracy_{wetland}\ (PA_{WL})\ (\%) &= \frac{A}{A+C} \times 100 = (1-EO_{WL}) \times 100 \\ User's\ accuracy_{wetland}\ (UA_{WL})\ (\%) &= \frac{A}{A+B} \times 100 = (1-EC_{WL}) \times 100 \end{aligned}$$

Overall accuracy (in this case, the percentage of pixels correctly classified by NWI) is one of the more commonly reported metrics in land classification accuracy assessments. It reflects the accuracy of all categories (in this case, wetland and non-wetland) but can be somewhat misleading if one class is much more prevalent than the other, as is the case with non-wetlands in NC. A better understanding of NWI accuracy can be gained by examining the additional accuracy metrics described above that focus on a single classification, such as wetlands. For example,  $EO_{WL}$  quantifies the frequency of predicting a non-wetland on the ground when a wetland does exist and  $EC_{WL}$  indicates the frequency of a wetland being predicted on the ground when one does not exist. These measures are captured by  $PA_{WL}$  and  $UA_{WL}$ , which historically have been the measures most commonly reported for land classification mapping.  $PA_{WL}$  is often described as accuracy from the mapper's perspective and describes the frequency at which a particular classification that exists on the ground is correctly identified by the map (NWI, in this case). A high  $PA_{WL}$  implies a low  $EO_{WL}$ . Conversely,  $UA_{WL}$  represents the perspective of the map user, and is the frequency that a specific classification will be found on the ground when it is indicated on the map. A high  $UA_{WL}$  implies a low  $EC_{WL}$ .

The odds-ratio statistic, like overall accuracy, provides a single number to represent the accuracy of all classifications. However, it is not sensitive to relative differences in prevalence between classes (Fielding 1997), making it useful for comparisons across data groupings or categories (such as size class or ecoregions), so the odds ratio (as opposed to overall accuracy) is used in this report. It is calculated as:

$$Odds \ ratio = \frac{A * D}{B * C}$$

The odds ratio has a lower limit of 0 and the upper limit depends on the sample size. It is undefined when *B* and/or *C* equal zero. Larger values of the odds ratio indicate higher accuracy.

# **Results and Discussion**

# Overlay models results

Overlay analyses were completed statewide for seven different models: nwi, hyd, snk, nwi + hyd, nwi + snk, hyd + snk, nwi + hyd + snk (Table 2; Figure 6). Detailed analysis of results from the nwi model were previously completed and reported in a separate report (NCDWR 2021), but significant findings included very high errors of omission for smaller wetlands (<1.0 ac.), which affected areas of the state where smaller wetlands were common, such as the Piedmont and Blue Ridge Level III ecoregions.

Almost all of the overlay models showed low odds ratios for almost all ecoregions and statewide (Figure 6). There was one outlier with an unexpectedly high odds ratio in the Blue Ridge ecoregion, but a manual review of spatial data suggested that this reflected nwi capturing the largest wetlands in this area while missing the great majority of the smaller wetlands (NCDWR 2021). The nwi model may also reflect a higher rate of correct identification of non-wetlands, since the odds ratio reflects the accuracy of all classifications. The higher rate of correct identification of non-wetlands may have contributed to the high odds ratio in the Blue Ridge ecoregion. Overall, the nwi had higher odds ratios for individual ecoregions than most of the other models, though the nwi model varied widely across ecoregions, suggesting it may have inconsistent reliability statewide. The hydrologic sinks (snk) provided the lowest overall performance based on odds-ratios, particularly in the eastern portions of the state (Southeastern Plains and Middle Atlantic Coastal Plain). The addition of other model variables to nwi did not lead to an increase in the odds ratio for any of the combined models (nwi + hyd, nwi + snk, nwi + hyd + snk).



#### Figure 6 Odds ratios for overlay models by ecoregion and statewide

The odds ratio reflects accuracy of correctly identifying both wetlands and non-wetlands. Accuracy of correctly identifying the wetland class only is better captured by the  $PA_{WL}$  (Figure 7) and  $UA_{WL}$  (Figure 8). While nwi appeared to be the best performing model overall when reviewing odds ratio results,  $PA_{WL}$  results indicate that nwi was one of the worst performers; only snk was slightly worse for any individual ecoregion. However, snk was extremely consistent across all ecoregions, even if the  $PA_{WL}$  tended to be low. For all other models, distinct differences existed in  $PA_{WL}$  between ecoregions, with much higher accuracies in the eastern portion of the state

(Southeastern Plains and Middle Atlantic Coastal ecoregions) than in the central (Piedmont ecoregion) or western (Blue Ridge ecoregion) areas. In contrast to what was seen with the odds ratio results, the combined models (nwi + hyd, nwi + snk, hyd + snk, and nwi + hyd + snk) tended to show improved accuracy in identifying the wetland class as compared to the individual models (nwi, hyd, and snk).



Figure 7 Producer's Accuracy for wetlands classification (PA<sub>WL</sub>) for overlay models by Level III ecoregion

This moderate to high  $PA_{WL}$  values for certain models and ecoregions suggest that there are often low errors of omission in the overlay models – in other words, most of the raster cells that are truly wetlands were correctly classified as wetlands in many cases. However, this is countered by low  $UA_{WL}$  results (Figure 8) for most models except for nwi, which suggests high errors of commission. In other words, while many wetlands are correctly identified by the combined models, the models also tend to overestimate the presence of wetlands, resulting in relatively large frequency of false positives. All models appeared to provide their best balance of over- and under-prediction in the Middle Atlantic Coastal Plain ecoregion, but over-prediction still seemed to be a significant issue even there. Over-prediction (i.e., low  $UA_{WL}$ ) for most models was much worse in the Blue Ridge and Piedmont ecoregions.



Figure 8 User's Accuracy for wetlands classification (UA<sub>WL</sub>) for overlay models by Level III ecoregion

Additional analyses were completed based on size class of the field-delineated features (<0.1ac., 0.1-0.25ac, 0.25-0.5ac, 0.5-1.0ac, >1.0ac). The odds ratios by size class (Figure 9) showed that all models performed extremely poorly for features smaller than 1.0 ac. However, the large increase for features >1.0 ac in size may

also be due in part to the inclusion of non-wetland features when calculating these values, since non-wetland features tended to be much larger than wetlands. The snk model was the only model that did not show this markedly different accuracy between features <1.0 ac. and >1.0 ac. in size.



Figure 9 Odds ratio for overlay models by groundtruthed feature size

The *PA<sub>WL</sub>* and *UA<sub>WL</sub>* (Figure 11) results showed that accuracy specifically for wetlands was also much lower for the smaller size classes for all models. The one exception was snk, which had uniformly moderate values for all size classes. This was the only variable that was calculated at the scale of the individual raster cells using a 20 ft. resolution DEM, rather than from converting existing vector data that were mapped at a much coarser scale (soils and NWI). This consistency in results from the snk model may also be at least partially due to the relatively impartial nature of its derivation (i.e., automated calculation from a DEM) in additional to the scale of the data, whereas the sources for nwi and hyd were the product of manual interpretation of aerial photos and similar sources, which may have been more subjective.

The combination of relatively high  $UA_{WL}$  and relatively low  $PA_{WL}$  values for all but the largest wetland size class implies that most models had large errors of omission for smaller wetlands. This same result was previously seen in an earlier NCDWR study that examined accuracy of NWI data for NC (NCDWR 2021). For large wetlands (>1.0 ac), the inverse was true – results suggested that nearly all of the overlay models had high errors of commission for large wetlands.

Because wetland size is strongly related to ecoregion, these results based on size class confirm that NWI, as well as hydric soils, are very likely to significantly underestimate wetlands in the Piedmont and Blue Ridge ecoregions mainly due to the higher prevalence of smaller wetlands in these areas, and that NWI and soils maps may be poorly suited to any projects that require identification of features <1.0 ac. in size, even when these variables are used in combination. However, the  $PA_{WL}$  and  $UA_{WL}$  did show improvement in the combined models (nwi + hyd, nwi + snk, hyd + snk, and nwi + hyd + snk) as compared to the single-variable models.



Figure 10 Producer's Accuracy for wetlands classification (PA<sub>WL</sub>) for overlay models by field-delineated wetland size class



Figure 11 User's Accuracy for wetlands classification (UA<sub>WL</sub>) for overlay models by field-delineated wetland size class

#### MaxEnt models results

A total of 27 model trials were completed for the NOP targeted study area. Results from two of these runs are discussed below: a baseline model (run 75) that used all available environmental variables, and a minimal model (run 76) that removed correlated variables and those that were not found to be valuable in the majority of modeling trials. Environmental variable rasters corresponded to the full rectangular extent of the targeted study area boundary, so there were some areas outside of the NOP boundary used by MaxEnt as background points (raster cells where wetlands had not been verified to be present) to characterize environmental conditions. As with the overlay models, only NCDOT groundtruthed points (wetland and non-wetland) from within the NOP focus area boundary were used to calculate the final accuracy metrics, though field-delineated wetland features from other agencies were used for model training.

The initial training and testing data sets were each comprised of an independent random subset of approximately 890 groundtruthed wetland points (about 10% of the total number of wetland points). However, early model runs appeared to suffer from overfitting to the environmental variables, presumably due to spatial autocorrelation of the training points. One example of overfitting was distinct banding in spatial results, which appeared to be due to overfitting of the precipitation variable (Figure 12). Many of these model runs using the initial training data set tended to have extremely good model diagnostics reported by the software (AUC and gain) and issues were only discovered when reviewing the spatial output. Much better results were obtained using a random subset of 100 groundtruthed wetland points from the NOP targeted study area as the training data, but the exact sample size and method for selecting subsamples for training and testing data are an area that would be worthy of additional research.





Figure 12 An extreme example of banding due to model overfitting in spatial model results (left) and precipitation contours for the same area (right) in the northern Piedmont of NC. Cooler colors (blues) indicate lower values and warm colors (red) indicate highest values of predicted likelihood of wetlands (top) or 30-year average annual precipitation (bottom). In top image, purple squares indicate location of groundtruthed model training data and black areas are null values due to raster cells with missing soils attributes.

The overfitting issue may also be partly due to the strong clustering of the training and testing data (Figure 13). A large proportion of the field-delineated data were associated with two large NCDOT projects in the central and southern portions of the focus area. The northern and easternmost portions of the NOP targeted study area appeared to have been poorly represented in the training/testing data. There was a total of 160 wetland features representing a total area of 63 ac. Wetland size ranged from 0.002 - 3.5 ac., with a mean of  $0.4 \pm 0.6$  ac. and median of 0.15 ac. The 70 non-wetland features totaled 4,401 ac., and ranged in size from 0.02 - 1,577 ac., mean 62.9 ± 216 ac., and median 5.2 ac.

The baseline model (run 75) resulted in a model with good diagnostics (AUC = 0.969, test gain = 2.971, averaged over 10 replicates). The relative importance of individual variables in the final complete model are summarized in Table 4. In spite of including 18 different variables, only five of them were major contributors to the final model: hydric, vegcomm, slope, elev, and tempmin. A potential issue with this "kitchen sink" approach is that the inclusion of too many variables may lead to modeling "noise" (such as the variability inherent in the data) rather than meaningful environmental conditions. Initial data reviews of the environmental variables found that a large number of variable were strongly correlated with each other. For example, PCA results showed a strong clustering of all soils variables (aws25, aws50, hydric, drain) and of curvature-related variables (curv, plan, and

tpi). Slope and elev, both of which were major contributors to the baseline model, were also correlated, though the other three variables that were major contributors were not.

Some users of machine learning modeling methods, such as MaxEnt, take the approach that the algorithms are able to compensate for covariance of model variables. This seemed to be the approach taken in the GRSM wetland modeling study (Pfennigwerth et al. 2019). The paper did not specify any tests for correlation or covariance and it appeared to include a number of environmental variables which were identical based on their descriptions (e.g., TPI and topographic ruggedness index [TRI]), would be expected to be correlated (e.g., Leaf-on Vegetation Class and Vegetation Class) or were found to be correlated in our study (e.g., soil attributes).



Review of the spatial output of the baseline/complete model (Figure 14) revealed a significant issue: there were widespread gaps in the results corresponding to missing values for certain soils attributes (aws25, aws50, and

Table 4 Variable contributions to MaxEnt			
model complete model (run75)			
Variable	Percent		
1 1 2			
nyaric	24.7		
vegcomm	18.5		
slope	14.0		
elev	10.7		
tempmin	10.5		
sinkdep	5.3		
tempmax	3.2		
drain	2.4		
vegclass	2.3		
tpi	2.1		
aws25	1.7		
precip	1.1		
plan	1.1		
prof	0.9		
twi	0.7		
aws50	0.6		
asp	0.2		
curv	0		

drain) that were primarily associated with the Water and Urban soil types. This effect was somewhat desirable for the Water classification, since it screened out the many ponds in this area, reservoirs, and large rivers, and appeared to have minimal effect on identification of the target feature type (wetlands) in undeveloped areas. However, the null values associated with Urban soil types precluded identification of wetlands in much of the southwestern portion of the focus area where some of the major cities and towns were located (Raleigh, Cary, Garner, Morrisville). Not being able to model areas where development pressures and potential wetland impacts are most likely to occur was considered to be a significant weakness of this model.

While the baseline model (run 75) used all available environmental variables, some modelers suggest that machine learning approaches should still use the traditional approach in statistical modeling of removing correlated variables prior to creating a model. This is thought to result in models that are parsimonious and that avoid fitting "noise" rather than ecologically meaningful relationships (Merow et al. 2013). This was the approach taken for model run 76 (minimal model).



6 5 10 20 Miles Figure 14 Spatial output of MaxEnt model run 75 (complete model)





Figure 16 Example of model outputs, showing comparison of NWI coverage versus model predictions by both MaxEnt models.

Variable selection for the minimal model was focused on reducing covariance between individual variables, as well addressing issues that were seen in previous runs, such as missing values for certain soil attributes and overfitting to precip. To address the issues with the soils, the hydric variable was used to represent soils, as it did not exhibit that same issue. Because precip was causing overfitting problems, and temperature data were highly correlated with precip, temperature was selected as a surrogate and climate conditions were represented by tempmin or tempmax. Model runs were iterative and variables included were modified based on the variable contribution and jackknife reports provided by the MaxEnt software and using best professional judgment. Minor adjustments to settings were also made (including modeling feature types used, particularly the threshold and hinge types). Most of the assessments of model trials were completed using the diagnostics provided by MaxEnt and visual review of spatial results in ArcGIS Pro.

Table 5 Variable contributions to MaxEnt model run 76 (minimal model)

Variable	Percent
valiable	contribution
hydric	32.7
vegcomm	20.4
tempmin	13.4
elev	11.3
sinkdep	10.7
slope	7.4
tpi	3.0
plan	1.2

The final minimal model included eight variables (Table 5), though many of the top contributors to the model overlapped those from the complete model (run 75). Mean model diagnostics from 10 replicates run with cross-validation were AUC =  $0.969 \pm 0.015$  and unregularized test gain of 3.028, which were comparable to the complete model. While the complete model seemed like it may have controlled for covarying variables, the complete model took much longer to run due to the increased number of variable combinations to be examined and the number of variables to be assessed through jackknifing.

The most notable feature in the spatial output of the minimal model (Figure 15) was the lack of areas of missing data, which was due to the exclusion of the problematic soil variables aws25, aws50, and drain.

However, the minimal model (run 76) appeared to show lower potential for wetlands in the northwestern and eastern portions of the NOP focus area than the complete model (run 75). These are the areas that had a distinct lack of training and testing data, so it cannot be determined whether the complete or the minimal model provided a more accurate picture for this area. This lack of field-delineations in these areas leads to a question of whether the conditions in these areas were sufficiently represented in the training data set. The minimal model also had distinct "halos" of moderate cloglog values in the southern portion of the NOP that were not present (or at least not as pronounced) in the complete model. Figure 16 shows an example of the outputs of both models overlaid with NWI, showing a greater spatial wetland prediction by the MaxEnt complete model than the minimal model. Both predicted greater wetland area(s) than NWI in the NOP.

One similarity between all variables included in the minimal model and the best performing variables in the complete model was that they were all from data sources that are updated, albeit irregularly in some cases. The vegcomm (from the GAP/LANDFIRE land cover dataset) is updated every five years. The tempmin is based on a 30-year average and is updated every 10 years. All other variables were derived from the legacy DEMs collected statewide in the early 2000's, but more recent and higher resolution DEMs are available from LiDAR collected within the last 10 years. Chances are good that LiDAR data collections will be repeated in the future, given the importance of having current and high-accuracy elevation data to the state's emergency management agencies (for flood mapping, for example). Methods such as those used in this study, that incorporate data collected repeatedly over time, provide an opportunity for regular updates of wetland maps without the extremely intensive requirements for time, training, and expertise required in manual review and digitization used currently in data sources such as NWI.

The outputs of the MaxEnt models were rasters showing a relative likelihood (as the "cloglog") of wetland presence for each raster cell. To calculate accuracy metrics for the MaxEnt models, these rasters of probabilities were converted to a presence/absence format, which required selection of a threshold. For each model, the cloglog values for each raster cell with a wetland or non-wetland classification assigned from field-delineations were exported for analysis. The 10<sup>th</sup>, 15<sup>th</sup>, 20<sup>th</sup>, 25<sup>th</sup>, and 30<sup>th</sup> percentiles of the cloglog for the wetland cells were calculated for each model. Each of the percentile values was then used to create a separate field indicating predicted wetland presence or absence based on the individual threshold and accuracy metrics calculated. As an example, the *PA*<sub>WL</sub> and  $UA_{WL}$  for each of the thresholds for the complete and minimal models are shown in Figure 17. This

shows the typical relationship in classification assessment – increasing  $PA_{WL}$  can come at the cost of lowering  $UA_{WL}$ . In this case, using the 10<sup>th</sup> percentile for either model would result in highest  $PA_{WL}$  but also the highest over-prediction of wetlands (i.e., lowest  $UA_{WL}$ ). Conversely, while the NWI  $UA_{WL}$  was higher for the targeted study area than any of the MaxEnt models, the  $PA_{WL}$  was quite low, suggesting that NWI both over- and under-predicts significantly.

All of the MaxEnt model thresholds resulted in very large increases in  $PA_{WL}$  relative to NWI, suggesting that the MaxEnt models were capturing many more actual wetlands in the landscape. While there was a decrease in  $UA_{WL}$  (therefore over-prediction) as  $PA_{WL}$  increased, it was much more modest than the corresponding drop in  $PA_{WL}$ .



Figure 17 Producer's Accuracy and User's Accuracy for nwi and for various thresholds (percentiles) for MaxEnt complete model (run 75) and minimal model (run 76)

For the overall accuracy metrics, odds ratios (Figure 18) were fairly consistent across all thresholds for each MaxEnt model. Odds ratios for all thresholds for run 75 (complete model) were very similar to the odds ratio for NWI. Odds ratios for all thresholds in run 76 (minimal model) were greater than the NWI result.



Figure 18 Odds ratio for nwi and for various thresholds (percentiles) for MaxEnt run 75 (complete model) and run 76 (minimal model)

The 30<sup>th</sup> percentile values for both the complete and minimal model were selected to create the presence/absence rasters since they provided a good balance of correct identification of wetlands, overprediction, and under-prediction. For the complete model (run 75), this threshold was 0.3787. For the minimal model (run 76), the threshold was 0.3285. The thresholds were used to reclassify the rasters of cloglog outputs and create new rasters showing predictions of wetland presence or absence.

The final presence/absence model outputs were combined with the field-delineated data in ESRI ArcGIS Pro and values exported to a text file for additional analyses based on wetland size class (< 0.5 ac. or  $\geq$  0.5 ac.) (Figure 19). Results showed that MaxEnt models as well as NWI showed inverse trends depending on the wetland size class, with under-prediction more prevalent in smaller features and overprediction more prevalent in larger features. Both MaxEnt models, however, outperformed NWI in identifying smaller wetland features (<0.5 ac) based on both  $PA_{WL}$  and  $UA_{WL}$ , and the differences between  $PA_{WL}$  and  $UA_{WL}$  for each MaxEnt model were much smaller than seen for NWI. For wetlands >0.5ac., NWI had a slightly higher  $UA_{WL}$  than the MaxEnt models and a smaller difference between  $PA_{WL}$ and  $UA_{WL}$ , but the MaxEnt models had higher  $PA_{WL}$ .



Figure 19 Producer's Accuracy and User's Accuracy of nwi, Maxent run 75 (complete model), and MaxEnt run 76 (minimal model) by wetland feature size class in NOP focus area

In summary, the MaxEnt modeling process using these particular variables did not show significant differences in model diagnostics or calculated accuracy measures whether using the entire set of variables or iteratively selecting only the most high-performing and valuable variables. While model diagnostics provided by the MaxEnt software were useful, high AUC and gain could be the result of overfitting or other issues, so visual review of the spatial output and calculation of classic accuracy measures were required for final assessment of the individual model runs. Several model variables were found to have significant impacts on results, so thorough reviews of environmental variables should be performed prior to and during the modeling process. Selection of a threshold to use for determining presence/absence in the final spatial data set should be supported by some specified criteria or rationale. In this study, I attempted to balance the risk of over-prediction and under-prediction based on *PA<sub>WL</sub>* and *UA<sub>WL</sub>*, but other applications may choose to more heavily weight either the error of commission or of omission.

When MaxEnt models were compared to the overlay models for all data (Table 6), the MaxEnt models had very similar results as hyd and nwi + hyd. MaxEnt  $PA_{WL}$  accuracy was an improvement over NWI but the MaxEnt models also had a small decrease in  $UA_{WL}$ . The MaxEnt approach, however, had the significant benefit of increasing accuracy for smaller wetland features (<0.5 ac).

Table 6 Accuracy statistics for the NOP targeted study area for MaxEnt and overlay models					
Model	PA	UA	Odds ratio		
MaxEnt complete model (NOP run 75 )	75%	17%	28.9		
MaxEnt minimal model (NOP run 76)	75%	18%	36.4		
nwi	47%	25%	29.6		
hyd	73%	14%	26.9		
snk	59%	9%	9.8		
nwi+hyd	79%	14%	33.6		
nwi+snk	73%	9%	16.1		
hyd+snk	84%	8%	22.6		
nwi+hyd+snk	86%	8%	26.1		

# ESRI Wetland Identification Model (WIM) results

We attempted to apply the WIM approach that has recently been added to the ESRI ArcGIS Pro ArcHydro toolbox to the same NOP focus area used for MaxEnt modeling. However, numerous errors were encountered using both the Wetland ID Workflow data model in the WIM toolbox and the individual script tools. We were unsuccessful in even creating the inputs required for the Random Forests modeling.

As part of the troubleshooting, the individual Python scripts that made up the individual script tools were reviewed to gain an insight into the geoprocessing workflow. Many of the processes used in WIM incorporate other standardized geoprocessing tools, but the recommended method for the initial smoothing of the DEM (Perona Malik) depended on functions in the SciKit Learn Python library. These functions apparently had limitations on the maximum size of the DEM it was able to process. A maximum size limit was not mentioned in the documentation at the time (ESRI 2020), but the method developer indicated that the size limit was significantly smaller than a USGS 12-digit hydrologic unit (HU) (Gina O'Neil, pers. comm., Feb. 3, 2021). The size of a HU-12 averages approximately 33 mi<sup>2</sup> in NC, or roughly equal to one-fifth of the area of the city of Raleigh. This size limitation may also mean that it is

only applicable in headwater systems: one of the required inputs (TWI) required calculation of the drainage area for each raster cell, so the tools would only be applicable to project areas whose entire watershed is below the size threshold required by the SciKit Learn functions.

Because of the previously unidentified size limitation and the technical issues with getting the geospatial tools to run successfully, this approach was abandoned and could not be assessed as part of this project. Based on existing research, the WIM seems to be a promising approach for remote identification of wetlands, and has the added benefit of being a "canned" toolset provided with ArcGIS Pro's ArcHydro toolbox add-in. The input data required (DEM and stream/surface water location, such as the National Hydrography Dataset) are also minimal and readily available. However, this particular toolset does not seem to be appropriate for the use of predicting wetlands over moderately sized or large areas.

The WIM is a fairly new method and will likely be further refined in coming years. Based on the currently available research, it appears to be a promising approach and could be immediately applicable to small project areas (e.g., private development, transportation projects). Much of the method could potentially be applied manually (either directly with existing geoprocessing tools or approximated using other workflows) to larger areas, using an alternative DEM smoothing method and the documentation and Python scripts as a guide.

# Conclusion

NWI has been shown to be a problematic data set for identifying wetlands in NC, though it is the only source of wetland maps available for the state and for much of the US. Prior work done by NCDWR has demonstrated that NWI has inconsistent accuracy across the state. NWI does an extremely poor job of identifying wetlands <1.0 ac. in size, and consequently NWI accuracy is very low in areas of the state where wetlands tend to small, such as the Piedmont and Blue Ridge ecoregions (NCDWR 2021). These findings were expected, based on the age of the NWI data and the mapping standards used to create it (Tiner 1997). The US Fish and Wildlife Service (USFWS), the creators and current stewards of the NWI, will no longer update the data set, and are relying on other agencies and entities to take on that role. Some other states have undertaken that effort but meeting the requirements of the wetland mapping standards for NWI updates (FGDC 2009) means that updates to NWI will continue to be problematic in many ways. For example, the mapping methods require that only wetlands >0.5 ac. in size be included on the map, and only those wetlands above that minimum mapping size be included in accuracy assessments. The mapping methods are based on those that were in use when NWI was first developed 40-50 years ago and require manual digitization and attribution of wetland features, which is a timeintensive – and therefore expensive – approach, which means that future updates will be infrequent. The NWI mapping standards also require identification of features not commonly considered "wetlands", such as deepwater, open water, and stream/river systems, and assignment of a wetland classification following a classification system (Cowardin 1979) that is complicated and rarely used outside of NWI, at least in NC. These additional requirements of including non-wetland habitats and assigning Cowardin classifications will increase time and expertise needed for updating NWI, and for information that is not usually of use to most wetland scientists, regulators, and the regulated community.

Exploration of more modern approaches to predict wetland locations and extent are common in the scientific literature, including several studies in NC or adjoining states. These take advantage of automated classification techniques and GIS and show potential for creating wetland maps outside of the NWI framework that will be more relevant to the majority of current users of NWI. They take advantage of advances in GIS software, spatial data collections, and computing power over the last several decades and can be quicker and cheaper to complete than manual wetland mapping techniques. GIS- based automated classification methods reduce potential sources of bias due to human interpretation of remote sensing data, and are also repeatable, which would allow more frequent updates to the data and therefore allow better tracking of changes to wetland resources over time.

This study focused on a cursory examination of two different approaches to creating updated wetland maps: simple spatial overlays and more complex machine-learning methods. The overlay approach used here was based on the one used in a study in the Level III ecoregion Southeastern Plains in Georgia (Martin et al. 2012). The Georgia study reported slightly different accuracy measures, including sensitivity and specificity. Sensitivity is equivalent to the  $PA_{WL}$  used in this study and specificity refers to the accuracy of correctly identifying non-wetlands (which was not considered in our study). The Georgia study found the NWI to have a specificity of 68%, which is well above the  $PA_{WL}$  of 49% that was reported for the Southeastern Plains in NC in a prior NCDWR report (NCDWR 2021). The Georgia study also found that a "complete" model using NWI, hydric soils, and topographic depressions performed best, resulting in a sensitivity of 88%. Our results from a similar overlay resulted in an increase in  $PA_{WL}$  to 81%, though the overlay model also resulted in decreased  $UA_{WL}$  from 60% to 25% for this ecoregion, suggesting that

although more wetlands are correctly predicted, the nwi + hyd + snk model also resulted in overprediction of wetlands.

Methods from a study in Virginia using Random Forests to identify wetlands (O'Neil 2018) serve as the basis for the Wetland Identification Method (WIM) tools recently incorporated into ESRI ArcGIS Pro (ESRI 2020). The Virginia study included one study area in the Middle Atlantic Coastal Plain Level III ecoregion, which also occurs in NC. The  $PA_{WL}$  values for this study showed an increase to 81% for the model as compared to the NWI baseline of 32%. However, there was also an accompanying decrease in  $UA_{WL}$  from 90% for NWI to 61% for the Random Forests model. Our baseline accuracy for NWI in this ecoregion was somewhat different than in Virginia, with a  $PA_{WL}$  of 53% and  $UA_{WL}$  of 68%. While we were unable to apply the WIM method for this project, the best performing overlay model for this ecoregion in our study was nwi + snk, which resulted in the best combination of increased  $PA_{WL}$  (70%) and minimally reduced  $UA_{WL}$  (49%), not quite as good as the improvements found in Virginia.

The NCDOT has had an ongoing research project for approximately ten years that is focused on developing spatial models for remote identification of wetlands that can be applied on a project-by-project basis. The most recent models were derived using a Random Forests approach in several Level IV ecoregions in eastern NC (NCDOT 2017, Wang 2015). NCDOT did not report the accuracy of NWI as a baseline for comparison, but NWI accuracy was reported for these Level IV ecoregions in a previous NCDWR report (NCDWR 2021) so those are presented for comparison (Table 7). Based on these findings, the Random Forests approach seemed to show improvements over NWI for *PA<sub>WL</sub>* in the Rolling Coastal Plain Level IV ecoregion, results were fairly similar to NWI for the Southeastern Floodplains and Low Terraces, and accuracy dropped slightly as compared to NWI in the Carolina Flatwoods.

	N	WI	NCDOT model		
Level III ecoregion Level IV ecoregion	PA <sub>WL</sub>	UA <sub>WL</sub>	PA <sub>WL</sub>	UA <sub>WL</sub>	
Southeastern Plains		60%	N/A	N/A	
Rolling Coastal Plain	42%	53%	63%	27%	
Southeastern Floodplains & Low Terraces	65%	69%	65%	52%	
Middle Atlantic Coastal Plain	53%	68%	N/A	N/A	
Carolina Flatwoods	59%	65%	50%	44%	

Table 7 Comparison of Producer's Accuracy and User's Accuracy for NWI (baseline) and NCDOT wetland model

While results from this and other studies showed somewhat mixed results, there are indications that models developed to support automated remote identification of wetlands may be a more efficient and cost-effective method for developing more useful and current wetland maps for the state. There are a number of technical issues that could not be addressed in this study, such as spatially correlated model variables, modification of modeling parameters, and limitations with currently available geoprocessing tools. However, these preliminary results suggest that use of more "unbiased" data sources for modeling (such as DEMs) as compared to maps derived from outdated methods like manual photointerpretation may address some key weaknesses previously identified in the NWI, particularly the poor accuracy for wetlands <1.0 ac. in size. Furthermore, automated methods would be realistically repeatable on a regular basis, ensuring that available wetland maps more closely represent current conditions.

# References

- Congalton, RG. 1991. "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data." *Remote Sensing of the Environment* (37): 35-46.
- Cowardin, L.M., V. Carter, F. Golet, and E. LaRoe. 1979. *Classification of wetlands and deepwater habitats of the United States*. Washington, DC: US Dept of the Interior, Fish and Wildlife Service. https://www.fws.gov/wetlands/documents/classification-of-wetlands-and-deepwater-habitatsof-the-united-states.pdf.
- Davis, L.R., K. Fishburn, H. Lestinsky, L. Moore, and J. Walter. 2019. US Topo Product Standard (ver. 2.0, February 2019). Vols. Book 11, Collection and Delineation of Spatial Data, chap. B2 in U.S. Geological Survey Techniques and Methods. Reston, VA: US Geological Survey. doi:https://doi.org/10.3133/tm11b2.
- Ducks Unlimited. 2018. "Technical Procedures for Updating the National Wetlands Inventory for Minnesota Central Project Region." Ann Arbor, MI. https://www.dnr.state.mn.us/eco/wetlands/nwi\_proj.html.
- Ducks Unlimited. 2016. "Updating the National Wetland Inventory in Northeast Minnesota Technical Documentation." Great Lakes/Atlantic Regional Office, Ann Arbor, MI. https://www.dnr.state.mn.us/eco/wetlands/nwi\_proj.html.
- Ducks Unlimited. 2008. "Updating the National Wetlands Inventory (NWI) for the Great Lakes (Wisconsin, Illinois, Indiana, Michigan, and Ohio) Project Workplan." Great Lakes/Atlantic Regional Office, Ann Arbor, MI. https://www.ducks.org/media/Conservation/GLARO/\_documents/\_library/\_gis/NWI\_Workplan. pdf.
- Ducks Unlimited. 2013. "Updating the National Wetlands Inventory in East-Central Minnesota Technical Documentation." Great Lakes/Atlantic Regional Office, Ann Arbor, MI. https://www.dnr.state.mn.us/eco/wetlands/nwi\_proj.html.
- Elith, J, S.J. Phillips, T. Hastie, M. Dudik, Y.E. Chee, C.J. Yates. 2011. "A statistical explanation of MaxEnt for ecologists." *Diversity and Distributions* 17: 43-57.
- ESRI. 2020. "Arc Hydro: Wetland Identification Model." Redland, CA. https://community.esri.com/t5/water-resources-blog/the-wetland-identification-model-wim-anew-arc-hydro/ba-p/884298.
- Federal Geographic Data Committee. 2009. "Wetlands Mapping Standard." FGDC Wetlands Subcommittee. https://www.fgdc.gov/standards/projects/wetlands-mapping/2009-08%20FGDC%20Wetlands%20Mapping%20Standard\_final.pdf.
- Fielding, A.H., and J.F. Bell. 1997. "A review of methods for assessment of prediction errors in conservation presence/absence models." *Environmental Conservation* 24 (1): 38-49.
- Gallant, AL. 2015. "The Challenges of Remote Monitoring of Wetlands." *Remote Sensing* 2015 (7): 10938-10950. doi:10.3390/rs70810938.

- Griffith, G.E., J.M. Omernik, J.A. Comstock, M. Schafale, W.H. McNab, D. Lenat, T.F MacPherson, T.F. MacPherson, J.B. Glover, and V.B. Shelburne. 2002. *Ecoregions of North Carolina and South Carolina.* Reston, VA: U.S. Geological Survey. https://www.epa.gov/eco-research/ecoregiondownload-files-state-region-4#pane-31.
- Gritzner, J. 2006. "Identifying wetland depressions in bare-ground LiDAR for hydrological modeling." *ESRI User Conference Proceedings.* http://proceedings.esri.com/library/userconf/proc06/papers/papers/pap 1225.pdf.
- Kloiber, S., and R. Macleod. 2019. "Lessons Learned from the Minnesota NWI Updates." *Wetland Mapping Consortium Webinar Series.* https://www.aswm.org/wetland-science/wetlands-onestop-mapping/5334-2019-past-wetland-mapping-consortium-webinars#mapping0904.
- Martin, G.I., L.K. Kirkman, and J. Hepinstall-Cymerman. 2012. "Mapping Geographically Isolated Wetlands in the Dougherty Plain, Georgia, USA." *Wetlands* 32: 149-160.
- McCauley, L.A., and D.G. Jenkins. 2005. "GIS-based estimates of former and current depressional wetlands in an agricultural landscape." *Ecological Applications* 15 (4): 1100-1208. https://www.researchgate.net/publication/228995334\_GISbased\_estimates\_of\_former\_and\_current\_depressional\_wetlands\_in\_an\_agricultural\_landscap e.
- Merow, C., M.J. Smith, and J.A. Silander. 2013. "A practical guide to MaxEnt for modeling species' distributions: what is does, and why inputs and settings matter." *Ecography* 36 (10): 1058-1069. doi:https://doi.org/10.1111/j.1600-0587.2013.07872.x.
- NCDOT. 2017. "Wetland Predictive Model Assessment for Kinston Bypass R-2553 LEDPA (memorandum)." June 15.
- NCDWR. 2021. *National Wetlands Inventory (NWI) Accuracy in North Carolina*. NC Dept. of Environmental Quality, Raleigh, NC: NC Division of Water Resources.
- O'Neil, G.L., J.L. Goodall, and W.T. Watson. 2018. "Evaluating the Potential for site-specific modification of LiDAR DEM derivatives to improve environmental planning-scale wetland identification using Random Forest classification." *Journal of Hydrology* 559 (April): 192-208. doi:https://doi.org/10.1016/j.jhydrol.2018.02.009.
- Pfennigwerth, A.A., J. Albritton, and T. Evans. 2019. "Using Spatial Modeling to Improve Wetland Inventories in Great Smokies National Park." *Natural Resources Journal* 39 (4): 482-488. https://boglearningnetwork.files.wordpress.com/2019/12/2019-pfenningwerth-et-al-usingspatial-modeling-to-improve-wetland-inventories-in-great-smoky-mountains-national-park.pdf.
- Phillips, S.J., M. Dudik, and R.E. Schapire. 2020. "MaxEnt software for modeling species niches and distributions (version 3.4.3)." Nov. Accessed Dec 2020. https://biodiversityinformatics.amnh.org/open\_source/maxent/.
- Phillips, SJ. 2017. A Brief Tutorial on Maxent. Biodiversity Infomatics. https://biodiversityinformatics.amnh.org/open\_source/maxent/Maxent\_tutorial2017.pdf.

- RTI International. 2011. "Assessing Geographically Isolated Wetlands in North and South Carolina the Southeast Isolated Wetlands Assessment (SEIWA) Final Report." Research Triangle Park, NC. http://www.ncwetlands.org/project/mapping-assessing-isolated-wetlands/.
- Saint Mary's University of Minnesota. 2018. "Technical Procedures for Updating the National Wetlands Inventory of Northwest Minnesota." GeoSpatial Services. https://www.dnr.state.mn.us/eco/wetlands/nwi\_proj.html.
- Saint Mary's University of Minnesota. 2015. "Technical Procedures for Updating the National Wetlands Inventory of Southern Minnesota." GeoSpatial Services. https://www.dnr.state.mn.us/eco/wetlands/nwi\_proj.html.
- Sutter, L. 1999. *DCM Wetland Mapping in Coastal North Carolina*. NC Dept of Environment and Natural Resources, Div. of Coastal Management. https://files.nc.gov/ncdeq/Coastal%20Management/documents/PDF/wetlands/WTYPEMAPDOC .pdf.
- Tiner, RW. 1997. "NWI Maps: What They Tell Us." *National Wetlands Newsletter* (Environmental Law Institute) 19 (2). https://www.fws.gov/wetlands/Documents/NWI-Maps-What-They-Tell-Us.pdf.
- US Environmental Protection Agency (US EPA). 2016. *National Wetland Condition Assessment 2011 Technical Report, EPA-843-R-15-006.* Washington, DC: Office of Water, Office of Research and Development. https://www.epa.gov/national-aquatic-resource-surveys/nwca.
- US Fish and Wildlife Service. 2021. *NWI Projects Mapper*. Accessed March 8, 2021. https://www.fws.gov/wetlands/data/mapper.html.
- Wang, S. 2015. Improvements to NCDOT's Wetland Prediction Model. UNCC WAM Research Team.
- Yang, L., J. Suming, P. Danielson, C. Homer, L. Gass, S. Bender, A. Case, et al. 2018. "A new generation of the United States National Land Cover Database." *ISPRS Journal of Photogrammetry and Remote Sensing* (146): 108-123. doi:https://doi.org/10.1016/j.isprsjprs.2018.09.006.