Wetland Mapping Tool Project
Phase 2 Report

Prepared by:

Dan Miller and Meghan Halabisky
TerrainWorks Remote Sensing and Geospatial Analysis Lab
University of Washington

dan@terrainworks.com halabisk@uw.edu

Prepared for:

Wetlands Scientific Advisory Group (WetSAG)
Cooperative Monitoring, Evaluation, and Research Committee

February 15, 2020

Introduction

The 2019-2021 Cooperative Monitoring, Evaluation, and Research (CMER) work plan[1] defines the Wetland Mapping Program strategy:

> "This program is intended to address gaps in existing data on the location, distribution, size, and geophysical characteristics of wetlands, especially for forested wetlands. More accurate spatial data could enhance the design and implementation of projects examining the effects of forest practices rules on wetland functions. Specifically, the data could provide context for (1) focusing research on wetlands and associated typed-waters that may be vulnerable to harvest and road impacts, and (2) assessing the spatial applicability (inference) of study findings to other landscapes. The use of remote sensing and associated geospatial modeling with GIS is a potentially viable tool to fill these data needs; however, no suitable GIS model is currently available for grouping wetlands by functional type or landscape position."

The Wetland Mapping Tool project seeks to provide an ArcGIS-based tool for using remotely sensed and other GIS data resources to map the estimated probability of wetland occurrence. The project is divided into two phases.

Background

Phase 1, completed in April 2018[2], involved development of a wetland intrinsic potential (WIP) tool. The WIP tool uses topographic, climatic, and geologic data to characterize five hydrogeomorphic indicators (Brinson, 1993): 1) depth to water table near streams and rivers, 2) depth to water table near lakes and ponds, 3) relative depth of closed depressions, 4) depth to an impermeable layer, and 5) a climate-topographic wetness index. Each indicator represents certain physical controls on processes of water flux through the landscape that create conditions for wetland formation.

Phase 1 focused on developing individual suitability curves for each of the hydrogeomorphic indicators using the range of values identified on the landscape. Each suitability curve is scaled from zero to one, with zero indicating no potential for wetland formation and one indicating a high potential. For each location on the ground, the suitability-curve values are combined to provide a postulated probability of wetland presence. The tool thus enables the translation of postulated controls on water flux to quantitative predictions of wetland occurrence that can be compared to field observations. The WIP tool developed during phase 1 provides a means of translating hypotheses about wetland formation into maps showing predictions of likely wetland occurrence. The WIP tool from phase 1 can be used without training data, but proved

[1] https://www.dnr.wa.gov/publications/fp_cmer_2019_2021_workplan_20190119.pdf?c4wk29
[2] Final report available at https://terrainworks.sharefile.com/d-s961f11c3eca47e58

to be difficult to put into practice and even with tuning still missed several forested wetland areas.

While Phase 2 builds off the hydrogeomorphic principles outlined in Phase 1, the methods and selected data inputs to the model are different. Phase 2, described in this report, uses machine learning trained on locations of wetland presence and absence to create a wetland probability map. For Phase 2, we built off insights gained from research led by the Washington State Department of Ecology (WA DOE) and developed under the EPA Grant Number CD01J09401, Improved Wetland Identification for Conservation and Regulatory Priorities (Halabisky, 2019), which was completed after Phase 1 results. The WA DOE-led project tested multiple data input layers, including those used in Phase 1, to map wetlands across an entire watershed. The WA DOE-led project provided quantitative information in regards to which datasets had highest model significance at predicting wetlands overall. However, the WA DOE-led project could not quantitatively assess model effectiveness specifically for forested wetlands because of a lack of reference data, which was the objective of the Phase 2 WIP tool. We visually explored all of the data inputs by overlaying them with known forested wetlands to qualitatively determine their effectiveness at detecting forested wetland areas.

In Phase 2 we selected only the variables that had model significance or qualitatively showed promise at picking up forested wetlands in the WA DOE-led effort. Multiple topographic indices calculated over a range of length scales were identified as strongly predicting wetland presence or absence, especially for forested wetlands, and are described below in the Phase 2 methods. The only variable from Phase 1 that was included in Phase 2 was the topographic wetness index. One additional variable from Phase 1, depth to water table near streams and rivers, was shown to have some significance in model improvement for the WA Dept. of Ecology led project, but was not included in the model testing and results. This depth-to-water metric was of lower importance and was not prioritized for inclusion because it required significant time investment to build into the WIP tool software. The depth-to-water metric could be added in the future or calculated separately and brought into the random forest model.

In addition to changing some of the input datasets for Phase 2, we explored the use of statistical models using machine-learning approaches (Halabisky, 2019). Random forest modelling, a machine-learning technique, provided the best performance and greatest flexibility for application across a diverse range of landscapes. The creation of input datasets and the random forest approach have been incorporated into the Phase 2 WIP tool ArcGIS Pro toolsets, and allow users to sequentially:

1. Calculate topographic attributes at multiple scales,
2. Calibrate a random forest model using point locations classified in terms of wetland presence or absence,
3. Develop a model of probability of wetland occurrence that can be applied and tested in other areas.

The ArcGIS toolboxes allow use of a wider variety of data sources than were available with the Phase-1 WIP tool for predicting wetland presence and provide means for evaluating the resulting models.

Random forest

Wetlands can be identified using indicators of vegetation, hydrology, topography, and soils. Remotely sensed multispectral imagery is used to identify wetland vegetation (e.g., Adam et al., 2009; Mahdavi et al., 2017) and presence of surface water. Lidar-derived elevation data is used to identify topographic features where wetlands form (e.g., Fink and Drohan, 2016; Lang et al., 2013; O'Neil et al., 2018). Lidar intensity can delineate surface water and wet soils (Lang and McCarty, 2009). Inclusion of soil characteristics can improve GIS-based wetland mapping (Buchanan et al., 2014; Knight et al., 2013). Even the abundance of down wood might influence the flow of surface water into wetlands or riparian areas (Janisch et al., 2011). Given the abundance of potential predictors, we require techniques capable of using and evaluating many variables for predicting wetland location.

Analysis of large and diverse datasets has benefited from the relatively recent development of machine-learning algorithms (e.g., Hastie et al., 2017; Maxwell et al., 2018). Specifically, there has been a shift to the use of powerful new machine learning algorithms that do not require assumptions about the statistical distribution of input data. Non-parametric supervised classification approaches to land cover mapping produce more efficient and accurate results than earlier supervised parametric classification methods (e.g. maximum likelihood) primarily because satellite image data values are not normally distributed (Wulder et al. 2019). Random forest modelling is the most commonly used non-parametric classification method (Breiman, 2001), which allows for the use of multiple, correlated input variables that are not normally distributed.

Random forest is an extension of the Classification and Regression Tree (CART, Breiman et al., 1984) approach for identifying correlations among many and potentially diverse attributes. For analysis of classified data (e.g., is a location a wetland or not), a decision tree is used in which the explanatory variables are divided into separate domains, so that each domain contains dependent variables (the training data) primarily of one class. The predicted class for a new location depends on which domain it falls into. CART is very effective at identifying subtle relationships, but suffers from high variance; that is, the result can be extremely sensitive to differences in the input data. The random forest algorithm addresses this weakness by using many analyses – that is, growing a large number of decision trees – for a single data set, with each analysis (tree) using only a randomly selected portion of the data. Each analysis predicts a class for each data point sampled from the training data set. Depending on the degree to which the training data classes can be isolated into separate domains within the range of explanatory variables (the degree to which the explanatory variables can discern controls on wetland formation), that prediction may be correct or incorrect. The result for each point is based on the majority result over all the analyses – over all the trees. Random forest has proven to be a

robust and effective method for discerning relationships between the types and variety of variables listed above and observed wetland locations (e.g., Chignell et al., 2018; Halabisky, 2019; Mahdavi et al., 2017; Maxwell et al., 2016; Tyralis et al., 2019).

Topographic indices

Wetlands form in locations where soil water accumulates and persists, so physical controls on soil water fluxes influence wetland formation. Topography imposes a primary control on soil-water fluxes and many studies have shown that topographic attributes derived from digital elevation data, particularly high-resolution data from lidar, are effective at identifying wetland locations (e.g., Lang et al., 2013; O'Neil et al., 2018). High-resolution lidar DEMs can thus provide data for identifying wetlands in locations where multispectral imagery may be ineffective or unavailable. Additional data types can be combined with topographic data to increase the accuracy of wetland predictions (Kloiber et al., 2015).

However, in order to identify wetlands using topographic metrics we first need to identify those topographic attributes most relevant to wetland formation and then translate the grid of elevation values provided by a lidar DEM to quantitative measures of those attributes. There are many ways to characterize topography. Wilson (2018) lists 17 different measures of curvature; Jasiewicz and Stepinski (2013) identify 498 distinct topographic forms that can be uniquely identified from a DEM. We must also identify the appropriate length scales at which to measure topographic attributes. Wetlands are often found in low-lying terrain, but is a 30-meter-wide depression as important as a 300-meter-wide depression, or a 3000-meter-wide depression? Likewise, does a depression on a valley floor have the same importance as a depression on a ridge top?

Studies to date have examined relatively few topographic attributes (although Maxwell et al., 2016, examined 21 terrain attributes in their analysis). Most research has focused primarily on topographic indices of soil wetness (e.g., Lang et al., 2013) and mapping of closed depressions (e.g., Wu and Lane, 2016). There have been several researchers who have modelled groundwater using an approximate measure of height above nearby channels referred to as depth-to-water (e.g., Murphy et al., 2007; White et al., 2012).

We tested the effectiveness of topographic measures of gradient, curvature, and local relief calculated at different length scales. Topographic metrics are calculated directly from a DEM, whereas calculation of the topographic wetness index and depth-to-water first require that the channel network be delineated, which entails multiple steps (e.g., Miller et al., 2015). For example, to calculate the topographic wetness index (TWI) one must use a DEM to calculate slope, flow direction, and flow accumulation using hydrological flow modelling tools, which are then built into an equation that calculates the TWI.

Gradient (or slope) indicates the change in elevation with distance: rise over run, or $S = dz/dx$, where S is the gradient and dz indicates the change in elevation over a horizontal distance dx. Gradient at a point can be calculated for any direction. We use the direction in which gradient is

largest; that is, along the fall line aligned with the hillslope aspect. Curvature gives the change in gradient with distance dS/dx (Figure 1).

Curvature parallel to the fall line, in the direction of steepest gradient, is referred to as profile curvature (Figure 1). Where gradient is decreasing as one moves downslope, (i.e., where profile curvature is positive), the velocity of shallow groundwater flow downslope decreases, forcing ground water toward the surface. Thus, large positive values of profile curvature may indicate likely zones of wet soils. Curvature in a direction perpendicular to the fall line, along a contour, is referred to as plan curvature (Figure 1). Plan curvature provides a measure of topographic convergence (positive values) and divergence (negative). Convergent topography acts to concentrate surface and subsurface water flow, so large positive values of plan curvature may also indicate likely zones of wet soils[3]. A DEM provides elevation values over a regular grid of points. Elevations between points must be interpolated. A variety of interpolation schemes have been devised, and the choice of algorithm used influences the resulting values (Florinsky, 1998). We represent the topography around a DEM grid point using a polynomial surface fit to the point and to 8 adjacent points (Zevenbergen and Thorne, 1987). The 8 adjacent points are placed on a circle of specified diameter about the central point (Shi et al., 2007). If an adjacent point on the circle does not fall exactly on a DEM grid point, we use bilinear interpolation to the nearest four grid points to estimate its elevation. This procedure allows us to calculate gradient and curvature for each DEM point measured over any length scale (down to the DEM grid size).

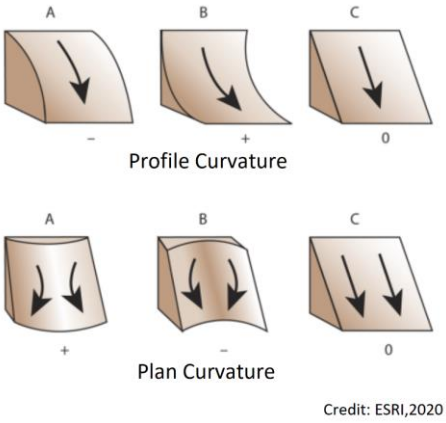

Figure 1. Examples of profile and plan curvature

We use local relief to indicate whether a point is in low- or high-lying terrain. As a measure of local relief, we use deviation from mean elevation (Wilson, 2018): $DEV = (z - \bar{z}_{dx})/sd_{dx}$, where z is elevation at the point of measurement, $\bar{z}_{dx}$ is mean elevation over a circle of diameter dx, and sd_{dx} is the standard deviation of elevation within that circle. Positive values of DEV indicate the point is higher than the mean of neighboring points (within the circle of diameter dx); negative values indicate the point is lower. Dividing by the standard deviation – a measure of how variable elevations are within the circle – acts to normalize DEV values so that depressions in gentle, low-relief terrain, like broad river valleys or the Puget lowlands, are recognized just as well as depressions in high-relief terrain, like alpine glacial cirques.

Unless the topography is totally flat, values of gradient, curvature, and local relief can vary depending on the distance (dx) over which they are measured. Over a few meters, say, features like tree-throw pits and gullies will affect topographic indices. Over tens of meters, these measures discern features like small bedrock hollows or hummocks in glacial and landslide

[3] https://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/curvature-function.htm

deposits. Over hundreds of meters we see effects of drumlins and ice-sheet outwash channels[4]. Over thousands of meters we see broad regional depressions. Landforms at multiple scales, spanning tens to thousands of meters, influence the flow paths of water through the landscape. To identify wetlands using topographic indicators the full range of length scales must be explored.

Methods

Study sites

Data collection and model evaluations were performed at four study areas in western Washington (Figure 2). Study areas include both low-relief terrain dominated by Holocene glacial and alluvial (river) deposits and higher-relief alpine terrain. The Puyallup study area spans both (Figure 3). The western portion consists primarily of continental ice-sheet deposits and landforms; the eastern portion extends into the foothills of Mount Rainier and includes steep U-shaped alpine glacial valleys. The adjacent Mashel basin lies primarily in higher-relief terrain, with few ice-sheet deposits but abundant alpine glacial landforms and deposits (Figure 3). Alpine zones in both the Puyallup and Mashel are underlain predominately by igneous and volcanic rock types. The Coulter Creek area consists almost entirely of continental ice-sheet deposits and landforms (Figure 4). The Hoh study area includes a broad valley filled with alluvial

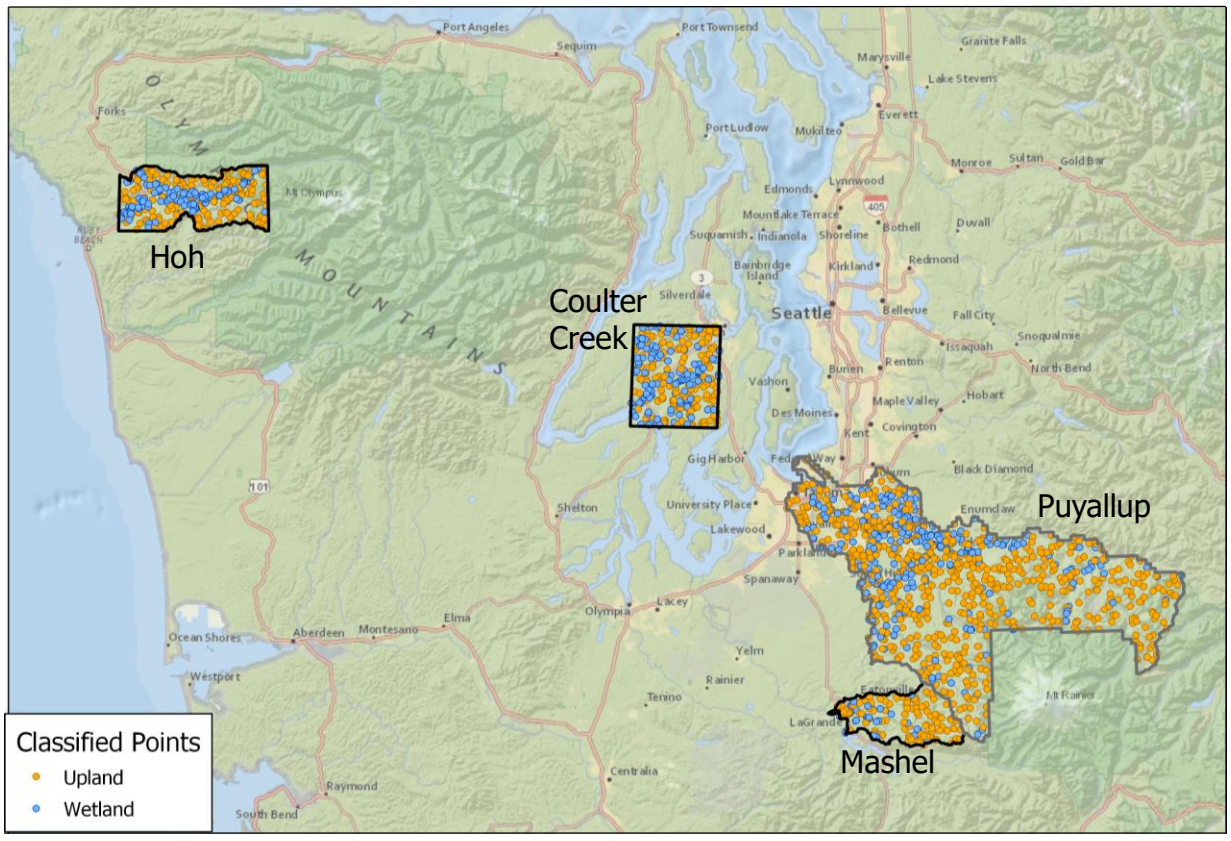

Figure 2. Study sites.

[4] See https://www.dnr.wa.gov/publications/ger_presentations_coe_glacial_landforms_puget_lowland.pdf

and alpine glacial deposits, with steep alpine zones predominately in marine sedimentary rocks (Figure 5).

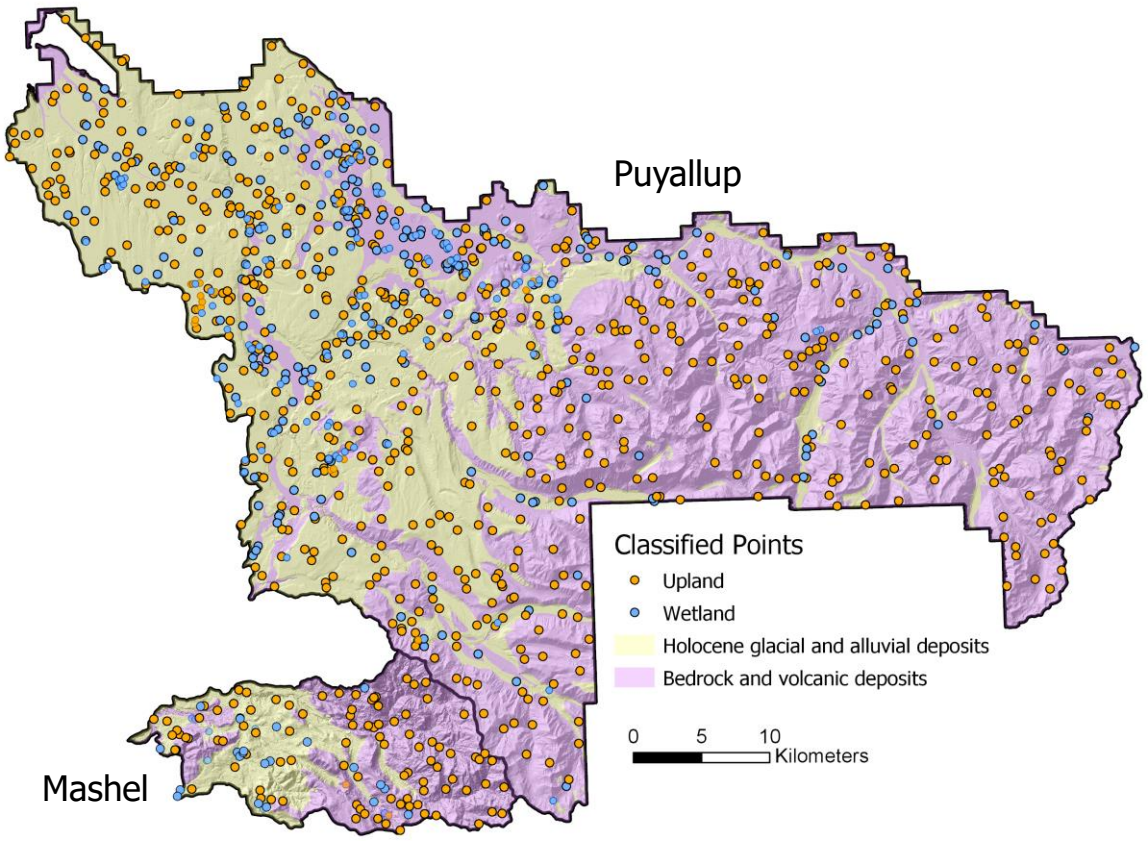

Figure 3. The Puyallup and Mashel study areas.

Precipitation in the lower elevations is predominately rainfall; all sites but Coulter Creek have snow at higher elevations. The Puyallup and Mashel have a similar range of mean annual rainfall, Coulter Creek is slightly dryer, and the Hoh is considerably wetter (Table 1, Figure 6).

Table 1	**Area (km^2)**	**Elevation (m)**			**Mean Annual Precipitation (mm)**		
		Min	**Max**	**Mean**	**Min**	**Max**	**Mean**
Puyallup	1967	0	2127	660	972	2896	1554
Mashel	231	452	4881	2211	912	2170	1677
Coulter Creek	384	0	1763	381	1280	1865	1423
Hoh	358	94	4085	1086	2762	4305	3335

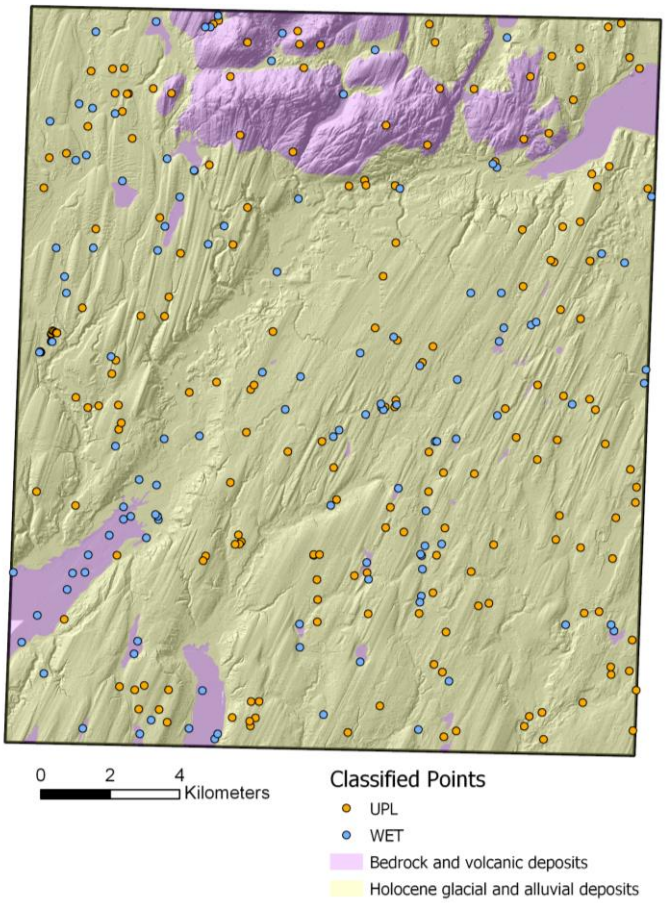

Figure 4. Coulter Creek study area.

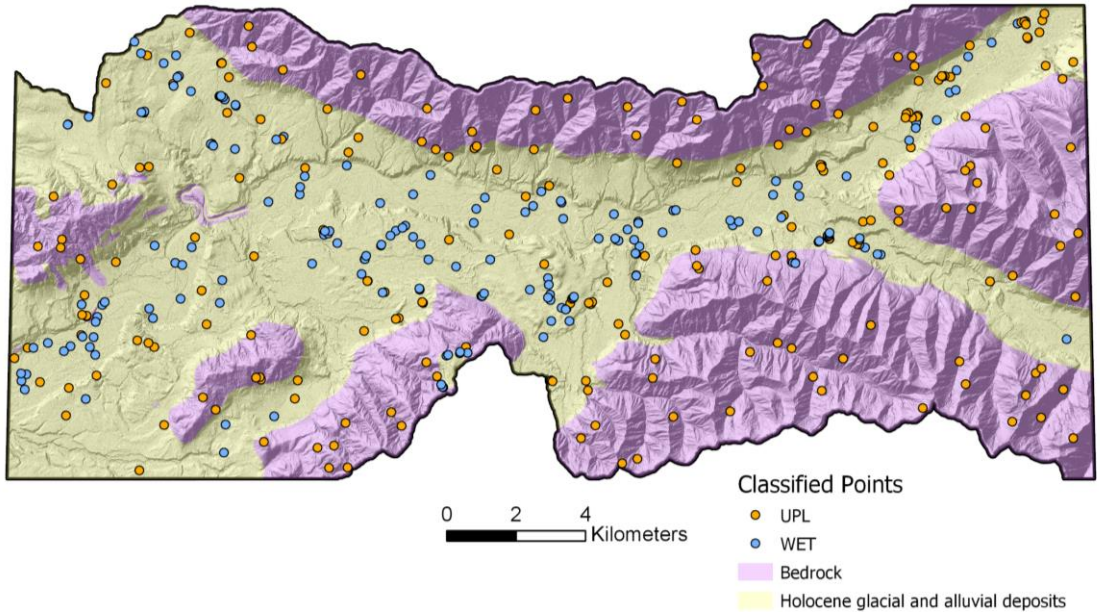

Figure 5. Hoh study area.

Data analysis

A primary goal of this project is to develop an ArcGIS Pro toolset for mapping wetland potential, therefore we have worked with data that are publicly available and analysis tools that are available directly within ArcGIS or can be implemented as scripts within ArcGIS toolboxes.

Topographic indices were calculated using compiled Fortan programs from the Netstream program suite (Miller, 2003). These programs implement the procedures described above for calculating gradient, curvature, and local relief over any length scale. The Surface Metrics python[5] script written for this project provides an ArcGIS Pro toolbox for using these programs to build the associated raster files. We calculated gradient, curvature, and local relief values over three length scales that we felt captured the range of variability across the landscape: 50 m, 150 m, and 300 mThe raster files were then used as input data for building the random forest models.

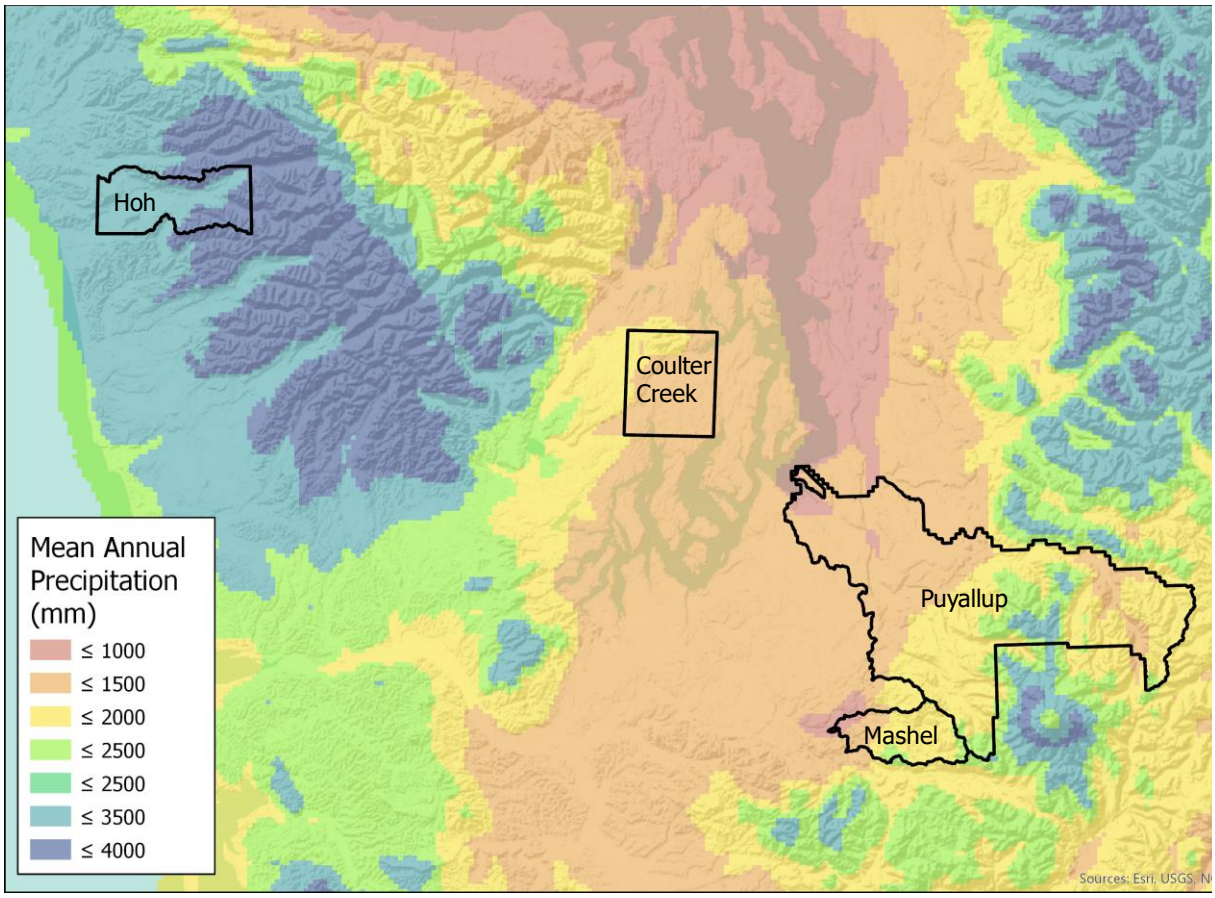

Figure 6. Mean annual precipitation across western Washington and the four study sites. Data from PRISM http://www.prism.oregonstate.edu/

[5] https://www.python.org/

The random forest model (Breiman, 2001) is implemented in the randomForest package[6] in the R statistical language[7]. We used the R-ArcGIS Bridge[8] to build the Wetland Tools ArcGIS Pro toolbox that implement scripts that call R functions to build and apply random forest models.

Data for training (calibrating) and evaluating the models consists of point locations classified as wetland or upland (see Data collection section below). The resulting classifications are verified through aerial photo interpretation and field surveys. The data collected for this project are stored as a point feature class within ArcGIS and are available for building and evaluating random forest models through the toolboxes developed for this project. The gradient, curvature, and local relief values for each length scale are determined for each point in the data set by overlaying the points on the rasters. Bilinear interpolation to the four nearest raster grid points is used to determine the value when the data points do not fall directly on a raster grid point (i.e., raster cell corner). Each data point then has a classification, wetland or upland, and a list of terrain attributes (gradient, plan curvature, profile curvature, and local relief at 50-, 150-, and 300-m length scales). This table of values is then used to build the random forest model.

The R random forest package provides two useful measures of model performance. One is the "out-of-bag"[9] error rate. The random forest consists of many individual decision trees created using a boot-strap sample (random sample with replacement) of the data points. This procedure is called "bagging"; the sampled points for any tree are "in the bag", the excluded points are "out of the bag". The prediction obtained from the average of the entire ensemble of trees in the forest is less sensitive to noise in the input data than any individual tree. Additionally, each tree can be used to predict the class of the out-of-bag points that were not used from the data set to build that tree. The majority class of all out-of-bag predictions for that point is compared to the observed class for that point. Repeated for all data points, this provides an estimate of model error in terms of the proportion of out-of-bag sample classes (wetland or upland) correctly predicted. The second indicator of model performance is a confusion matrix[10], which shows how many data points were correctly classified, how many wetland points were classified as upland, and how many upland points were classified as wetland.

For any location, the resulting random forest model can then use the terrain-attribute values at that location to calculate a probability that the location is wetland or upland. The model can thus build a new raster showing the predicted probability that a wetland will be found at each DEM grid point. Likewise, the model can predict probability of wetland occurrence for a new set of data points. The predicted probability can then be compared to the observed class (wetland or upland) to provide a test of model predictions. A random forest model can thus be built using classified point data and interpolated to predict wetland occurrence within the same

[6] https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
[7] https://www.r-project.org/
[8] https://r-arcgis.github.io/
[9] https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf
[10] https://en.wikipedia.org/wiki/Confusion_matrix

basin, or extrapolated to predict wetland occurrence in another basin, and new data points can be collected to evaluate model performance.

Data collection

Phase 1 GRTS sampling

We initially trained and validated our model for the Puyallup and Mashel study areas using data developed under the EPA Grant Number CD01J09401, Improved Wetland Identification for Conservation and Regulatory Priorities (Halabisky, 2019). This dataset was created using a modified multi-layer stratified-random sampling design based on Generalized Random Tessellation Stratified (GRTS) sampling protocol (Stevens and Olsen, 2004) implemented in spsurvey[11] in the software program R[12]. Points were stratified first using slope and accessibility and then second on landcover class. We determined low slope and high slope areas using a threshold calculated from the 99th percentile of mean slope for all wetland polygons in the current National Wetlands Inventory[13] (NWI) for the Puyallup and Mashel study areas. We defined accessible regions as all public lands and areas between 15 and 60 meters of all public roads to allow for roadside surveys.

This provided four strata:

- High slope - Inaccessible
- High slope – Accessible
- Low slope – Accessible
- Low slope – Inaccessible

We used the secondary stratification only for low-slope areas (accessible & inaccessible) based on the following 6 land cover classes: developed, agriculture, grass/bare, forests/shrub & wetland. This secondary stratification was derived from the NOAA C-CAP[14] landcover dataset. We sampled each of the four strata in the following proportions: High slope-Inaccessible - 1/6 of total sample points, High slope-Accessible -1/6 of total sample points, Low slope-Inaccessible - 1/3 of total sample points, Low slope-Accessible - 1/3 of total sample points. We used these proportions to ensure enough sample points fell in wetland areas, which are generally in low slope areas, while still sampling in areas with a lower likelihood of having wetlands to ensure good coverage.

Phase 2 stratified random sample

We derived the sample points for the Coulter Creek and Hoh study area by first running the random forest model trained on the Puyallup study area. We used the Puyallup trained model to stratify our sample points, because it provided an efficient way to identify potential wetland areas as well as areas where there is high model uncertainty (i.e., probability near 0.5). This

[11] https://cran.r-project.org/web/packages/spsurvey/spsurvey.pdf
[12] https://www.r-project.org/
[13] https://www.fws.gov/wetlands/
[14] https://coast.noaa.gov/digitalcoast/tools/lca.html

provided a raster of wetland probability from 0 to 1. We stratified 400 sample points equally into four strata based on the preliminary wetland probability raster: 0 – 0.25, 0.25 – 0.5, 0.5 – 0.75, 0.75 – 1.0. By sampling in areas of high model uncertainty we felt we could improve final model results with fewer areas falling in the middle two strata.

Data sets

The data sets created for each of the study areas were evaluated using the same two-stage approach. Each sample point was first assessed using aerial imagery and other available datasets including: Google street view imagery, Google Earth historic imagery, lidar data from the Washington State Department of Natural Resources lidar viewer, and pre-existing wetland inventories (i.e., NWI, the Pierce County wetland inventory[15], and the NOAA C-CAP data). If a point could not be determined as a wetland or an upland in aerial imagery or any other available datasets, it was marked as unknown. A secondary assessment in the field (i.e., on the ground) focused on a proportion of those points. We added ancillary data points for wetlands observed in aerial imagery verification that were not identified in the NWI and those that we came across while assessing points in the field. We also added ancillary data points for non-wetlands that were mapped in NWI as wetlands. All ancillary data points were collected opportunistically and only used as training data. We were unable to identify any slope wetlands to include in the training or validation dataset. Therefore, the model could not predict or validate the presence of slope wetlands.

Table 2. Number of GRTS sample points for the Puyallup watershed assessed in the office and field, number of ancillary points added to the dataset (top table); GRTS and ancillary points used for training and GRTS points used for validation (bottom table).

GRTS	Field	Office	Total	**Ancillary**	Field	Office	Total
Wetland	16	335	351	Wetland	36	73	109
Upland	25	630	655	Upland	23	0	23
Unknown	1	232	233				
Total	42	1197	1239		59	73	132

Training	GRTS	Ancillary	Total	**Validation**	GRTS
Wetland	171	109	280	Wetland	75
Upland	461	23	484	Upland	193
Total	632	132	764	Total	268

We assessed 1,239 GRTS points and added 132 ancillary points for the Puyallup watershed (Table 2). Of the 1,239 GRTS points, 42 points were followed up and assessed by visiting them in the field. Of the 132 ancillary points, 59 were validated in the field.

[15] https://gisdata-piercecowa.opendata.arcgis.com/datasets/cwi-wetlands-delineation

Before using sample data points for our analysis, we removed any point that could not be determined as being in a wetland or upland. We also removed points if we felt they were overrepresented in the dataset. In the Puyallup watershed, for example, 103 points were located in Lake Tapps and classified as wetland, creating a disproportionate number of points representing a lacustrine wetland. Of the assessed GRTS points, we reserved 193 upland points and 75 wetland points for validation. We derived the Puyallup training dataset from the remaining GRTS sample points and supplemented with 132 ancillary points. The training dataset had 484 upland points and 280 wetland points.

Table 3. Number of sample points used for training and validation for the Hoh Coulter Creek, and Mashel study areas assessed in the office and field. Field points include ancillary points collected opportunistically as part of the field effort.

Hoh	Field	Office	Total
Unknown	13	90	103
Upland	84	156	240
Wetland	64	114	178
Total	161	360	521
Coulter Creek			
Unknown	0	125	125
Upland	17	177	194
Wetland	19	100	119
Total	36	402	438
Mashel			
Upland	19	76	95
Wetland	6	18	24
Total	25	94	119

We did not divide the sample data points into training and validation for the Mashel, Coulter Creek, and Hoh study areas (Table 3). Rather, because of the relatively small number of points, we used all the data collected for each of the study areas as input data to train the model. We also included all ancillary data collected opportunistically in the field. Evaluation of models is based on the "out-of-bag" error. Because of the limitation of the validation datasets, we only used the results to provide a coarse comparative metric to results from models trained using the NWI or the Puyallup model.

Results

Model performance

The random forest model for the Puyallup watershed had an overall accuracy of 96.6% (Figure 7). The wetland error of commission (false positives) was 4.3% and the error of omission (missed wetlands) was 8.0%. In contrast, the current NWI for the Puyallup watershed had an overall accuracy of 88.1% and an error of commission of 2.1% and an error of omission of 41.8%. The random forest model for the Puyallup mapped four times the area of wetlands within forest lands than the NWI. The random forest model trained on the Puyallup identified slope, dev150, dev300 as the three variables that contributed the most importance to the model as measured by the Gini importance or the mean decrease in impurity (MDI) (Figure 8). MDI is a commonly used metric that can be derived from a random forest model. MDI counts the times a feature (e.g., slope) is used to split a node in a decision tree, weighted by the number of samples it splits.

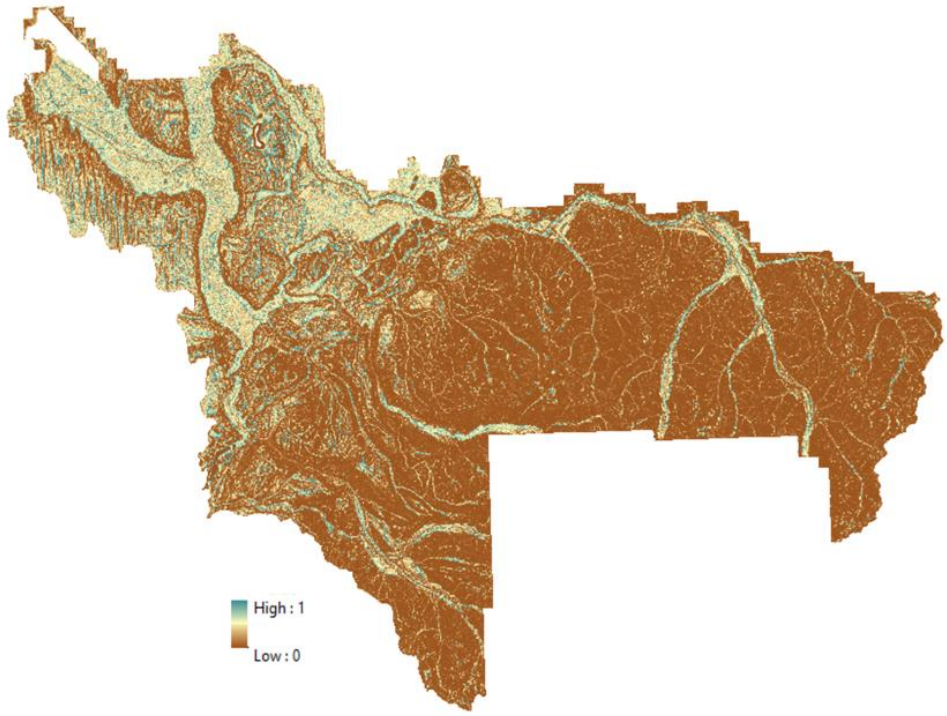

Figure 7. Results from the WIP tool random forest model. The raster output produces a probability that a pixel is a wetland from 0-1. Pixel values close to 1 (Blue pixels) have a high probability of being a wetland, while pixel values close to 0 (brown pixels) have a low probability.

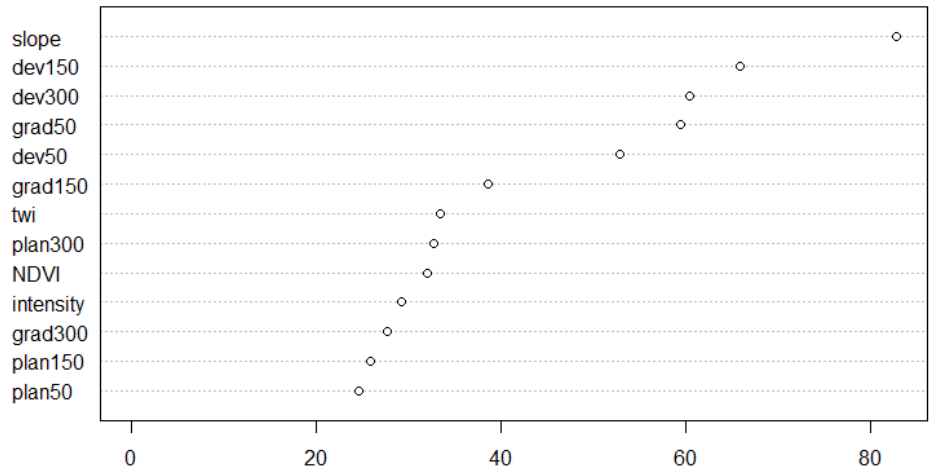

Figure 8. Results from the WIP tool random forest model. One output of a random forest model is the Gini coeffeicient which is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The variables at the top of the chart contributed the most definition in the random forest model and are considered to be of higher model importance.

When the Puyallup model was extrapolated to the other study areas it provided potential locations of wetlands across each study area. The overall accuracy of the random forest model trained using the Puyallup data when run on the Mashel watershed had an overall accuracy of 96% and an error of omission of 21% (Figure 9). When the random forest model was run using training data specific to the Mashel the overall accuracy only had a slight improvement (97% overall accuracy, 16% error of omission). However, both models had much higher overall accuracy than the NWI for the area, which was 86.5% with an error of omission of 54%.

The model trained on the Puyallup watershed could be extrapolated to the Mashel watershed without much decrease in accuracy (96% OA v. 97% OA). However, the Puyallup model did not perform as well for the two other study areas. The Mashel watershed is neighboring the Puyallup watershed and has similar topographic and wetland characteristics, while the Hoh and Coulter Creek study area are very different than the Puyallup watershed.

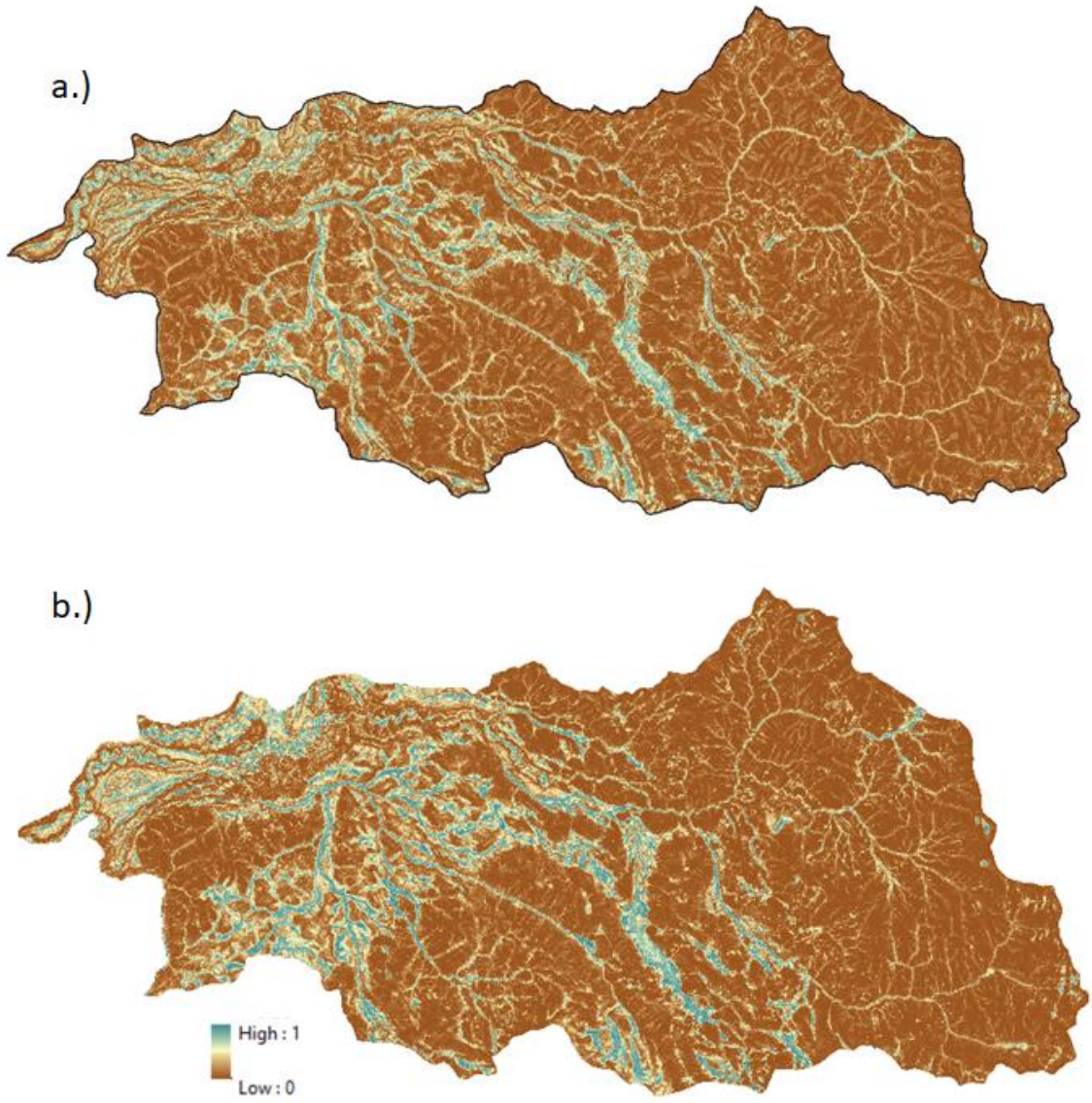

Figure 9. The model trained on the Puyallup data performed well when extrapolated to the Mashel watershed (a.). The model only slightly improved when data from the Mashel was used to create a new random forest model for the Mashel (b.)

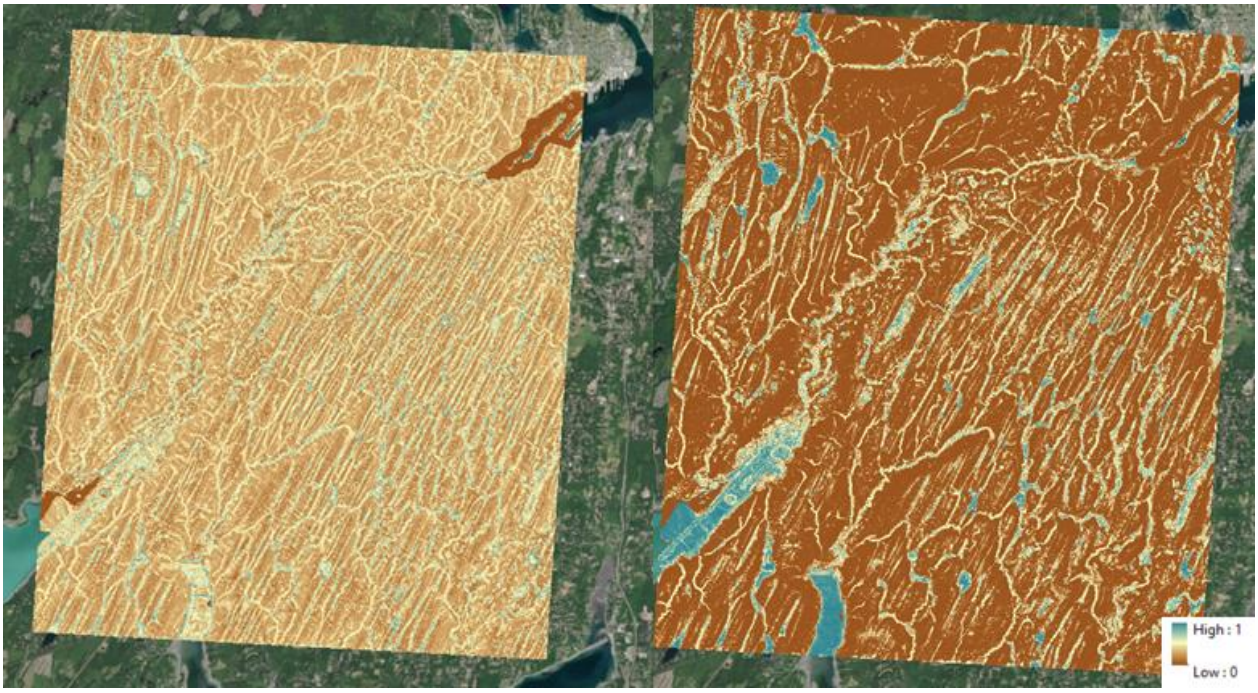

Figure 10. The model provided greater certainty when trained on the sample data collected in Coulter Creek. The left image shows results created by extrapolating the Puyallup model; the right image shows refined model results created from the sample training data specific to the study area.

While the preliminary model did not perform as well for the two other study areas it was useful at identifying wetlands missed in the NWI and was helpful for sample point stratification. The preliminary random forest model results improved when we refined our models for the Coulter Creek and Hoh watershed study areas using the sample data specific to each study area (Figure 10 & 11). The raster probabilities moved toward greater certainty of being a wetland or an upland.

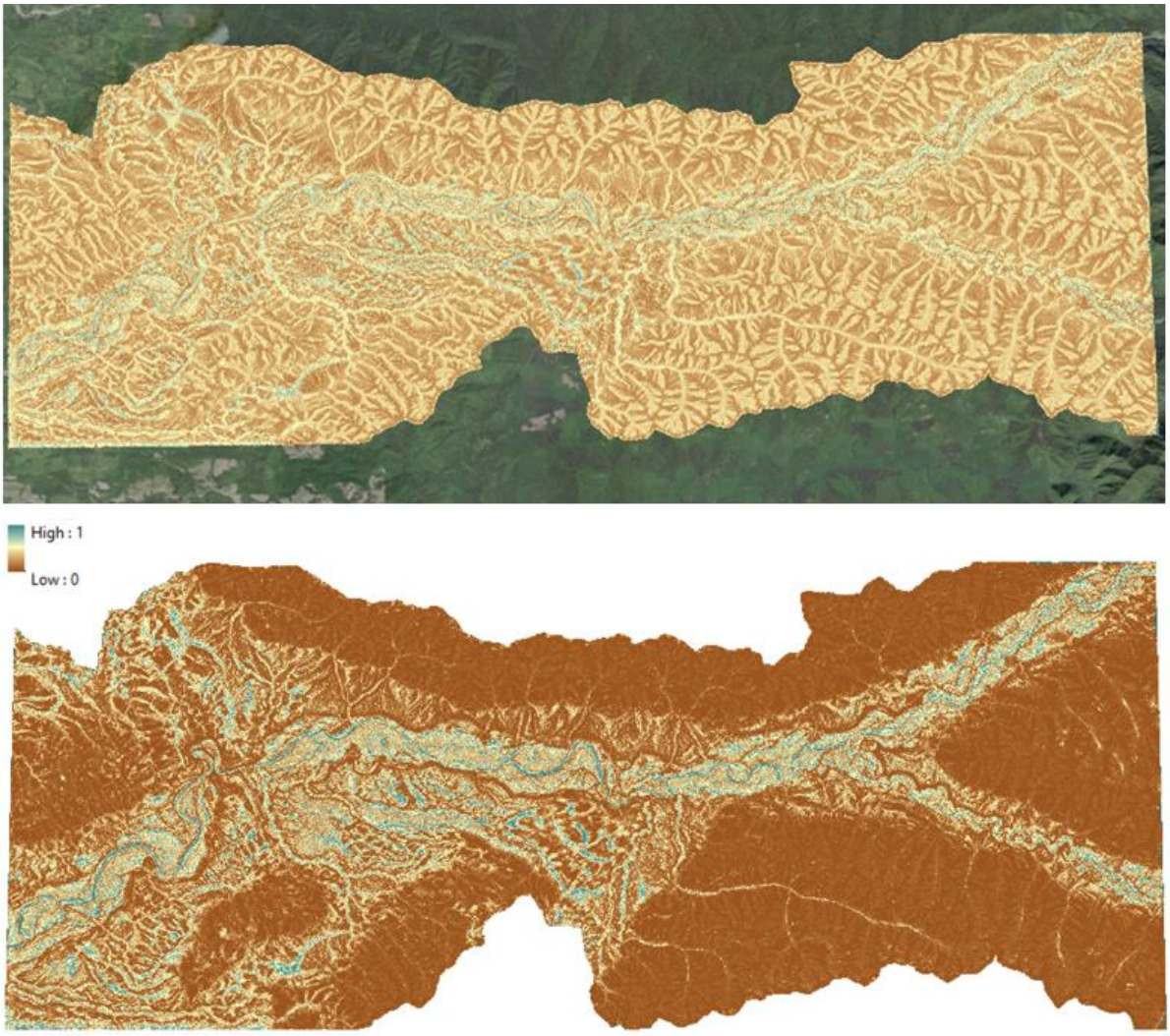

Figure 11. Model results improved and moved to greater certainty when the random forest model was trained on the sample data collected in the Hoh study area. The top image shows model results created by extrapolating the Puyallup model to Hoh study area. The image on the bottom shows refined model results created from the sample training data collected specific to the Hoh study area.

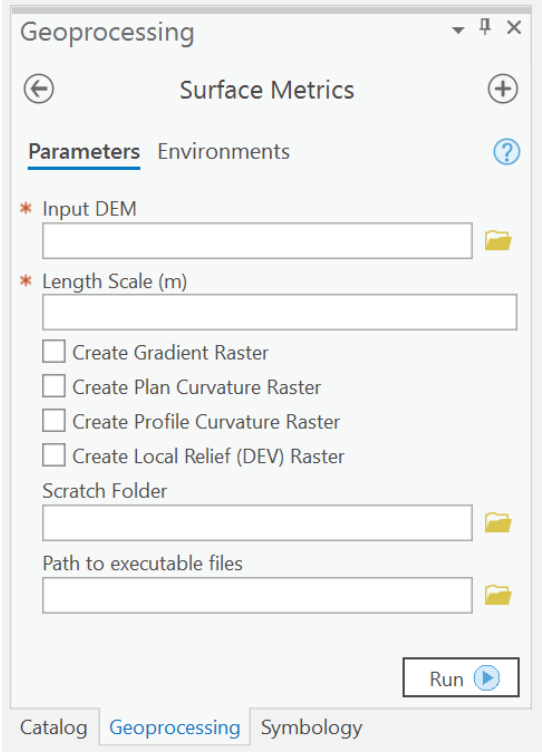

Figure 12. Example of ArcGIS toolbox for creating terrain-attribute raster files.

Toolboxes for ArcGIS Pro

For this project, we have developed two toolboxes for ArcGIS Pro: DEMutilities and Wetland Tools. Currently, the DEMutilities toolbox contains one script tool: Surface Metrics. The dialog box for this script is shown in Figure 12. This dialog box provides an interface familiar to ArcGIS users. Descriptions of each item in the dialog box are provided in the metadata for the tool, which are displayed in ArcGIS when the toolbox is opened.

The Wetland Tools toolbox contains two scripts: Build Random Forest and Run Random Forest. The Build Random Forest script is used with a set of classified (wetland or upland) points to train a random forest model. The Run Random Forest model is used to apply an existing model (built with the build Random Forest script) to other locations and/or with other classified point data sets. The dialog boxes are shown in Figure 13.

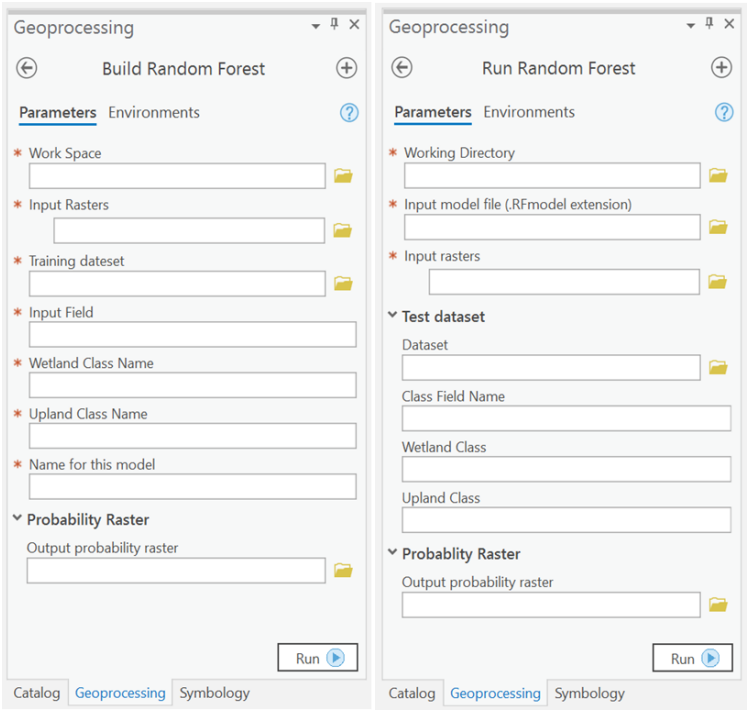

Figure 13. ArcGIS Pro dialog boxes in the Wetland Tools toolbox for building and running random forest models.

Discussion

Model interpretation

Each tree in the random forest model built with the Build Random Forest script has divided the values found in the input rasters into a large set of domains within which the input training data points are primarily of a single class (Wetland or Upland). The degree to which these domains isolate each class depends on the degree to which the information contained in the rasters delineates spatial controls on wetland occurrence. When we apply a random forest model to predict wetland class (Wetland or Upland) for each cell in a DEM, the model takes the value for that cell from each of the input rasters and runs it through each tree; that is, it determines if that point falls in a domain classified as Wetland or Upland. The random forest contains many trees (typically hundreds to thousands, the script currently uses 200) and each tree provides a "vote" – a predicted class – for each DEM cell. The proportion of wetland votes for a cell is interpreted as the predicted probability that the cell is a wetland. Upon completion of our model run we selected a probability equal to or greater than 50% to classify each cell as a wetland to build the accuracy assessments presented in this document. However, a user can select a higher or lower probability depending on if they want to reduce errors of omission or errors of commission. For example, if a user is concerned with overmapping wetlands (commission errors) they can select a higher cutoff such as 80%. One of the outputs of the random forest model is a curve that shows how accuracy changes depending on the cutoff used and can be helpful to select a cutoff to balance these errors and optimize overall accuracy. Certainly, a user does not need to select any cutoff and can use the probability output as a screening tool to identify areas of high, medium, and low probability of being a wetland.

The predicted probability and resulting classification depend on how well the information provided by the input rasters identifies physical controls on wetland occurrence. We can gauge model performance by how well it predicts the class of the training data (using the out-of-bag error estimate and confusion matrix obtained when the model was built) and by seeing how well the model can predict the class of other classified point locations (using the Run Random Forest script).

The Build Random Forest tool can be used to examine model performance for different sets of input rasters. The Surface Metrics tool can be used to build rasters for terrain attributes characterized over different length scales. Together, these tools can be used to identify the combination of terrain attributes and length scales that best identify wetland class for training and test data sets. The model output also includes a table showing the relative importance of each raster in determining wetland class.

The importance of training data

The ability of a random forest model to predict wetland occurrence primarily depends on how well the data used to train the model represent the range of wetland types and locations that exist on the ground. Any biases or errors of omission in the training data will produce results with the same bias and error. If the training data does not include some wetland type, slope

seeps for example, then the resulting model will not provide any indication of the probability of encountering that wetland type. Deriving a probabilistic un-biased sample to create a training dataset when you don't know where all of the wetlands are to begin with presents an interesting challenge. Additionally, wetlands are considered rare features on the landscape and a purely random sample would result in a training dataset that had too few wetlands sample points to build a good model. However, there is evidence in the scientific literature that correcting for this by oversampling wetlands on the landscape creates imbalanced results that over-predict wetlands (Halabisky 2017).

For this project we tested out two approaches to create a balanced training dataset despite not having a complete map of wetland locations. For our Puyallup model we used a GRTS sample design developed by the EPA to stratify our sample points using land cover and slope. The idea was that increasing our sample size in areas of low slope would increase our chance of sampling in a wetland. However, the GRTS sample design is complex and may be difficult for many users to implement. For the Hoh and Coulter Creek study areas we ran a preliminary model trained on the National Wetland Inventory and then we stratified our sample using the preliminary model output. The sampling methods for the Hoh and Coulter Creek was easier to implement and provided a good distribution of wetland and upland samples.

In order to balance the need for adequate samples that fall in wetlands without skewing the sample too much towards wetlands, we recommend aiming for a sample that has between one-third to one-half of the points falling in wetlands and the remaining points falling in upland. This project did not focus on trying to assure a balanced sample of wetland types (e.g. Cowardin class) because of the complexity of doing that without a priori knowledge of the location and distribution of wetland classes.

Extension of model to new locations

As noted above, the ability of a random forest model to predict wetland occurrence depends on how well the data used to train the model represent the range of wetlands types and locations that exist on the ground. A model trained on one study area, but run on a different study area, will produce accurate results if the two study areas are similar. A key point is that not only will model values from the variables themselves vary for different study areas, but often the variables themselves will vary in importance as well. For instance, in one watershed that contains many surface-water driven wetlands, the topographic wetness index may be the most important variable that describes the variability between wetlands and uplands, but in another study area the slope may be ranked as a more important variable.

Model availability

The Fortran programs used to build the raster data sets are licensed under the Gnu Public License[16], version 3. The python and R scripts for the DEMutilities and Wetland Tool ArcGIS Pro

[16] https://www.gnu.org/licenses/gpl-3.0.en.html

toolboxes are posted to a public github repository at https://github.com/DanMillerM2/ForestedWetlands. TerrainWorks maintains all software developed during collaborative projects. Bug reports, comments, and feature requests for these toolboxes can be submitted at http://www.terrainworks.com/contact-us.

Future

We have designed the tools for this project expecting that they will evolve over time. The scripts and software are licensed as open source and publicly available via github. The random forest model can readily accommodate new terrain attributes as explanatory variables and the scripts in the Wetland Tools toolbox can accommodate any input grid that can be imported to ArcGIS. We have started with a simple set of terrain attributes in the Surface Metrics script and expect that users will experiment with other input data types, such as soil depth and conductivity that can be extracted from SSURGO[17] and STATSGO data. Over time, we hope to add additional raster outputs to the DEMutilities toolbox, including the topographic wetness index (Beven and Kirkby, 1979), depth-to-water[18], and topographic position (Weiss, 2001). We also expect that additional functionality will be added to the scripts in the Wetland Tools toolbox, including options for tuning the random forest model, additional metrics such as cross validation for model evaluation, ability to generate a stratified random set of points for model validation, and options to calculate the probable wetland area for any specified polygon using the probability raster.

We are conducting further testing and validation of the WIP tool in forested areas in Washington State as part of a 2019 NASA Carbon Monitoring Science grant, titled “Teal Carbon: A stakeholder driven monitoring of forest wetland carbon”. Wetland probability outputs from these models will be available upon completion through the NASA CMS website[19]. The study areas for these three locations are the Hoh watershed, the Mashel watershed, and the Colville National Forest.

Conclusion

Forested wetlands have proven challenging to identify using remotely sensed multispectral and optical data because the vegetation unique to wetlands can be hidden below forest canopy. We have shown that topographic attributes alone, derived from high-resolution lidar DEMs, can be used to quantify probability of wetland occurrence with high confidence. Lidar data is already widely available across Washington[20] and in several years such data should be available for the entire state. The methods and GIS-based tools developed with this project enable use of the data to map wetland potential and to evaluate and improve resulting wetland-probability maps using photo- and field-verified data. We expect that the capabilities of these tools will expand

[17] https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/office/ssr12/tr/?cid=nrcs142p2_010596
[18] Rather than the wet-areas algorithm of Murphy et al. (2007), we expect to use the height-above-channel algorithm described in this blog post: http://netmapchallenges.blogspot.com/
[19] https://carbon.nasa.gov/cgi-bin/available_archived_products.pl
[20] http://lidarportal.dnr.wa.gov/

over time as users determine the most effective topographic attributes for identifying wetlands and identify other applications for the resulting models.

References

Adam, E., Mutanga, O., and Rugege, D., 2009, Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review: Wetlands Ecology and Management, v. 18, no. 3, p. 281-296.

Beven, K. J., and Kirkby, M. J., 1979, A physically based, variable contributing area model of basin hydrology: Hydrological Sciences Journal, v. 24, no. 1, p. 43-69.

Breiman, L., 2001, Random Forests: Machine Learning, v. 45, p. 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., 1984, Classification and regression trees, Monterey, CA, Wadsworth & Brooks / Cole Advanced Books & Software.

Brinson, M. M., 1993, A hydrogeomorphic classification for wetlands: U.S. Army Corps of Engineers.

Buchanan, B. P., Fleming, M., Schneider, R. L., Richards, B. K., Archibald, J., Qiu, Z., and Walter, M. T., 2014, Evaluating topographic wetness indices across central New York agricultural landscapes: Hydrol. Earth Syst. Sci., v. 18, p. 3279-3299.

Chignell, S. M., Luizza, M. W., Skach, S., Young, N. E., Evangelista, P. H., Pettorelli, N., and Fatoyinbo, L., 2018, An integrative modeling approach to mapping wetlands and riparian areas in a heterogeneous Rocky Mountain watershed: Remote Sensing in Ecology and Conservation, v. 4, no. 2, p. 150-165.

Fink, C. M., and Drohan, P. J., 2016, High Resolution Hydric Soil Mapping using LiDAR Digital Terrain Modeling: Soil Science Society of America Journal, v. 80, no. 2, p. 355.

Florinsky, I. V., 1998, Accuracy of local topographic variables derived from digital elevation models: Int. J. Geographical Information Science, v. 12, no. 1, p. 47-61.

Florinsky, I. V., 2016, Digital Terrain Analysis in Soil Science and Geology, Acedemic Press, Elsevier, 486 p.:

Halabisky, M., 2019, Improved wetland identification for conservation and regulatory priorities: University of Washington.

Hastie, T., Tibshirani, R., and Friedman, J. H., 2017, The Elements of Statistical Learning, Springer, Springer Series in Statistics, 745 p.:

Hengl, T., and Reuter, H. I., 2009, Geomorphometry. Concepts, Software, Applications, *in* Hartemink, A. E., and McBratney, A. B., eds., Developments in Soil Science, Elsevier.

Janisch, J. E., Foster, A. D., and Ehinger, W. J., 2011, Characteristics of small headwater wetlands in second-growth forests of Washington, USA: Forest Ecology and Management, v. 261, p. 1265-1274.

Jasiewicz, J., and Stepinski, T. F., 2013, Geomorphons - a pattern recognition approach to classification and mapping of landforms: Geomorphology, v. 182, p. 147-156.

Kloiber, S. M., Macleod, R. D., Smith, A. J., Knight, J. F., and Huberty, B. J., 2015, A Semi-Automated, Multi-Source Data Fusion Update of a Wetland Inventory for East-Central Minnesota, USA: Wetlands, v. 35, no. 2, p. 335-348.

Knight, J. F., Tolcser, B. P., Corcoran, J. M., and Rampi, L. P., 2013, The effects of data selection and thematic detail on the accuracy of high spatial resolution wetland classifications: Photogrammetric Engineering & Remote Sensing, v. 79, no. 7, p. 613-623.

Lang, M., McCarty, G., Oesterling, R., and Yeo, I.-Y., 2013, Topographic metrics for improved mapping of forested wetlands: Wetlands, v. 33, p. 141*155.

Lang, M. W., and McCarty, G. W., 2009, Lidar intensity for improved detection of inundation below the forest canopy: Wetlands, v. 29, no. 4, p. 1166-1178.

Mahdavi, S., Salehi, B., Granger, J., Amani, M., Brisco, B., and Huang, W., 2017, Remote sensing for wetland classification: a comprehensive review: GIScience & Remote Sensing, v. 55, no. 5, p. 623-658.

Maxwell, A. E., Warner, T. A., and Fang, F., 2018, Implementation of machine-learning classification in remote sensing: an applied review: International Journal of Remote Sensing, v. 39, no. 9, p. 2784-2817.

Maxwell, A. E., Warner, T. A., and Strager, M. P., 2016, Predicting Palustrine Wetland Probability Using Random Forest Machine Learning and Digital Elevation Data-Derived Terrain Variables: Photogrammetric Engineering & Remote Sensing, v. 82, no. 6, p. 437-447.

Miller, D., Benda, L., DePasquale, J., and Albert, D., 2015, Creation of a digital flowline network from IfSAR 5-m DEMs for the Matanuska-Susitna Basins: a resource for update of the National Hydrographic Dataset in Alaska.

Miller, D. J., 2003, Programs for DEM Analysis, Landscape Dynamics and Forest Management, General Technical Report RMRS-GTR-101CD: Fort Collins, CO, USA, USDA Forest Service, Rocky Mountain Research Station.

Murphy, P. N. C., Ogilvie, J., Connor, K., and Arp, P. A., 2007, Mapping wetlands: a comparison of two different approaches for New Brunswick, Canada: Wetlands, v. 27, no. 4, p. 846-854.

O'Neil, G. L., Goodall, J. L., and Watson, L. T., 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, v. 559, p. 192-208.

Shi, X., Zhu, A.-X., Burt, J., Choi, W., Wang, R., Pei, T., Li, B., and Qin, C., 2007, An experiment using a circular neighborhood to calculate slope gradient from a DEM: Photogrammetric Engineering & Remote Sensing, v. 73, no. 2, p. 143-154.

Stevens, D. L., and Olsen, A. R., 2004, Spatially Balanced Sampling of Natural Resources: Journal of the American Statistical Association, v. 99, no. 465, p. 262-278.

Tyralis, H., Papacharalampous, G., and Langousis, A., 2019, A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources: Water, v. 11, no. 5.

Weiss, A. D., 2001, Topographic position and landforms analysis, ESRI Users Conference: San Diego, CA.

White, B., Ogilvie, J., Campbell, D. M. H., Hiltz, D., Gauthier, B., DChisholm, H. K., H., Wen, H. K., Murphy, P. N. C., and Arp, P. A., 2012, Using the cartographic depth-to-water index to locate small streams and associated wet areas across landscapes: Canadian Water Resources Journal, v. 37, no. 4, p. 333-347.

Wilson, J. P., 2018, Environmental Applications of Digital Terrain Modeling, John Wiley & Sons, New Analytical Methods in Earth and Environmental Science, 336 p.:

Wu, Q., and Lane, C. R., 2016, Delineation and Quantification of Wetland Depressions in the Prairie Pothole Region of North Dakota: Wetlands, v. 36, no. 2, p. 215-227.

Zevenbergen, L. W., and Thorne, C. R., 1987, Quantitative analysis of land surface topography: Earth Surface Processes and Landforms, v. 12, p. 47-56.