12-7-2022

Random Forest Modeling Approach to Predict Streamflow Intermittence of Tennessee Headwaters using Flow Conditioned Parameter Grids

Berkay Tok

RANDOM FOREST MODELING APPROACH TO PREDICT STREAMFLOW INTERMITTENCE OF TENNESSEE HEADWATERS USING FLOW CONDITIONED PARAMETER GRIDS

by

Berkay Tok

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Earth Sciences

The University of Memphis

December 2022

Abstract

Water quality impairment in small tributaries due to soil erosion and stream degradation of West Tennessee is an ongoing problem. A method to model streamflow permanence can assist stream restoration work by supplementing ground monitoring and providing better targeting of conservation attempts in the most vulnerable areas. This project applied a random forest model by incorporating climatic and landcover data as predictors to create streamflow permanence data for the West Tennessee tributaries (Lower Mississippi-Hatchie Hydrologic Unit, HUC 4-801). Specifically, the applicability of the Flow Conditioned Parameter Grids (FCPG) process is tested to study if the process improves prediction results compared to raw predictor results. In addition, the model's ability to capture the effect of headwater lakes in increasing the probability of streamflow permanence in downstream reaches is investigated using two pairs of streams from the Tennessee Department of Environmental Conservation (TDEC) database of watershed water quality assessments. With the various predictor variable configurations tested in the model, an average of 25 percent Mean Squared Error (MSE) accuracy is acquired in the prediction of the streamflow permanence status of west Tennessee streams. The results showed processing FCPG layers did not provide an increase in prediction accuracy for this study. Validation of the model results using the test stream pairs was inconclusive. While the model did predict streamflow permanence downstream of one headwater lake and intermittent streamflow in the other stream of the pair. it predicted perennial flow for both streams in the second pair, regardless of the presence of a headwater lake. This project provides scalable and replicable methods using machine learning and remote sensing data to predict streamflow permanence at a 25% MSE rate.

Keywords: Streamflow Permanence, Random Forest, Remote Sensing, Headwaters

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction

Importance of headwater streams monitoring

Headwater streams are defined as first and second-order streams (Wohl, 2017) and represent the majority of stream channel length (Godsey & Kirchner, 2014; Leopold et al., 1964). Headwaters are one of the most endangered river ecosystems and continue to be threatened by land-use changes and river engineering (Wohl, 2017). Headwater streams influence downstream parts of the river network in various ways such as retainment or transmission of sediment and nutrients, organic and inorganic carbon, wood (Sando & Blasch, 2015), and the creation of habitats for diverse organisms (Wohl, 2017).

Previous research shows that headwater streams are crucial to hydrological networks (Sando & Blasch, 2015). Disturbances in headwater streams may affect the condition of the surrounding areas (Costigan et al., 2016). Removal of riparian vegetation from headwater streams for agricultural and urban development affects the flow between headwaters and downstream (Winter, 2007) and in turn, causes degradation of downstream systems (Wasser et al., 2015). Such disturbances occurring naturally or by human intervention may also influence downstream fish populations (Sando & Blasch, 2015). Previous research suggests that it is important to monitor the dynamics and condition of headwater streams, but in-situ monitoring of headwater streams is a challenging task, especially in small tributaries at regional scales (Eberts et al., 2019). Due to typically shallow headwater depths, these streams are vulnerable to blockage by small structural changes and disconnection (Wohl, 2017), and the geographically widespread nature of these streams along with the narrow and fragmented spatial configuration of riparian corridors make their status of degradation and disturbance difficult to monitor (Wasser et al., 2013).

In early studies of hydrology, it is understood that the length and density of headwater systems dynamically change year around and first-order streams were found to have the most variance (Blyth & Rodda, 1973; Day, 1978). Godsey and Kircher (2014) indicate that even stream order can show differences seasonally by two Strahler orders. For example, a first-order stream may become a second-order stream or completely dry out, while a second-order stream may become a first-, third-, or even fourth-order stream as drainage in the watershed changes as the seasonal variations occur. In-situ monitoring of headwater streams is a costly, slow process and bound to low sampling rates, resulting in low accuracy. In addition to the cost factor, some of these areas are even more challenging to map by hand due to logistical factors such as the steepness of the terrain or the density of surrounding forested areas (Godsey & Kirchner, 2014). More frequent observations could be useful when evaluating restoration efforts and determining which tributaries' surface flow might be affected by varying environmental factors such as land-cover, land-use, climate change, human activities, etc. (Godsey & Kirchner, 2014).

Streamflow Permanence in West Tennessee

Streamflow permanence is the flow status of a river that is defined by having a surface-water presence (Costigan et al., 2016) which affects the surrounding riparian corridors. In recent decades, there has been an increased interest among researchers in streamflow permanence and prediction, since smaller streams and their status are believed to be one of the biggest factors affecting hydrological systems. Smaller streams are also the least-researched and because of this, may also be the last to benefit from conservation efforts (Jaeger et al., 2021)

The prediction of streamflow permanence is important to understand the overall state of these smaller streams and their riparian corridors and allows restoration and preservation attempts to be focused or redirected to more important areas of the watershed. As with climate changes, it has become challenging to predict the streams that would have year-round water flow. Even though streamflow permanence studies have been conducted on small scales, recent studies show that models that combine the physical characteristics of headwaters with regional climatic data could be instrumental in gaining an understanding of regional-scale streamflow permanence which allows for annual and monthly changes to be reflected in the streamflow permanence model and its outcomes (Pate et al., 2020). Furthermore, stream integrity is usually lower in areas where urban and agricultural development is high (Costigan et al., 2016; Wasser et al., 2015). This is true for the smaller streams that flow through the densely-agricultural region of west Tennessee. These streams make up the tributary watersheds of the

Mississippi River in the Lower Mississippi-Hatchie hydrologic unit. Due to various factors such as soil erosion and stream channel degradation, water quality is a continuing issue in these areas. According to a 2015 report prepared for the U.S. Environmental Protection Agency (EPA), most watersheds in West Tennessee have a low health index and are prone to further degradation (K. Matthews et al., 2015).

Streamflow Permanence Modeling Approaches

Previous efforts to identify streamflow permanence include collecting direct observations using citizen scientists' contributions (Jensen et al., 2017; Turner & Richter, 2011), and implementing sensor-based technologies such as electrical resistance sensors (Goulsbra et al., 2014) and state loggers (Bhamjee & Lindsay, 2011). Airborne remote sensing methods can be incorporated for replicable monitoring strategies, especially for regional-scale studies (Hayes et al., 2014; Michez et al., 2017), while Johansen et al., (2010) found that remote sensing is the only viable technique for measurements in large spatial extent assessments. In the light of streamflow permanence monitoring efforts in rivers, it is suggested that compiling various sources of datasets would be beneficial to understanding the condition of headwater streams and developing a new model to understand the dynamics of streamflow permanence.

Among various modeling approaches in the field of hydrology, a machine learning approach, random forest, is a tree-based learning algorithm that is computationally efficient for large datasets (Oshiro et al., 2012). The random forest model creates decision trees from a set of samples (Biau & Scornet, 2016) and prediction is done by aggregation of multiple decisions (nodes) from decision trees (Svetnik et al., 2003). After the selection of a sample, the algorithm pulls a random assortment of features and begins testing a series of splits using each feature to predict a class (i.e., “wet”, or “dry”). The “wet” and “dry” classes are based on streamflow intermittency, which is defined as streams that have at least one dry (no) flow observation throughout the year. If an area has multiple observations at different times of the year, a single dry observation is sufficient to consider it as an intermittent stream and belongs to the “dry” class for this study’s purposes. Similarly, if a stream is observed to have surface flow in the dry months of the year, it is considered to be a perennial stream, and belongs to the “wet” class. Ephemeral streams are excluded from this study since generally these streams only flow when there is precipitation.

Jaeger et al. (2019) created the Probability of Streamflow Permanence (PROSPER) model where every variable is modified with the process called Flow Conditioned Parameter Grids (known as FCPG) for the prediction of streamflow permanence in the Northeast Pacific region. The FCPG approach accumulates predictor parameters according to the elevation and flow direction data of basins. The PROSPER model was able to predict streamflow permanence with 17%-22% prediction error rates. Shen et al., (2022) also applied a random forest model for error correction in prediction for large-scale models such as the PCRaster GLOBal Water Balance model (PCR-GLOBWB), which is a grid-based hydrological prediction model by Utrecht University, and their results showed improvement of the estimation of errors. Pham et al., (2020) found that compared to multiple linear regression models, random forest performed better in the prediction of streamflow status in snowmelt-driven watersheds. Other studies compared various machine learning models and neural networks and found extreme machine learning kernels had superior performance compared to other models (Li et al., 2019).

Objectives

The objective of this study is to produce a stream permanence dataset of Mississippi River tributaries in west Tennessee using a random forest model. Specifically, this work will test if a random forest model can predict the streamflow permanence in the headwater streams of these tributary rivers using moderate-resolution datasets of climatic and landcover predictors by applying the FCPG process used in Jaeger et. al (2019). The random forest learning algorithm is trained by samples of "wet" and "dry" classes of stream reaches identified from the database tables of the National Hydrography Dataset (NHD) Plus - High Resolution (HR) version and verified by aerial photographs. Additionally, the trained model is tested on two pairs of streams to examine if headwaters with and without lakes can be identified using streamflow permanence predictions. These stream pairs have been assessed by TDEC as part of their watershed water quality management and stewardship program (https://tdeconline.tn.gov/dwr/). Pair 1 is comprised of Meridian and

Bond Creeks in the Forked Deer watershed (Figure 1) and Pair 2 is made up of Spring and Piney Creeks in the Hatchie watershed (Figure 1). Meridian Creek has headwater lakes and is fully supporting of all uses (i.e. fish and aquatic life, human, livestock, agricultural, and industrial), while Bond Creek is not. Similarly, Spring Creek has some headwater lakes and is fully supporting, and Piney Creek has no headwater lakes and is not. TDEC defines 5 support categories: fully supporting waters meet all the quality criteria for their assigned use categories, supporting waters support some of their assigned use categories but have not been assessed for all uses, not assessed waters have not been assessed or relevant data is outdated, impaired waters have a Total Maximum Daily Load completed but not required, and not supporting waters are assessed and found to have one or more water quality issues.

As a result of this study, it is expected to have a better understanding of headwater streams in the area and create models that can be utilized by state agencies to improve and optimize management and monitoring efforts as new data become available.

Study Area

This study is focused on the Lower Mississippi-Hatchie Hydrologic Unit Code (HUC) 4-0801 region. This area has been chosen due to ongoing water quality issues in the area as assessed by previous reports (K. Matthews et al., 2015). The area consists of various landcover types with agriculture being the dominant land use. Most streams in the area are low gradient with silt or sandy bottoms, resulting in channelization and degradation of aquatic habitat.

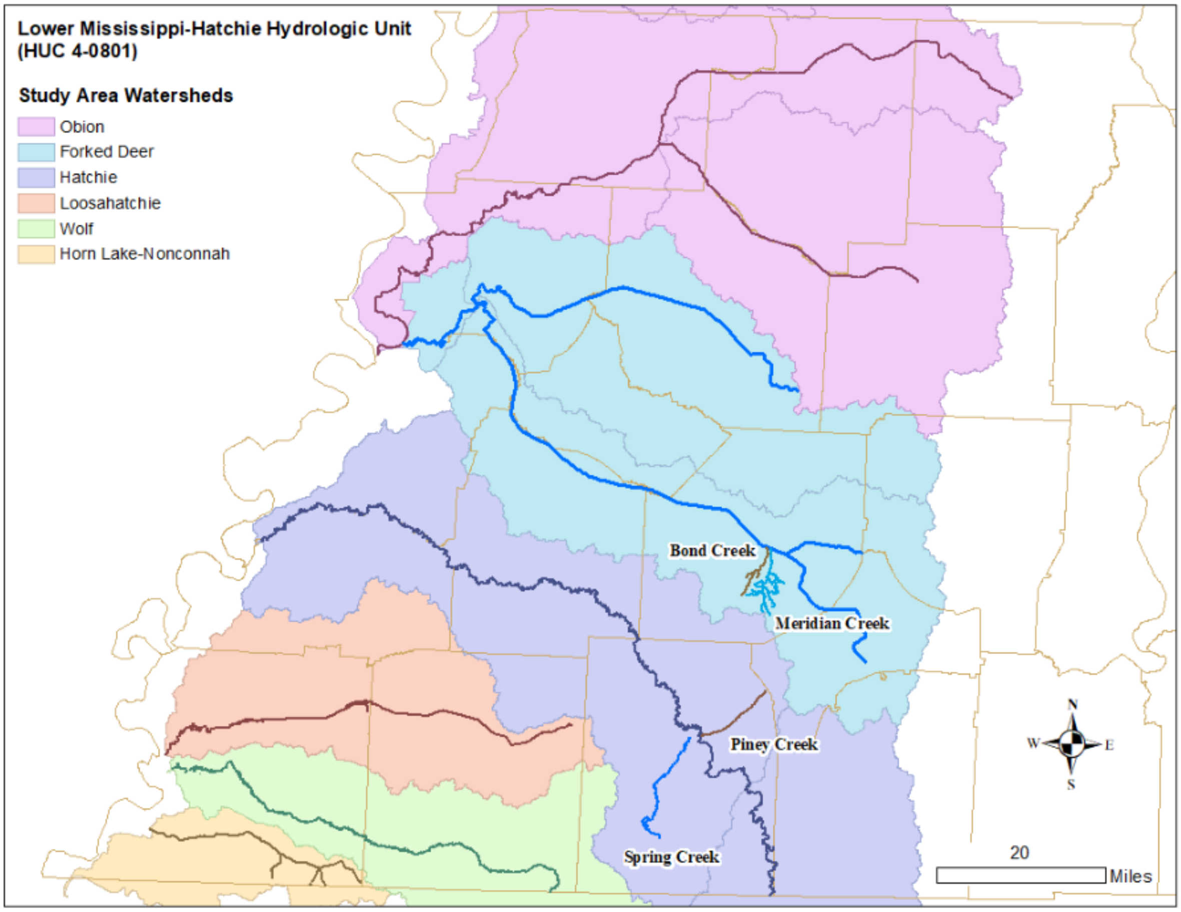


Figure 1 Area of Study

Data Preparation

Since streamflow status can change significantly in each month of the year, monthly datasets are chosen for the analysis to discern these changes in streamflow permanence. All data utilized in this study are shown in Table 1.

Climate predictor variables in this study include annual monthly precipitation (PPT) and maximum (Tmax) and minimum temperature (Tmin) from 2010 to 2020 collected from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) from Oregon State University PRISM Climate Group (*PRISM Climate Group at Oregon State University*, 2014). The PRISM model takes data from close to 13,000 stations for precipitation and 10,000 for temperature and analyzes them to produce climate datasets (Daly et al., 2008).

Evapotranspiration (ET) is the combination of transpiration (water use) by vegetation and evaporation from the soil surface (Senay et al., 2013). For ET data, the Operational Simplified Surface Energy Balance (SSEBop) model data is utilized. SSEBop uses various datasets to model ET, including elevation, temperature correction coefficient, land surface temperature, and air temperature.

For land-cover/land-use data, the National Landcover Datasets (NLCD) from 2011, 2016, and 2019 are utilized. The major landcover and land use classes for the study area are Cultivated Crops (11624800 cells), Deciduous Forest (5617030 cells), Woody Wetlands (3797521), Pasture/Hay (3684800 cells), Mixed Forest (1622866 cells), Developed Open Space (1432130 cells), Evergreen Forest (973899 cells) and Open Water (939525 cells).

Topography data is taken from the NHDPlus dataset which is derived from the USGS 3DEP program with 30m resolution.

Category	**Data source**	**References**
Physiographic Land Use and Land Cover	National Landcover Dataset (NLCD)	Fry et al., 2011; Homer et al., 2007; Homer et al., 2015
Climate (Temperature and Precipitation)	PRISM Climate data	Daly et al. 2017
Evapotranspiration	Operational Simplified Surface Energy Balance (SSEBop)	Senay et al., 2013

Table 1 Datasets to compile predictive variables

All predictor variables have been reprojected to the NAD83 Conus Albers (EPSG 5070) coordinate system, resampled to 30m resolution, and clipped to the Lower Mississippi-Hatchie study area, HUC 4-0801, using the FPCG Tools library's batch resampling tools. The FCPG Tools library can be downloaded through the USGS GitLab repositories (https://code.usgs.gov/StreamStats/FCPGtools) and the relevant code to reproduce this study's results can be found in the appendix.

Methods

Observation Dataset

Having a set of accurate observation points to train machine learning models is an important first step in the prediction and classification of the streamflow intermittency of a river. For the United States, a dedicated streamflow dataset is not readily available. While there are streamflow permanence datasets available, these datasets have regional inconsistencies and most stream gauges have a location bias toward large rivers, making them not feasible for small headwater studies (Jaeger et al., 2021).

Upon inspection of various nationwide streamflow datasets such as USGS Gages, CrowdWater, Stream Tracker, EPA, and FLOwPER (Jaeger et al., 2021), streamflow observations for the Mid-South region of the United States were insufficient for the identification of headwater streams. The HUC4-0801 study area contains some stream gauges in West Tennessee, on the Forked Deer, the Obion, and a few other streams, but since these are larger rivers with permanent streamflow, they were not fit for the study purposes. Therefore, the definition of intermittent and perennial streams from flowlines in the NHDPlus-HR attribute table was utilized to derive an intermittent/perennial stream observations dataset. The NHD-Plus attribute table includes Feature Type (F-type) and Feature Codes that encode detailed information from the USGS topographic maps. Streams with F-type “Stream – Intermittent” are sampled and randomly selected to represent “dry” observations. For “wet” observations, streams with F-Type “Stream –

Perennial" are selected, converted into points, and sampled randomly. The sampling of the point datasets is done using the R statistical language and the Tidyverse package for data wrangling. The resulting training dataset is further verified using various aerial satellite imagery such as Google Earth (https://earth.google.com) and the National Agricultural Imaging Program (NAIP) collection of digital aerial photos (Earth Resources Observation And Science (EROS) Center, 2017) to achieve better accuracy. Figure 2 shows two examples of dry and wet data points used to train the models.

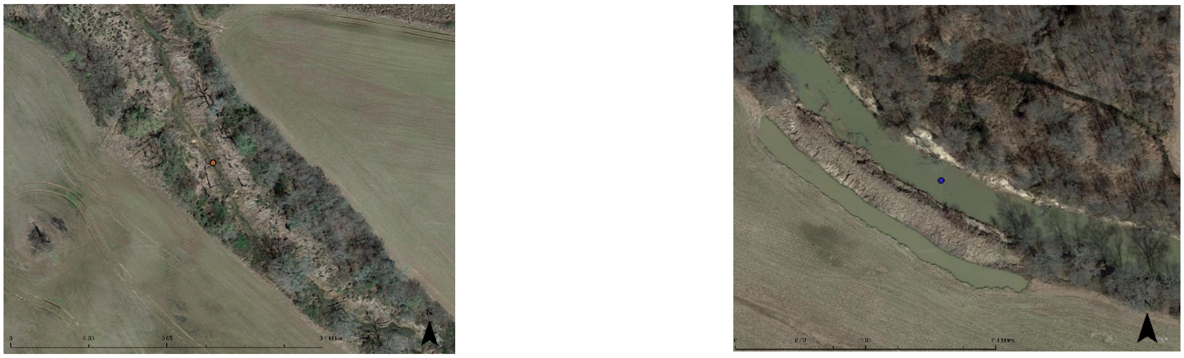


Figure 2 An example of "dry" (left) and "wet" (right) observation points verified by aerial photograph identification.

Model Development

Digital Elevation Model Processing

Data gathered from the USGS national map portal is imported into the ArcGIS Pro software to be preprocessed. Individual raster digital elevation model (DEM) layers are mosaiced to cover the entire study area of HUC4-0801. As the first step of hydrological modeling of the DEM, the Fill tool is used to fill the sinks and depressions in the elevation model (Planchon & Darboux, 2002). Sinks should be filled to achieve the correct delineation of hydrological networks. Leaving sinks and peaks as they come in the raw data can cause discontinuities in the elevation model. For a USGS 30m resolution DEM, between 0.9 to 4.7 percent of the pixels were found to be sinks (Tarboton et al., 1991). Filling sinks in the DEM is an iterative process. Fixing one sink can create other sinks or peaks. The tool iterates until it is unable to identify a sink or a peak anymore. Therefore, to assure the reliability of the data from the beginning, iteration of the peaks and sinks using the fill tool is crucial.

After the fill iteration step, the flow direction raster is calculated from the DEM raster using the eight-direction (D8) model. Flow direction calculation is an important step in hydrological modeling because it has the ability to determine which cells would likely flow into their neighboring cells. Eight available values represent the flow direction of a cell. In an 8-direction pour point model, each cell takes a number indicating to which of the 8 surrounding cells water will flow from the current cell. This process is then followed by the flow accumulation calculation (Jenson & Domingue, 1988).

The final step of preprocessing hydrological data is flow accumulation. In this process, each cell is assigned a number that is equal to the number of cells that flow into that cell. After the flow accumulation raster is calculated, the raster calculator tool is used to define a threshold number of cells that can flow into specific cells to be met or exceeded, creating a stream. For this process, a conditional operator inside the raster calculator is used as

$$Con(\text{“}FlowAccumulation\text{”} > 1000, 1$$

Equation 1 Flow Accumulation Threshold

These steps for calculation of flow accumulation are explained in O’Callaghan & Mark (1984) and approaches to choosing an appropriate threshold for the calculation of stream networks are discussed in Tarboton, et al. (1991). For this study area, and resolution of 30 m per pixel, a threshold of 1000 has been chosen after testing a range of values. In the comparison of aerial satellite imagery with the resulting stream networks, 1000 was the number that matched the most visible headwater streams from the true color imagery. Figure 3 shows the results of this process.

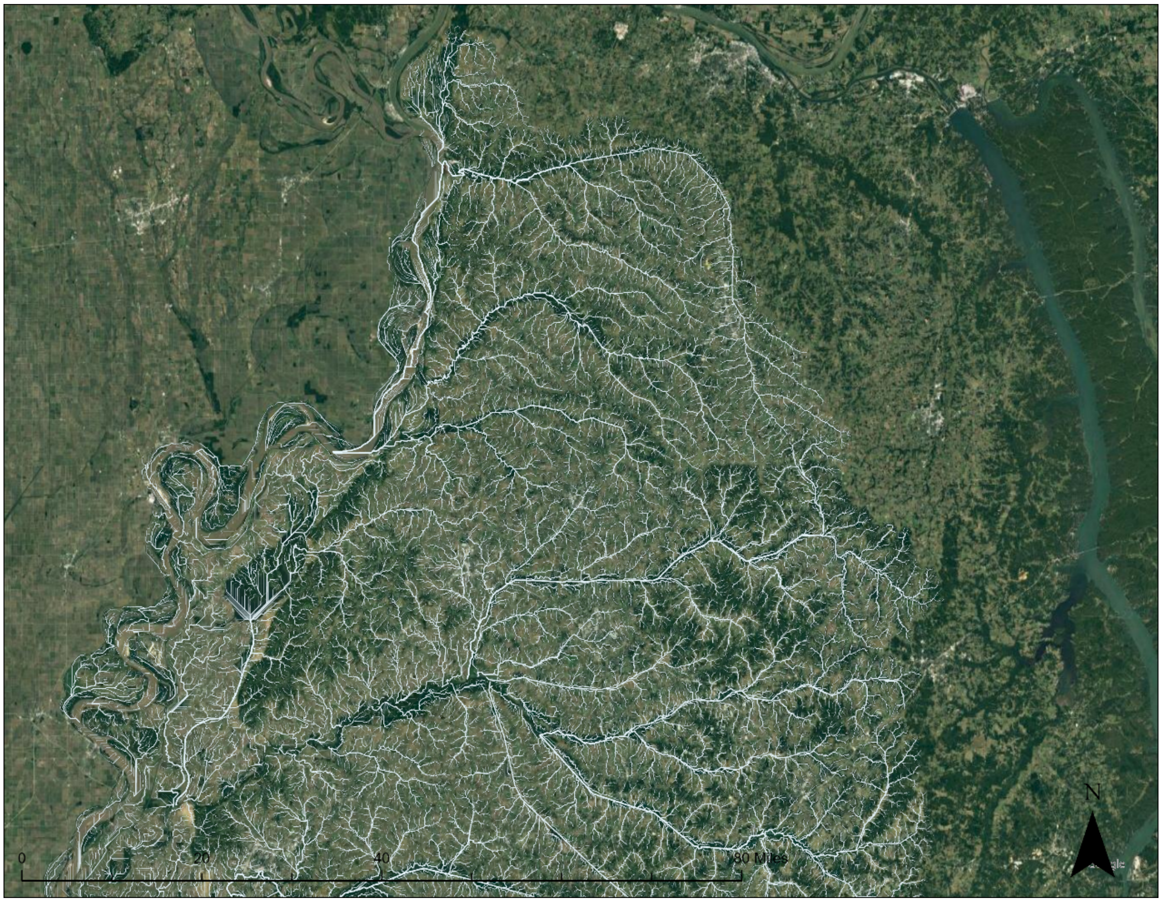


Figure 3 Stream Network Delineation from DEM

Random Forest

Random forest is a machine learning algorithm that can be used for probabilistic predictions or classification tasks. It works by combining multiple decisions from a set of decision trees. Random forest is popular in the remote sensing field due to its accuracy in classification and being resistant to overfit issues (Pal, 2005).

A random forest model was trained by 611 observation points created from the NHD Plus-HR Flowlines data as a point feature class in the HUC4-0801 area. Of these, 321 points were recorded as “dry” observation points and 290 are recorded as “wet”. The resulting streamflow intermittency observation points are exported as a multipoint feature class dataset in ArcGIS pro software to be used as training observation points.

To test the FCPG process in the model, various configurations of predictors are tested for this study. The first training model tested the prediction parameters from PRISM (n=396) and SSEBop (n=132) from 2010 to 2020 and landcover categories from NLCD for 2011, 2016, and 2019 without any FCPG process applied to any of the predictors. For the second model, the FCPG-processed climatic layers, along with landcover categories without the application of the FCPG process, were used. The third training model used FCPG-processed climatic prediction layers, and the binarized, accumulated, and flow-conditioned NLCD landcover layers for 2011, 2016, and 2019. The binarization separates all landcover classes into binary rasters so they can be accumulated using the FCPG Tools accumulation function. These different combinations were used to test whether the FCPG process is a viable treatment for both categorical and continuous variables, as assessed by the model performances. The flow diagram for these steps is shown in Figure 4 below and Figure 5 represents the processes applied to different models.

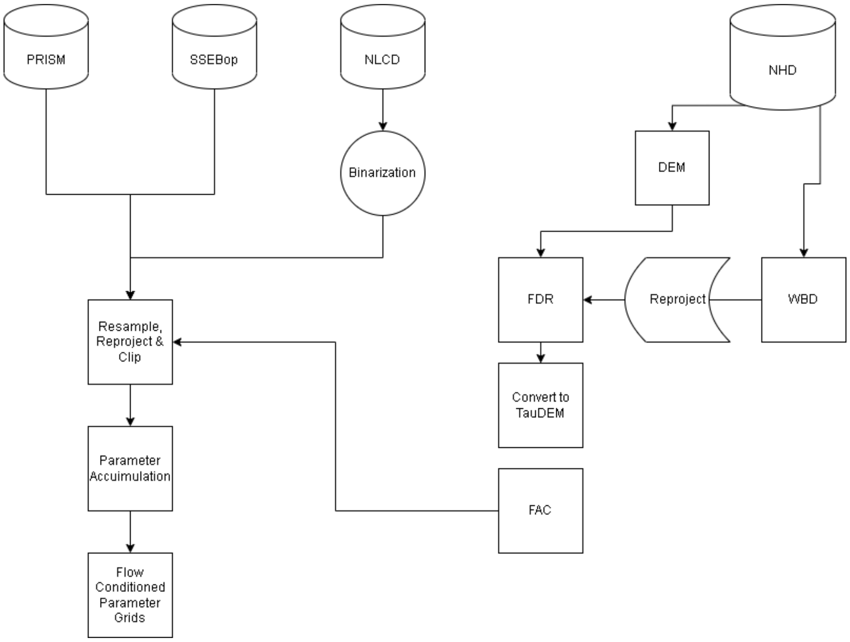


Figure 4 *A flow diagram for the analysis steps*

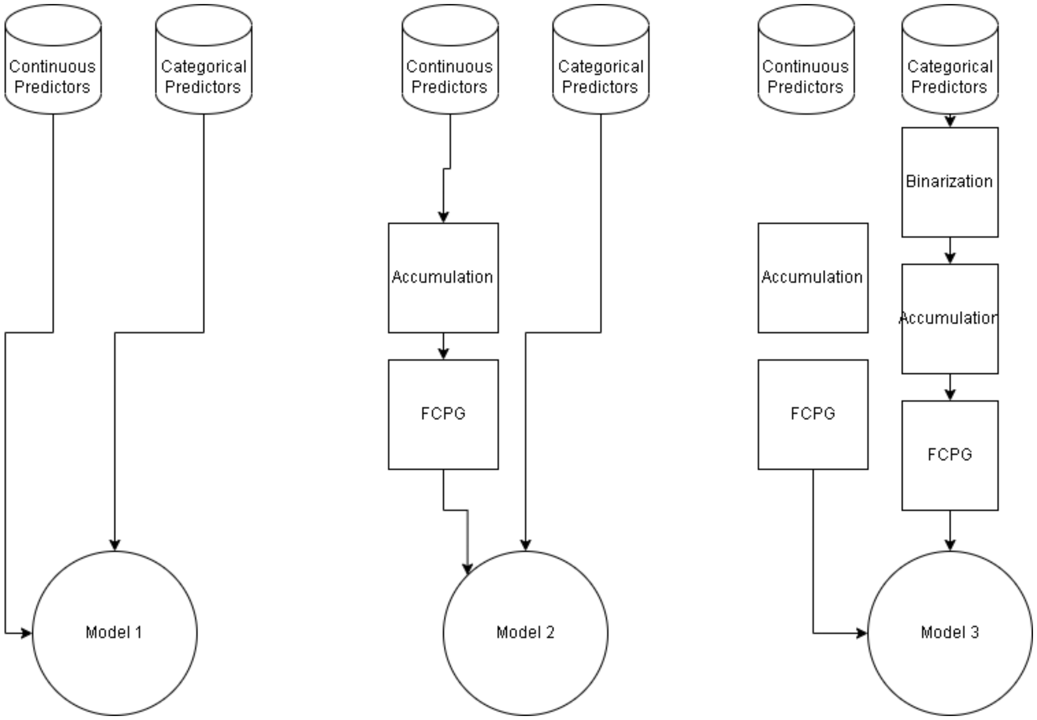


Figure 5 Processes for Prediction Variables

Model Evaluation

One of the most useful aspects of the random forest process is the creation of variable importance measurements for each predictor variable (Strobl et al., 2007). Variable importance is quantified by the rate of decrease in classification error and can suggest the most useful predictor variables. The model performance is determined by the creation of error estimates such as the Mean Squared Error (MSE) Out-of-Bag (OOB) error (Belgiu & Drăguţ, 2016), as well as F1 scores in which both false positives and false negatives are considered during the calculation of the score (Seo et al., 2021), and the Matthews' Correlation Coefficient (MCC) (Matthews, 1975). MSE is the average squared difference between the estimated values and the actual value. Lower MSE scores generally mean better prediction accuracy. The F1 score combines precision and recall values of a random forest model and it can provide further insight into a model to assess if the model is affected by classification issues. The OOB error process separates a certain amount of data to be tested against the model to validate the model's prediction success. For every decision tree, several observations get selected to make up the tree, and the remaining observations become the OOB points. The benefits of using OOB error validation are data leakage prevention, lowered variance, and good performance for small- to medium-sized datasets. In this study, 10% of the data was split for model evaluation, and the sensitivity and accuracy of the model output are reported in the results section.

The sensitivity calculation is shown in Equation 2:

$$TP \,/\, (TP + FN)$$

Equation 2 Sensitivity Calculation

where TP is True Positive, and FN is False Negative.

The accuracy calculation is shown in Equation 3:

$$(TP + TN) \,/\, (TP + FP + TN + FN).$$

Equation 3 Accuracy Calculation

where TN is True Negative and FP is False Positive, and TP and FN are as before.

MCC is the same as the calculation of Phi in statistics and is shown in Equation 4:

$$(TP * TN - FP * FN) \,/\, \sqrt{}(TP + FP)(TP + FN)(TN + FP)(TN + FN)$$

Equation 4 MCC Calculation

For two-class predictions, the model's discrimination ability is presented in a confusion matrix. For a two-class prediction, four classes represent the results (Brown, 2018). The MCC score is only high when all 4 confusion matrix categories (True Positive [TP], True Negative [TN], False Positive [FP], False Negative [FN]) scored a high value (Matthews, 1975).

Flow Conditioned Parameter Grids

FCPG Tools are a Python 3 library and require Python and TauDEM tools (Tarboton, 1997) to be installed. Most of the installation dependencies are managed by a conda environment file that can be downloaded from corresponding USGS repositories. Conda is a tool offered by Anaconda software distribution that allows users to create Python environments and manages package dependencies.

Three hundred and ninety-six prediction rasters from PRISM and 132 prediction rasters from the SSEBop datasets are batch processed to be used in the random forest model using the FCPG Tools (Barnhart et al., 2020) batch resampling function. This function requires a list of input rasters, a flow direction grid, and an output directory to save the processed rasters. The flow direction raster calculated in ArcGIS Pro was used to resample, reproject, and clip all the prediction layers. The process took approximately 7 hours of raw processing time on a Xeon E3 1270 v2 processor with 32 GB ECC error-correcting RAM modules.

Using FCPG Tools' batch accumulation function, previously resampled and clipped parameter rasters are accumulated using the TauDEM D8 tool. The function requires an input list of rasters to be accumulated using a flow direction raster of the area. The process took 10 hours of raw processing time on a Xeon E3 1270 v2 processor with 32 GB ECC RAM modules.

As the final step of the FCPG process, the batch FCPG function from the FCPG Tools library is used to calculate Flow Conditioned Parameter Grids for the predictive layers. The process took 5 hours of raw processing time on a Xeon E3 1270 v2 processor with 32 GB ECC RAM modules.

Results

Random Forest Outcomes

Random forest model performance is evaluated by MSE OOB error rates in addition to F1, MCC, Sensitivity, and Accuracy scores. All three model results are shown in Table 2 below, followed by a section explaining each model's predictor importance.

Number of Trees	**50**	**100**
Model 1 MSE	26.454	24.363
Intermittent	24.419	21.33
Perennial	28.747	27.804
Number of Trees	**50**	**100**
Model 2 MSE	26.869	24.533
Intermittent	24.188	21.26
Perennial	29.636	28.093
Number of Trees	**50**	**100**
Model 3 MSE	29.951	28.456
Intermittent	27.381	24.697
Perennial	32.8	32.557

Table 2 Results of all tested models

Model 1

For the first model tested in this study, climatic predictor variables of precipitation, minimum and maximum temperature, and evapotranspiration as well as NLCD landcover classification layers are fed to the model without being processed through FCPG Tools. NLCD layers are fed to the model as categorical variables. The model resulted in an MSE OOB rate of 26.4 at 50 trees and 24.3 at 100 trees. Prediction for intermittent streams was 24.4 at 50 trees and 21.3 at 100 trees. The perennial prediction rate was 28.7 at 50 trees and 27.8 at 100 trees.

For this model, the most important predictors were the landcover classes for all 3 years, the precipitation from November 2020, and the minimum temperature for April

2020 from the PRISM dataset. The selection of November 2020 was particularly interesting since it was one of the most important variables in the second model as well. The results for the variable importance of Model 1 are shown in Figure 6 and corresponding Table 3.

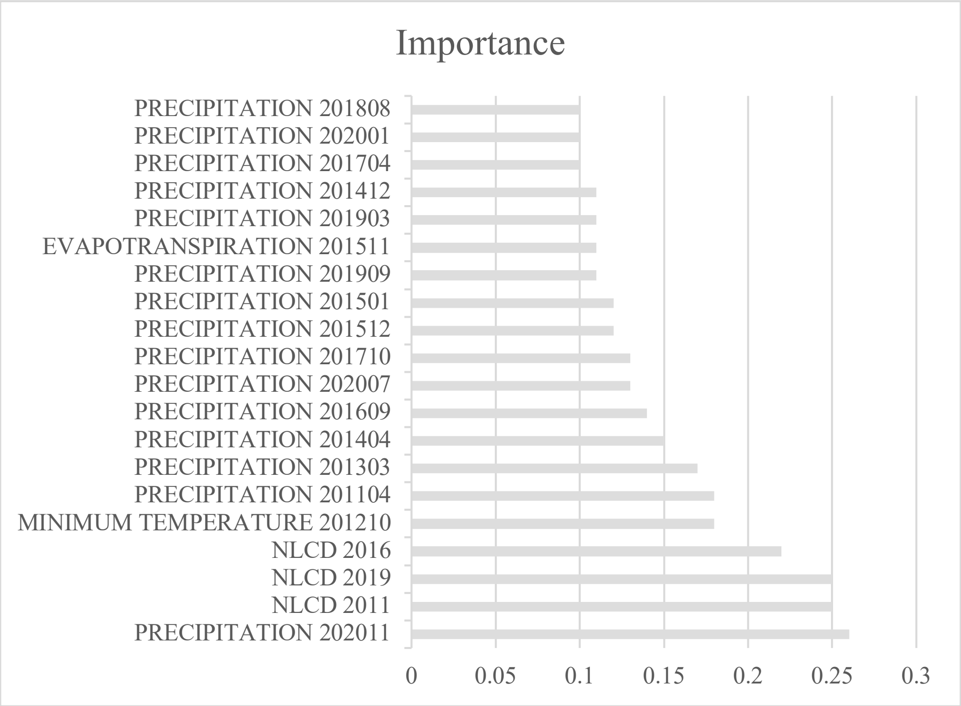


Figure 6 Variable Importance for the Model 1

Top Variable Importance	
Variable	Importance
Precipitation 202011	0.26
NLCD 2011	0.25
NLCD 2019	0.25

Top Variable Importance	
NLCD 2016	0.22
Minimum temperature 201210	0.18
Precipitation 201104	0.18
Precipitation 201303	0.17
Precipitation 201404	0.15
Precipitation 201609	0.14
Precipitation 202007	0.13
Precipitation 201710	0.13
Precipitation 201512	0.12
Precipitation 201501	0.12
Precipitation 201909	0.11
Evapotranspiration 201511	0.11
Precipitation 201903	0.11
Precipitation 201412	0.11
Precipitation 201704	0.1
Precipitation 202001	0.1
Precipitation 201808	0.1

Table 3 Variable Importance Table for Model 1

The model had a 0.74 accuracy for both intermittent and perennial classification predictions and similar sensitivity and F1 scores at 0.74 for intermittent and 0.73 for perennial stream classifications. Classification diagnostics for Model 1 are shown in Table 4 below.

Validation Data: Classification Diagnostics				
Category	F1-Score	MCC	Sensitivity	Accuracy
Intermittent	0.74	0.48	0.74	0.74
Perennial	0.73	0.48	0.73	0.74

Table 4 Classification Diagnostics for Model 1

Model 2

The second model uses the same variables as the first model, but only climatic variables were processed with the FCPG process; the NLCD predictors were fed to the model as categorical data. MSE OOB was 26.8 at 50 trees and 24.5 at 100 trees. The intermittent prediction rate was 24.1 at 50 trees and 21.2 at 100 trees for the second model.

The most important variables were NLCD 2011 landcover type, precipitation for November 2020 which was also one of the most important variables in the first model, NLCD 2016 landcover type, and precipitation for April 2017. 2020 was not a drought year but 2017 was classified as an exceptionally drought year. The result for the variable importance of Model 2 is shown in Figure 7 and corresponding Table 5.

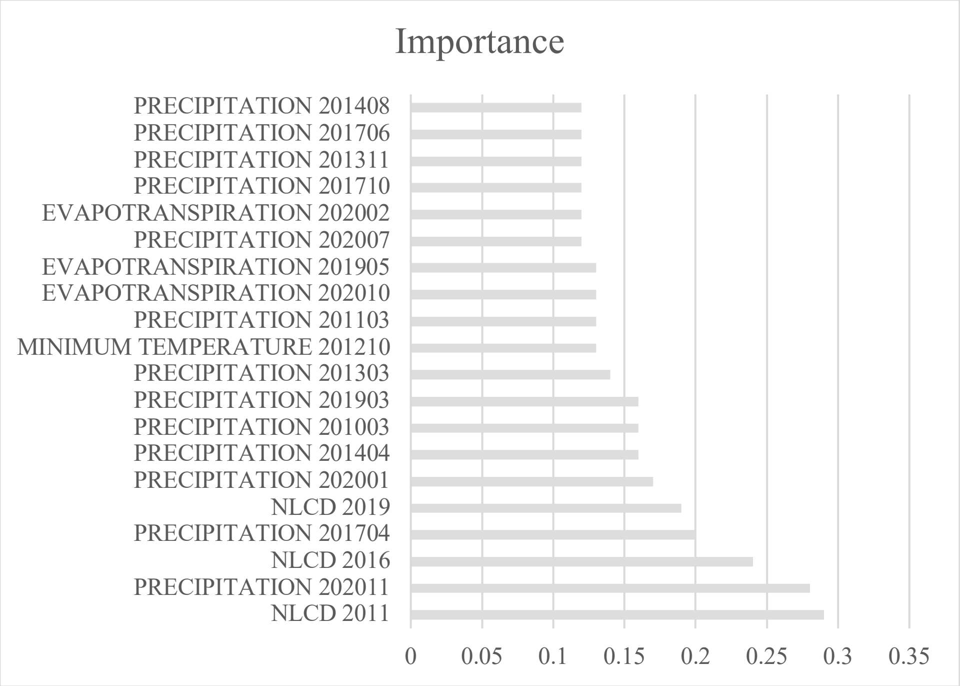


Figure 7 Variable Importance for Model 2

Top Variable Importance	
Variable	Importance
NLCD 2011	0.29
Precipitation 202011	0.28
NLCD 2016	0.24
Precipitation 201704	0.2
NLCD 2019	0.19
Precipitation 202001	0.17
Precipitation 201404	0.16
Precipitation 201003	0.16
Precipitation 201903	0.16
Precipitation 201303	0.14
Minimum temperature 201210	0.13
Precipitation 201103	0.13
Evapotranspiration 202010	0.13
Evapotranspiration 201905	0.13

Top Variable Importance	
Precipitation 202007	0.12
Evapotranspiration 202002	0.12
Precipitation 201710	0.12
Precipitation 201311	0.12
Precipitation 201706	0.12
Precipitation 201408	0.12

Table 5 Variable Importance Table for Model 2

On the second model, the accuracy score for both intermittent and perennial streams increased to 0.84, while the sensitivity score was 0.91 for intermittent streams but 0.74 for perennial streams. The F1 and MCC scores were similar for both intermittent and perennial streams at 0.67 for MCC and 0.8 for F1 scores. Classification diagnostics for Model 2 are shown in Table 6 below.

Validation Data: Classification Diagnostics				
Category	F1-Score	MCC	Sensitivity	Accuracy
Intermittent	0.86	0.67	0.91	0.84
Perennial	0.8	0.67	0.74	0.84

Table 6 Classification Diagnostics for Model 2

Model 3

In the last model, both the climate and landcover predictors were processed using FCPG tools, after binarization of the individual NLCD years instead of using these predictors as categorical data. The binarization process separates all landcover classes into binary rasters so they can be accumulated using the FCPG Tools' accumulation function, i.e., the Forest category produces a binary raster where forested areas are represented with True (1), and non-forested areas are False (0). The addition of FCPG landcover layers resulted in an increase in MSE OOB error to 29.9 at 50 trees and 28.4 at 100 trees. The prediction error rate for intermittent streams was 27.3 at 50 trees and 24.6 at 100 trees. For perennial streams, the prediction error rate was 32.8 at 50 trees and 32.5 at 100 trees.

The most important layers for predicting streamflow permanence in this model were the precipitation layers from April 2017, November 2020, September 2014, October 2017, July 2015, July 2020, and April 2014 followed by the evapotranspiration predictor from October 2011. Results are presented in Figure 8 and shown in Table 7. 2011 and 2017 were particularly dry years and this may have contributed to April 2017 being the most important variable in this model as well as the high importance of evapotranspiration for October 2011.

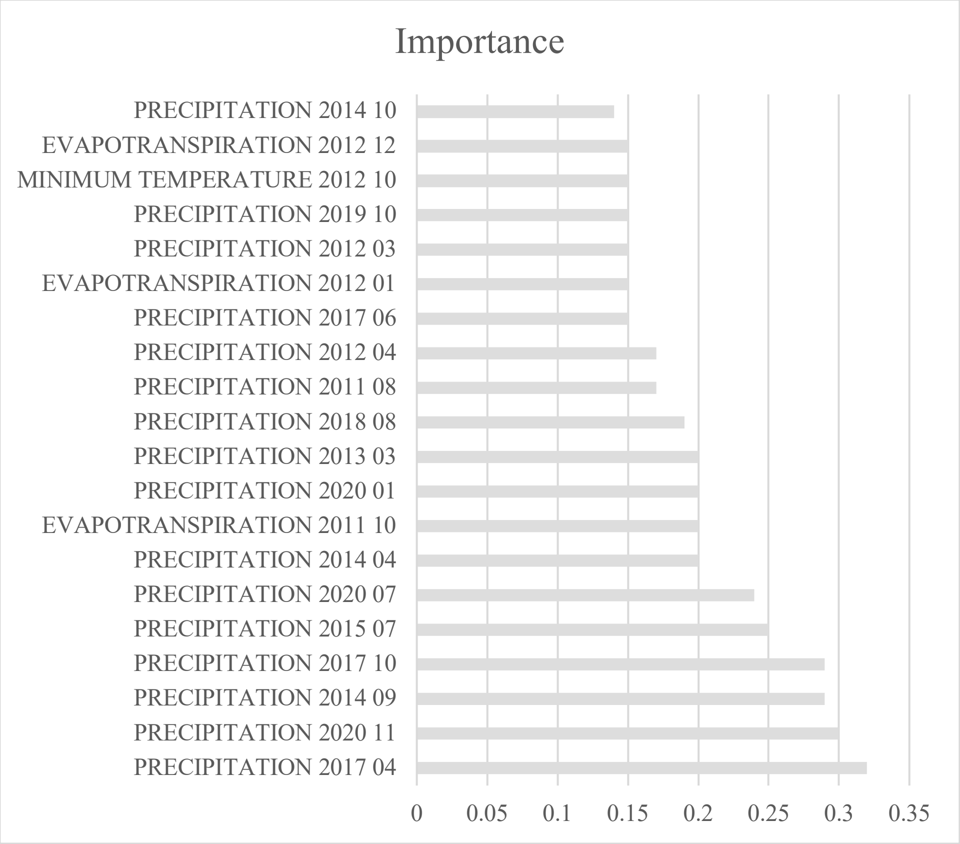


Figure 8 Variable Importance for Model 2

Top Variable Importance

Variable	Importance
Precipitation 2017 04	0.32
Precipitation 2020 11	0.3
Precipitation 2014 09	0.29
Precipitation 2017 10	0.29
Precipitation 2015 07	0.25
Precipitation 2020 07	0.24
Precipitation 2014 04	0.2
Evapotranspiration 2011 10	0.2
Precipitation 2020 01	0.2
Precipitation 2013 03	0.2

Top Variable Importance	
Precipitation 2018 08	0.19
Precipitation 2011 08	0.17
Precipitation 2012 04	0.17
Precipitation 2017 06	0.15
Evapotranspiration 2012 01	0.15
Precipitation 2012 03	0.15
Precipitation 2019 10	0.15
Minimum temperature 2012 10	0.15
Evapotranspiration 2012 12	0.15
Precipitation 2014 10	0.14

Table 7 Variable Importance Table for Model 3

On the third model, accuracy decreased to 0.72 for both categories, sensitivity was 0.7 and 0.75 for intermittent and perennial streams, respectively. The MCC scores were low at 0.44 for both categories and F1 scores had a relatively big gap between them at 0.75 for intermittent and 0.68 for perennial streams. A low MCC score suggested a decrease in prediction accuracy for this model which could be related to low elevation variation in the study area. Classification diagnostics for Model 3 are shown in Table 8 below.

Validation Data: Classification Diagnostics				
Category	F1-Score	MCC	Sensitivity	Accuracy
Intermittent	0.75	0.44	0.7	0.72
Perennial	0.68	0.44	0.75	0.72

Table 8 Classification Diagnostics for Model 3

Overall, Model 1 performed best due to consistent high accuracy and sensitivity scores for both intermittent and perennial data points (Table 2). This model used no FCPG processing in setting up the predictor variables and both continuous predictors and categorical variables are fed to the model only after clipping, reprojecting, and resampling of these variables.

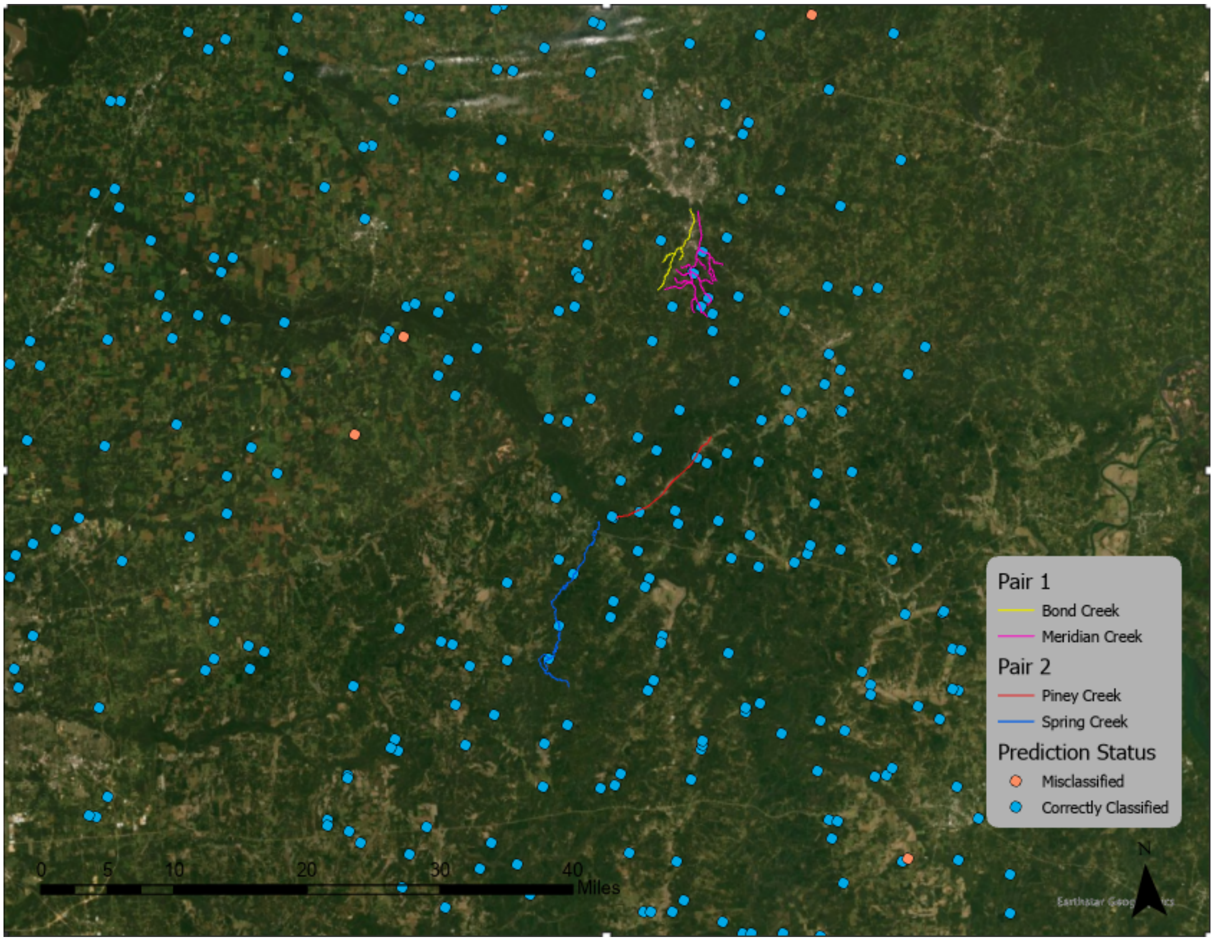


Figure 9 Training Data Prediction Results Blue Dots are Correct Classifications; Orange Dots are Misclassifications from the first model

Validation of the 10% of training data split off for this purpose resulted in 24 correct and 9 incorrect intermittent classifications, and 21 correct and 7 incorrect perennial classifications. Prediction results for the entire training observation data are shown in Figure 9.

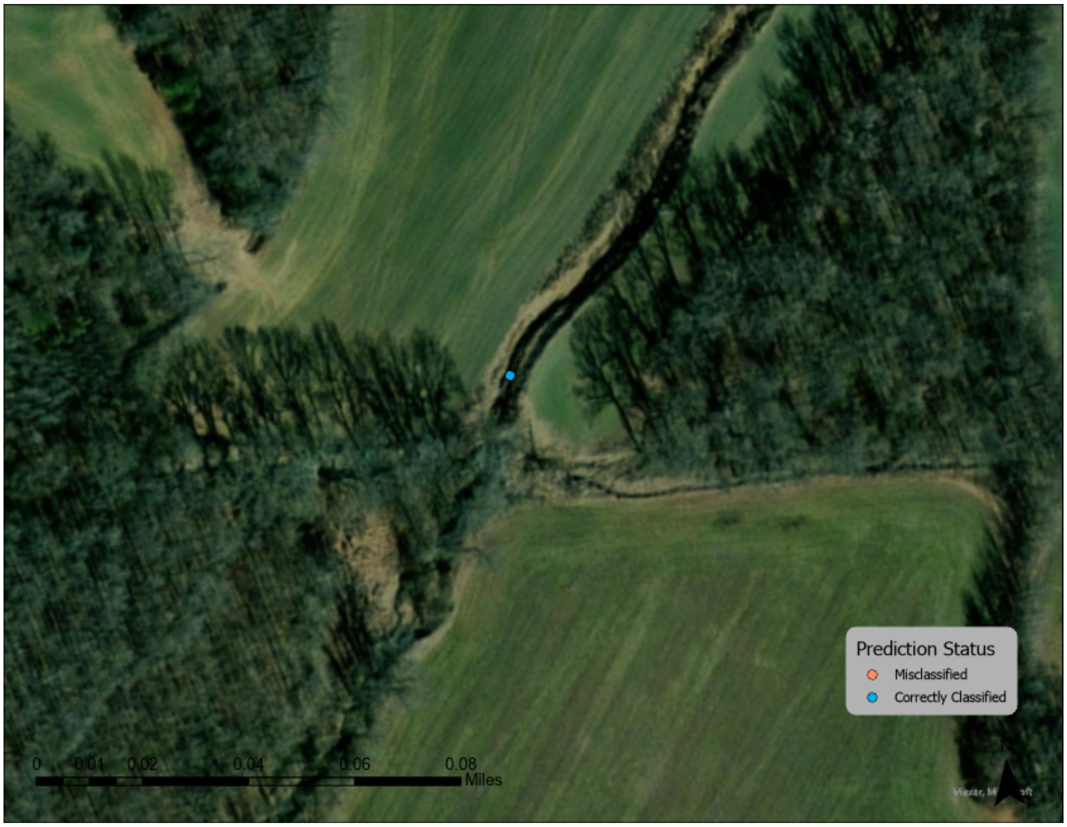


Figure 10 Detail view of a correctly classified stream: an intermittent stream classified as intermittent

Figure 10 shows a 2nd order stream that is classified as an intermittent stream by the random forest classifier. Agricultural activities might have contributed to the classification of this observation as an intermittent stream.

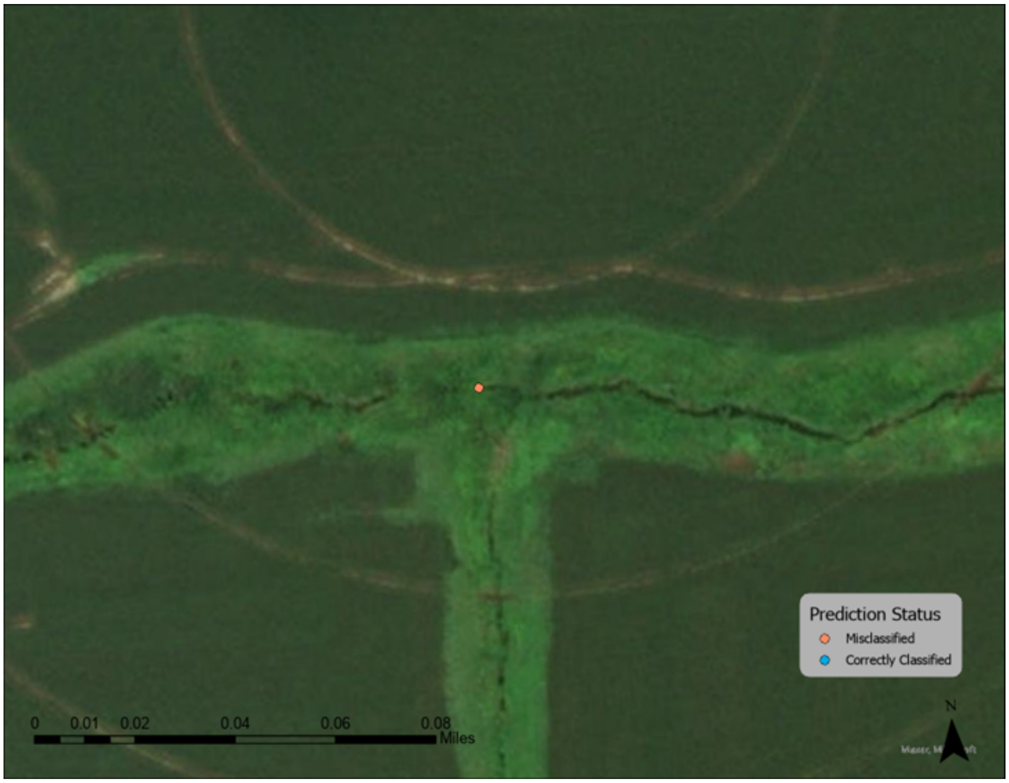


Figure 11 Detail view of an incorrectly classified stream, classified as perennial, but an intermittent stream

Figure 11 shows a stream that is classified as a perennial stream, in disagreement with the imagery data. Validation of the NHD classification of this stream reach using satellite imagery shows small agricultural drainage ditches flowing through crop fields served by a center pivot irrigation system (Figure 11). Lighter green areas are vegetated margins separating the different crop fields. This stream is a small-order tributary or even a headwater, and as such would be expected to be dry during the late summer and early fall months. Such areas were predicted to pose a challenge to the model, in part because the resolution of the input data may be too coarse to correctly classify flow conditions in these small streams. Further investigation of the riparian corridors of these streams using vegetation indices derived from fine-resolution, multiband imagery, may provide additional information to the model to increase the accuracy of its predictions in small headwaters.

After deliberation of the validation data on the three training models, Model 1, which is the best performing model, is used to predict the entire study area. This model is considered the best performing model due to the consistency in the prediction of intermittent and perennial streams. Two predictions are made to visually inspect the result of the model output, one as a raster for the entire study area and another one as a feature class for the flow lines.

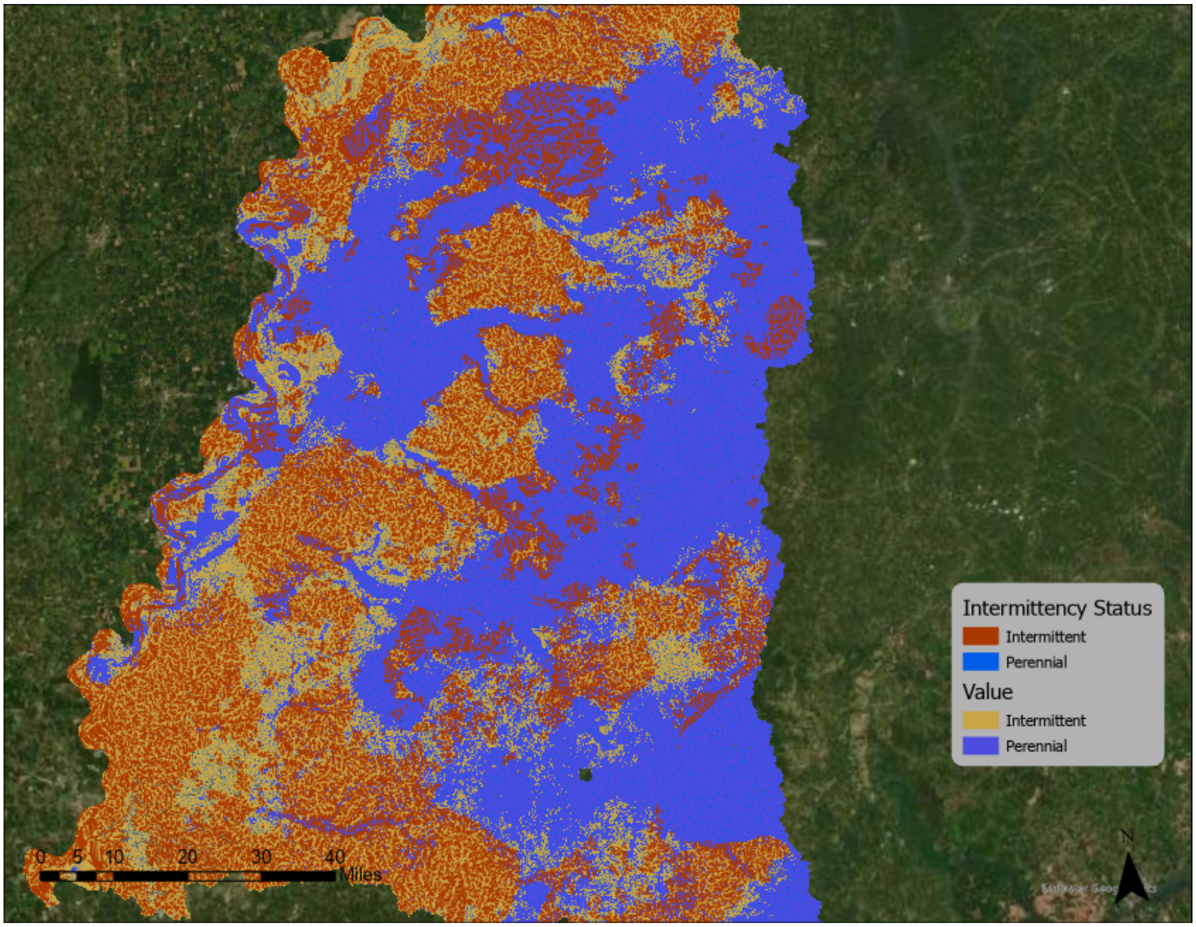


Figure 12 Final Prediction Raster and Flow Lines Feature Class for the Study Area

The predictions made by the model should be taken as areas with a high probability of either having intermittent or perennial streams (Figure 12) rather than specific predictions per stream, mostly due to the lack of input variable resolution. Purple areas on the map are perennial prediction areas and the orange color represents areas where the model predicts a higher expectancy of streams with intermittent flow. Superimposed are the feature class predictions for the same area, with blue lines representing perennial streams and red lines the predictions for streams with a higher possibility of going intermittent.

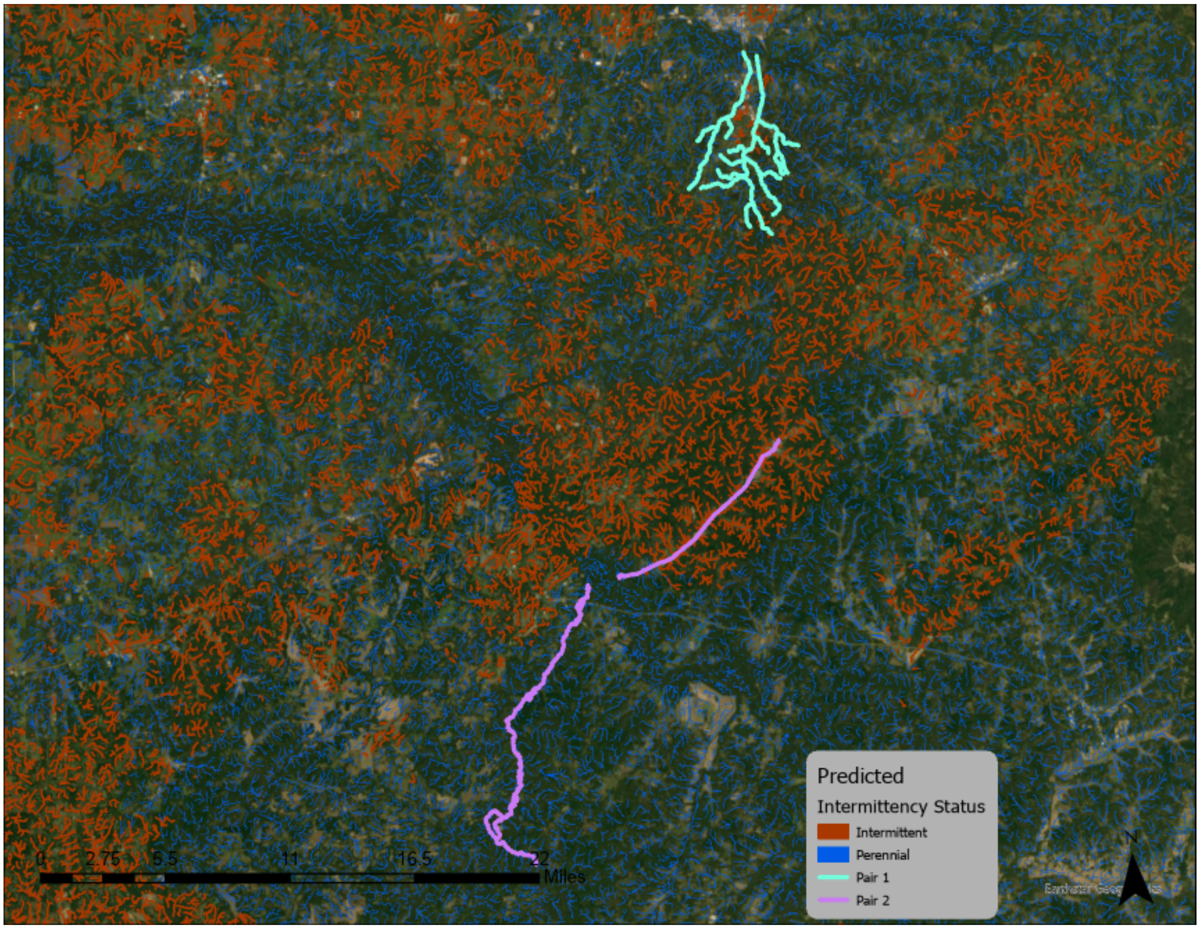


Figure 13 Predictions of Stream Network Results for the Study Area

Random forest model results presented as a flow line feature class showing the prediction of stream permanence for individual streams in the study area are shown in Figure 13. These results indicate which areas tend to contain environmental effects that could result in more intermittent streams and which areas have more perennial streams, although many headwaters are predicted as perennial. This can be attributed to training data used in creating the model, and in-situ verification of this data can provide answers to why the model predicted some headwaters as perennial. The areas with intermittent predictions could be affected by agricultural land uses and other activities as well as the elevation and slope of these areas. In general, smaller-order streams (upstream reaches) are more likely to be intermittent than are higher-order streams (downstream reaches) because, in addition to receiving water input from precipitation, surface runoff, and baseflow from the subsurface, downstream areas receive input from all upstream reaches. Because of this, these higher-order streams tend to be perennial unless there is extreme water loss through evaporation, or more likely, loss due to infiltration to the subsurface via a sandy streambed. This could be investigated using the model by incorporating variables into it that reflect other water budget inputs and outputs to and from the basin.

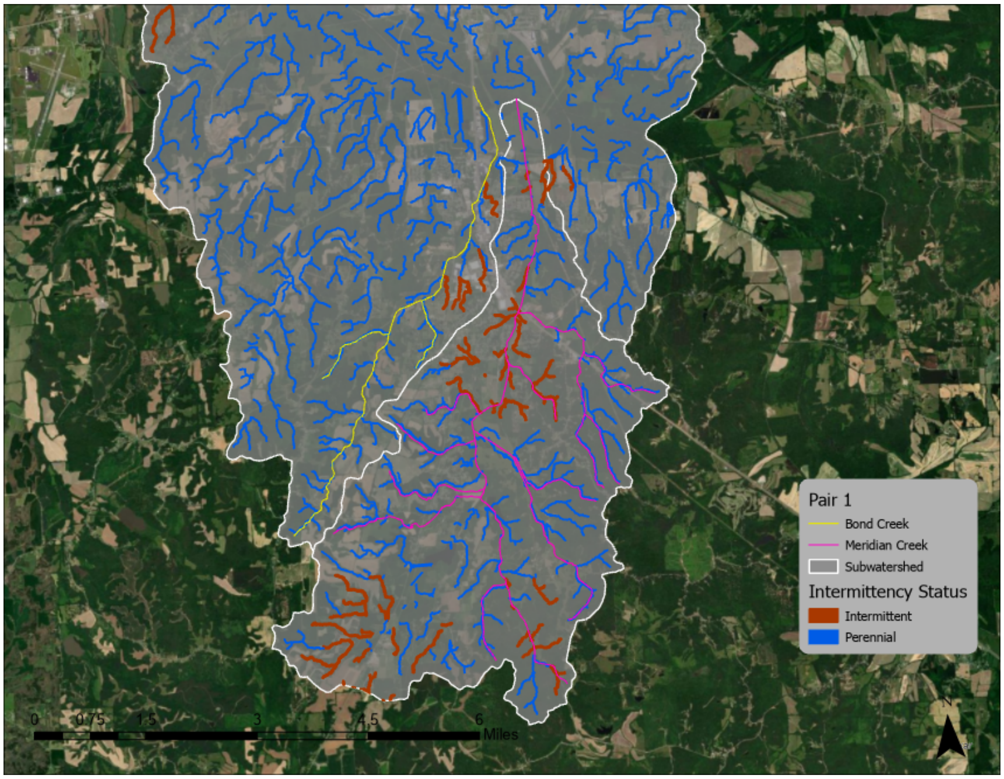


Figure 14 Model Predictions for Pair 1, Meridian, and Bond Creek

In addition to the predictions for the entire study area, two pairs of streams were selected from the TDEC watershed health assessments to test the model's ability to discriminate between a stream with a headwater lake and one without. Classification results suggest that most of Bond Creek are perennial streams except for some tributaries, while Meridian Creek has parts that are classified as intermittent streams that can dry out in the dry months of the year. These predicted intermittent sections of Meridian Creek are both in upstream reaches above the headwater lakes as well as further downstream below them, where the flow would be expected to be perennial. Bond Creek is a small stream with no headwater lakes, and so perennial flow here was not expected. Figure 15 shows this stream pair in the TDEC dataset, where results of their *in-situ* verifications have been displayed. Bond Creek was assessed as not supporting of all its uses (red lines, Figure 15) which is an indication that this stream experiences intermittent flow. In contrast, Meridian Creek is fully supporting of all its uses downstream of the lakes, but not supporting in the headwater reaches above the lakes. This means that the headwater lakes supply enough flow that downstream reaches are likely perennial, while above the lake there may be intermittent flow. For this test pair, it is not conclusive to say that intermittency is predicted correctly according to the existence of a headwater lake. Thus, the presence of headwater lakes in the streamflow permanence prediction was not a reliable classifier in this study's scope. Visual assessment of the model output suggests an overprediction of permanence. More specific prediction and classification studies using distance-based model inputs and distance-based accumulation methods offered by FCPG Tools may be needed to identify and classify headwater lakes and how they affect the streamflow of first and second-order streams.

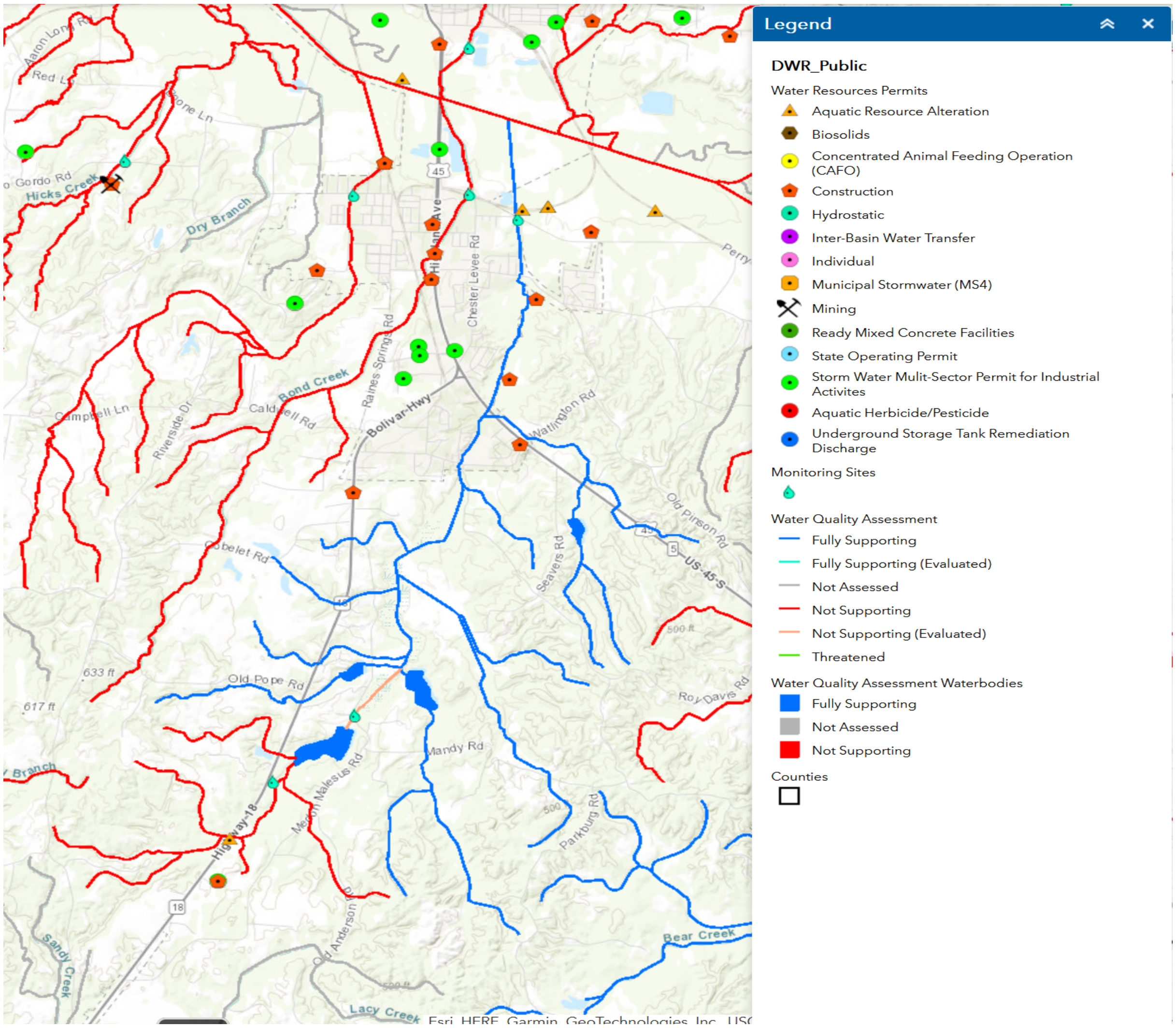


Figure 15 TDEC Assessments of Pair 1 Streams

In Figure 16, the second test pair is shown for visual comparison of streamflow permanence predictions. Piney Creek stream reaches are mostly shown as intermittent while Spring Creek reaches are mostly predicted as perennial. These results are what would be expected given the *in-situ* assessments from the TDEC dataset (Figure 17).

While the headwater lakes in Spring Creek were not assessed water bodies, they are present throughout the watershed. The assessed stream reaches of Spring Creek are all classified as fully supporting of all their uses, and therefore would be likely to have perennial flow. Only the mainstem of Spring Creek was assessed (red line, Figure 17), but this was classified as not supporting of all its uses. Given its size and lack of headwater lakes, this stream is likely to experience intermittent flow. The model predictions and in situ assessments align for the second test pair, but given the results from Pair 1, it is not conclusive that the model is making the correct prediction for the expected reasons. Further examination of both the model results and the TDEC dataset along with the identification of more headwater lakes is needed to fully assess the credibility of the model predictions of flow permanence in smaller streams below a lake.

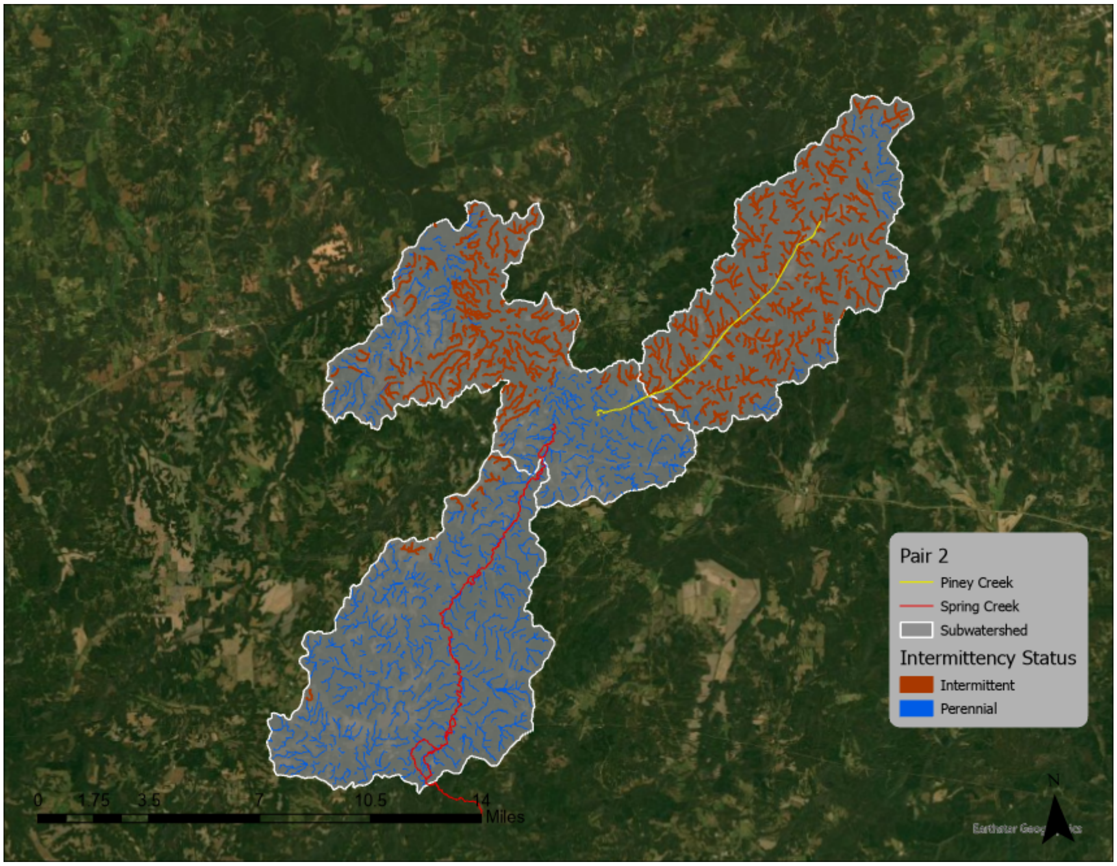


Figure 16 Test Pair 2, Piney Creek and Spring Creek with corresponding subwatersheds

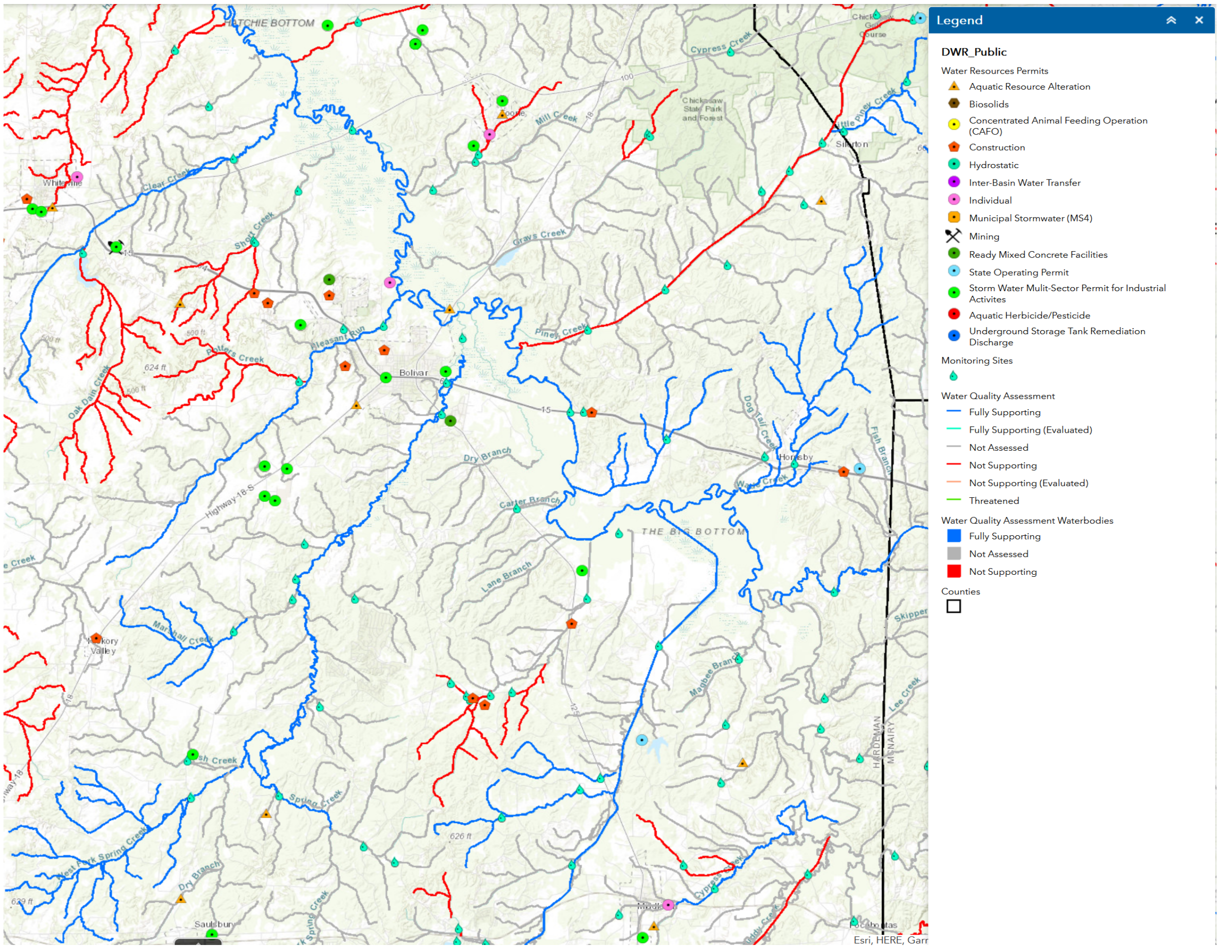


Figure 17 Test Pair 2 in the TDEC dataset

In Figure 18, the Hatchie River is shown. The streams of the Lower Hatchie River watershed are marked as intermittent by the model even though the mainstem of the Lower Hatchie is perennial. This could be attributed to the coarse resolution of the predictor variables, or that pure climate and land cover input variables are insufficient to predict the flow accumulation that provides the perennial flow in downstream reaches of larger (higher-order) streams. This could also be due to the lack of proper observation and training data for the model, although the HHD Plus HR dataset generally has intermittent and perennial flow classifications on major rivers. Figure 19 shows a zoomed-in figure of these areas. While the surrounding small tributaries of the Lower Hatchie River likely are intermittent, their accumulated flow combined with the perennial flow from upstream would make the Lower Hatchie mainstem perennial as well. This result, combined with the large amount of predicted area of streamflow permanence in the eastern portion of the study area, which are the headwaters for the rivers here (Figure 12) warrants further investigation.

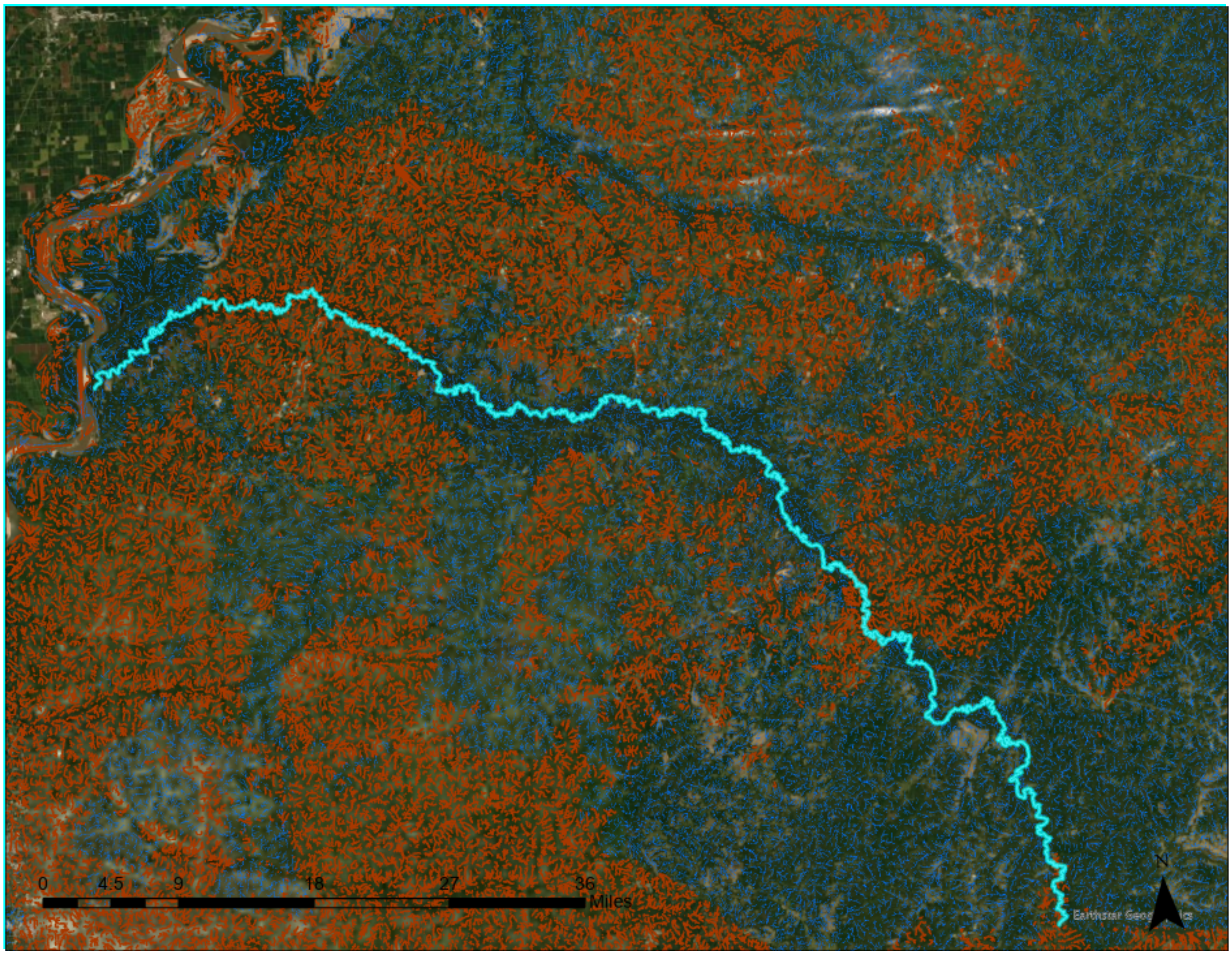


Figure 18 Model predictions in the Hatchie River watershed

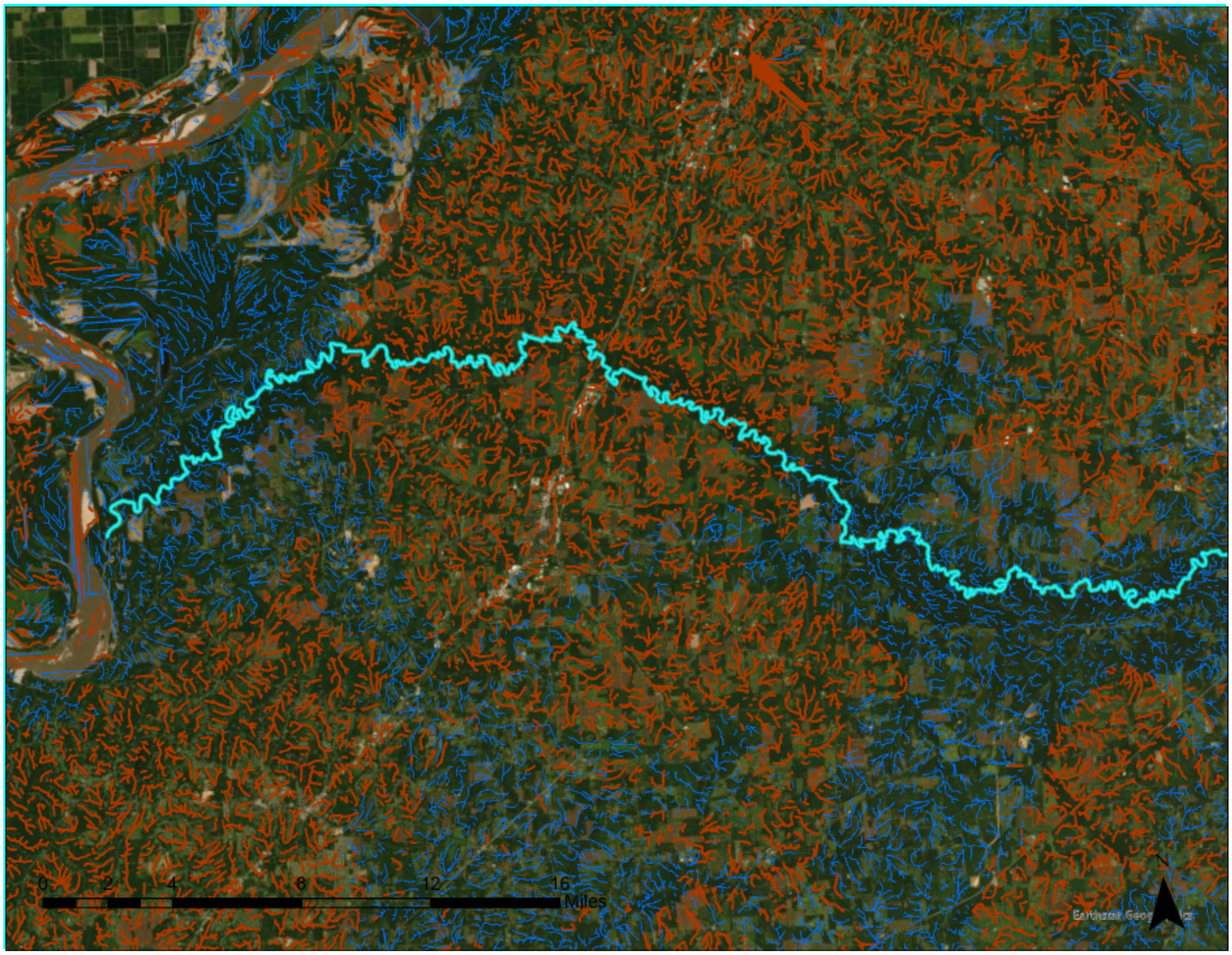


Figure 19 Intermittent Parts of Lower Hatchie River

Discussion and Further Steps

In this study, streamflow permanence is analyzed using a random forest model with various predictor data inputs that include precipitation, minimum and maximum temperature, and evapotranspiration on ungauged headwater streams. Previous studies about streamflow intermittency classification and prediction use both advanced machine learning techniques and deep learning applications for the prediction of streamflow status (Bolanos et al., 2016; Jiang & Wang, 2019; Li et al., 2019; Peng et al., 2020; Razavi & Coulibaly, 2013). While some of these studies are built to predict the intermittency status of streams and hydrological networks from scratch (Ha et al., 2021; Sando & Blasch, 2015), others attempt further feature engineering by selecting the most important variables to improve model performance and efficiency for better analysis and predictions (Shen et al., 2022).

When considering the streamflow, the total water budget in a basin is affected by precipitation and evapotranspiration as tested in this study, but many other variables affect the streamflow including conveyance evaporation, imported water, stream outflow, infiltration, and runoff. In this study, soil and permeability data are mostly excluded except for the NLCD land use/ land cover data, but these variables also can be incorporated into predictive models to improve model performance and make more reliable predictions.

For a comprehensive list of catchment attributes, Razavi & Coulibaly (2013) serve as a reference point to define these potential variables and datasets to add to further streamflow permanence studies that observe flow status in ungauged streams. These variables include area, elevation, porous aquifers, river network density, catchment drainage area, leaf area index, percentage of woody vegetation in the catchment, plant water-holding capacity, and soil moisture deficit. In addition to these variables, the model settings can also be adjusted. For example, tree size in the random forest model can be increased for better prediction results, given there is enough processing power available. Random forest uses the results of many decision trees to create a final prediction. All of the random forest models tested in this study have been built on both 50 and 100 trees, but other studies have found that increasing the tree sizes also increased prediction accuracy when using a random forest algorithm (Peng et al., 2020). In addition, to adjusting tree size, training a model on gauged networks to determine key factors in flow density can also improve feature engineering in machine learning applications. Processing predictor parameters in conjunction with the elevation of corresponding watersheds is yet another approach that could be utilized in streamflow permanence prediction studies.

FCPG Tools (Barnhart et al., 2020) are produced to process the accumulation of predictor parameters using elevation and flow direction data. In this study, this FCPG accumulation process is applied to test the potential benefits of replacing raw predictive data with flow-conditioned variables in a random forest model for West Tennessee tributaries. Out-Of-Bag Error and other metrics are used to assess the performance of each classification model, as well as observation and verification of the headwater pairs from West Tennessee watersheds. Previous studies successfully incorporated random forest models with a 20-30% prediction MSE OOB error for streamflow status (Jaeger et al., 2019; Li et al., 2019). This study's results are comparable to those with a 24% MSE OOB error rate for the best performing model, where no predictor variables were processed using FCPG Tools, and 28% for the worst performing model, in which both continuous and categorical variables were accumulated using FCPG Tools. The results of these models indicate that in this study's scope, the FCPG process did not make a significant difference in the classification of streamflow intermittency status. This may be attributed to low elevation variability in this study's area of interest. Further investigation of the reasoning behind these results can be conducted to understand what else can be done differently to improve these predictions.

In addition, there was a weak relationship in this study between streamflow status and the existence of headwater lakes. While the existence of headwater lakes does affect the streamflow permanence of streams, additional predictors may be needed for a more precise classification than those predictor parameters used here. The ArcGIS Random Forest Classification and Regression tool offers the addition of distance-based training variables to be considered for the model and FCPG Tools also provide distance-based accumulation methods that can be used for testing the effects of the spatial relation of headwater lakes and streamflow permanence.

Meridian and Bond Creeks (Figure 14) did not match the prediction of streamflow with their corresponding headwater existence status as would be expected given the TDEC water quality assessments (Figure 15). However, the predictions for Piney and Spring Creeks (Figure 16) did (Figure 17). Further investigation is needed to understand the relation between headwater streams that have lakes upstream and headwater streams without headwater lakes in order to predict the status of the corresponding streams. Applying a wider selection of prediction variables may also significantly improve predictions of the correlation between streamflow permanence and headwater lakes. For example, Jaeger et al., (2021) indicated that many factors contribute to the prediction of streamflow status, and observing these variables can assist in predicting the presence or absence of surface water flow. In this study, the variables that had the greatest impact on streamflow status across all models were the precipitation predictors. Variable importance tables showed the occurrence of the precipitation variable inputs in the months following each July in all three models. During the dry season, baseflow supplies the streamflow This could mean that model's variable importance in choosing the summer growing season as an important predictor of baseflow as the streamflow source. Additionally, west Tennessee has periodic dry years and, in some years, just dry summers. This may indicate that some streams only become intermittent in very dry years. These results suggest that while the model is capturing the correlation between seasonal changes and streamflow, it may benefit from the addition of distance-based accumulation processes to better route the flow inputs through the network.

Streamflow prediction has many contributing factors in addition to precipitation. Minimum and maximum temperature and evapotranspiration variables were both tested in this study but showed lower importance compared to precipitation. Further research incorporating other variables can assist in predicting streamflow and the assessment of stream health status. These analyses of streamflow predictions can serve in understanding the health status of basins. Water budget and water balance simulations can be used as an additional input to such prediction studies by using other variables for water that flows into the basin and water that exits from the basin. Also, soil type and permeability are important factors when calculating streamflow status and there are datasets such as STATS2GO (Soil Survey Staff, 2022) that can provide additional training variables for such predictive models. Accurate predictions of streamflow status could also provide critical insight into studying the implications of climate change in hydrological networks, although only to the extent that climate variables are considered important to the model.

This study found there is a need to fill some data gaps in the recently released NHDPlus HR v2 hydrography dataset. NHDPlus HR v2 contains missing attribute data for feature types of streams in this study's area of interest. In specific regions, there are no intermittency data; these streams are not assessed and not marked as perennial, intermittent, or ephemeral. This study's results could fill this data gap for these areas until the NHDPlus HR v2 dataset is assessed and marked in these areas. Figure 20 shows the areas with missing intermittency information.

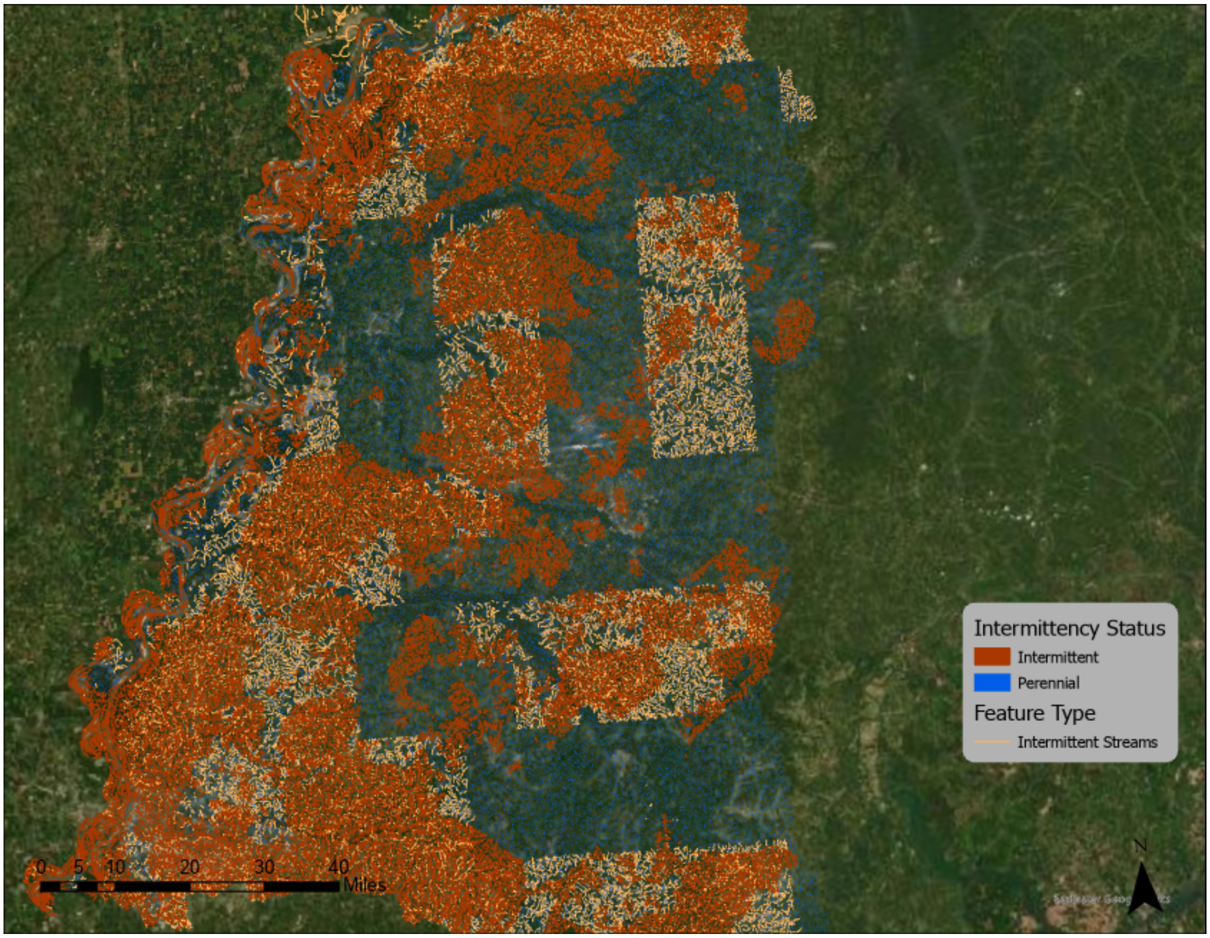


Figure 20 Streams with missing intermittency data with prediction results of the random forest model

In this study, a training dataset is derived from the NHDPlus HR v2 dataset using intermittent/perennial attributes of the streams in West Tennessee basins to test if the environmental variables such as precipitation and evapotranspiration can be useful in predicting stream impairment. In addition, two pairs of streams are tested to see if predictions of the models can distinguish between streams with headwater lakes and streams without headwater lakes. Results show acceptable prediction accuracy for streamflow status but not for permanence below headwater lakes. Due to the limitation of this study, in-situ verification was not a viable option. In order to better assess the model performance, in-situ verification of the training data may be needed. In addition, FCPG Tools accumulation functions are utilized, and the results did not improve using FCPG tools. The FCPG study currently in progress has not been published yet, and given the results of the new investigation, the results of this study may be reassessed once the FCPG paper is released. As a result of these analyses, it is understood that the NHD Plus HR dataset has data gaps, and further attempts and studies are needed to fill this gap. Furthermore, a nationwide streamflow permanence dataset is needed for further investigation of streamflow permanence. Results of these models also indicated that precipitation is a strong predictor of streamflow, similar to elevation data. Even though it may lower the prediction accuracy, it may be useful to isolate precipitation and conduct similar tests with other predictors to understand other contributing variables in the prediction of streamflow status without the dominant effects of precipitation. In conclusion, this study expects that these results can be useful for state agencies, environmental/conservation groups, and other entities to further understand streams that may dry at any time throughout the year and which environmental factors could be causing impairment of streams. These replicable and scalable methods can offer an understanding of spatio-temporal dynamics that influence streamflow permanence in West Tennessee streams.

Budget

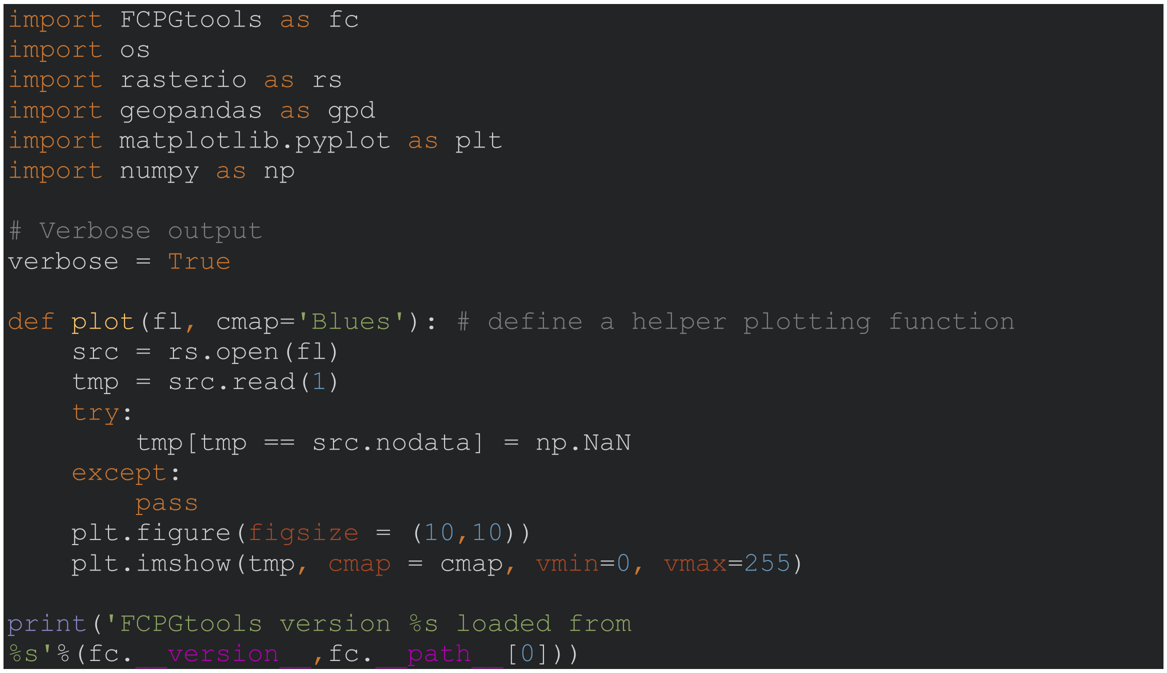
This thesis work is supported by a USGS104b grant from The Tennessee Water Resources Center (TNWRRC).

Acknowledgments

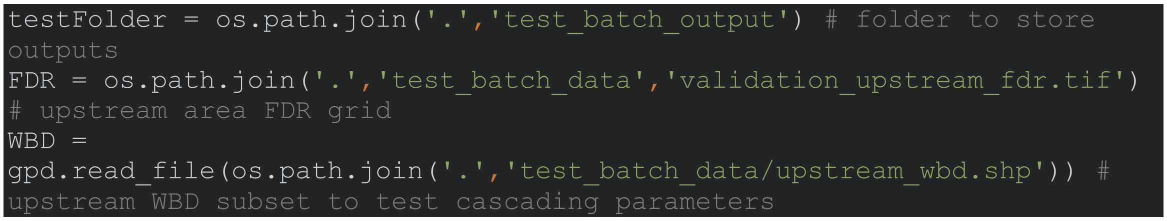
Flow-Grid Conditioned Grid Tools are part of the public domain because they are produced by employees of the U.S. Government. As the authors of the study, we are thankful to Theodore B. Barnhart, Roy Sando, and other USGS employees who produced FCPG tools and provided support to understand their usage.

Appendix

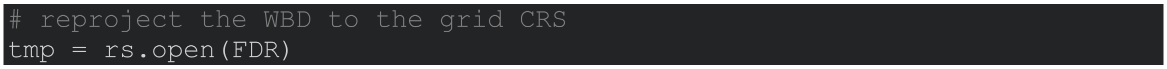
The codes used in this study can be found in the appendix. Further information and more code examples can be found in the official Flow Conditioned Parameter Grid GitHub or GitLab repositories hosted by the USGS



```python
import FCPGtools as fc
import os
import rasterio as rs
import geopandas as gpd
import matplotlib.pyplot as plt
import numpy as np

# Verbose output
verbose = True

def plot(fl, cmap='Blues'): # define a helper plotting function
    src = rs.open(fl)
    tmp = src.read(1)
    try:
        tmp[tmp == src.nodata] = np.NaN
    except:
        pass
    plt.figure(figsize = (10,10))
    plt.imshow(tmp, cmap = cmap, vmin=0, vmax=255)

print('FCPGtools version %s loaded from
%s'%(fc.__version__,fc.__path__[0]))
```



```python
testFolder = os.path.join('.','test_batch_output') # folder to store
outputs
FDR = os.path.join('.','test_batch_data','validation_upstream_fdr.tif')
# upstream area FDR grid
WBD =
gpd.read_file(os.path.join('.','test_batch_data/upstream_wbd.shp')) #
upstream WBD subset to test cascading parameters
```



```python
# reproject the WBD to the grid CRS
tmp = rs.open(FDR)
```

```
dstCRS = tmp.crs.to_proj4()

WBD.to_crs(crs=dstCRS, inplace=True)
```

```
# define output paths for TauDEM flow direction
FDRTau = os.path.join(testFolder,'FDRtau.tif')

# define output paths for TauDEM flow accumulation
upstreamFAC = os.path.join(testFolder,'upstreamFAC.tif') # path for the
output FAC grid.

# reclassify ESRI flow directions to TauDEM
fc.tauDrainDir(FDR, FDRTau, verbose=verbose)

# calculate flow accumulation
fc.tauFlowAccum(FDRTau,upstreamFAC, cores=4, verbose=verbose)
```

```
# define output filepaths for NLCD
NLCDAccum = os.path.join(testFolder,'NLCDAccum.tif')
NLCDFCPG = os.path.join(testFolder,'NLCDFCPG.tif')
# NLCD layer import
nlcd11_11 = os.path.join(testFolder, 'nlcd11binarized',
'nlcd_2011_land_cover_l48_20210604rprj12.tif')
```

```
# calculate single parameter accumulation for NLCD
nlcd_accum = fc.accumulateParam(nlcd11_11, FDRTau, NLCDAccum)
finalNLCDFCPG = fc.make_fcpg(NLCDAccum, upstreamFAC, NLCDFCPG, verbose
= verbose)
```

```
# assign input folder path
nlcd_batch_folder = os.path.join('.','test_batch_data', 'nlcdbin')

# emtpy list to populate with parameter filenames
nlcd_batch_list = []

for filename in os.listdir(nlcd_batch_folder):
    name, ext = os.path.splitext(filename)
    if ext == '.tif': #change filetype here
        infile = os.path.join(nlcd_batch_folder, filename)
        nlcd_batch_list.append(infile)
```

```
nlcdbin = os.path.join('.','test_batch_data', 'nlcdbin')
```

```
nlcdlist = []

for filename in os.listdir(nlcdbin):
    infile = os.path.join(nlcdbin, filename)
    nlcdlist.append(infile)
```

```
accumParams = fc.accumulateParam_batch(nlcdlist, FDRTau, testFolder
,cores = 4, verbose = verbose)
batch_accumout_folder =
os.path.join('.','test_batch_output\ssebop_accum_out') # folder to
store outputs
accumParams = []

for bfilename in os.listdir(batch_accumout_folder):
    binfile = os.path.join(batch_accumout_folder, bfilename)
    accumParams.append(binfile)
# batch FCPG output folder path
batch_FCPG_folder =
os.path.join('.','test_batch_output\ssebop_FCPG_out') # folder to store
outputs

# batch FCPG

upstream_cpgs =
fc.make_fcpg_batch(accumParams,upstreamFAC,batch_FCPG_folder, verbose =
verbose)
usLCbinary = fc.cat2bin(LCupstream, testFolder, verbose=verbose)
```

References

Barnhart, T. B., Sando, R., Siefken, S. A., McCarthy, P. M., & Rea, A. H. (2020). *Flow-Conditioned Parameter Grid Tools: U.S. Geological Survey Software Release*. https://doi.org/10.5066/P9W8UZ47

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bhamjee, R., & Lindsay, J. B. (2011). Ephemeral stream sensor design using state loggers. *Hydrology and Earth System Sciences*, *15*(3), 1009–1021. https://doi.org/10.5194/hess-15-1009-2011

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Blyth, K., & Rodda, J. C. (1973). A stream length study. *Water Resources Research*, *9*(5), 1454–1461. https://doi.org/10.1029/WR009i005p01454

Bolanos, S., Stiff, D., Brisco, B., & Pietroniro, A. (2016). Operational Surface Water Detection and Monitoring Using Radarsat 2. *Remote Sensing*, *8*(4), 285. https://doi.org/10.3390/rs8040285

Brown, J. B. (2018). Classifiers and their Metrics Quantified. *Molecular Informatics*, *37*(1–2), 1700127. https://doi.org/10.1002/minf.201700127

Costigan, K. H., Jaeger, K. L., Goss, C. W., Fritz, K. M., & Goebel, P. C. (2016). Understanding controls on flow permanence in intermittent rivers to aid ecological research: Integrating meteorology, geology and land cover. *Ecohydrology*, *9*(7), 1141–1153. https://doi.org/10.1002/eco.1712

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., & Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, *28*(15), 2031–2064. https://doi.org/10.1002/joc.1688

Day, D. G. (1978). Drainage density changes during rainfall. *Earth Surface Processes*, *3*(3), 319–326. https://doi.org/10.1002/esp.3290030310

Earth Resources Observation And Science (EROS) Center. (2017). *National Agriculture Imagery Program (NAIP)* [Tiff]. U.S. Geological Survey. https://doi.org/10.5066/F7QN651G

Eberts, S. M., Woodside, M. D., Landers, M. N., & Wagner, C. R. (2019). Monitoring the pulse of our Nation’s rivers and streams—The U.S. Geological Survey streamgaging network. In *Monitoring the pulse of our Nation's rivers and streams—The U.S. Geological Survey streamgaging network* (USGS Numbered Series No. 2018–3081; Fact Sheet, Vols. 2018–3081). U.S. Geological Survey. https://doi.org/10.3133/fs20183081

Godsey, S. E., & Kirchner, J. W. (2014). Dynamic, discontinuous stream networks: Hydrologically driven variations in active drainage density, flowing channels and stream order. *Hydrological Processes*, *28*(23), 5791–5803. https://doi.org/10.1002/hyp.10310

Goulsbra, C., Evans, M., & Lindsay, J. (2014). Temporary streams in a peatland catchment: Pattern, timing, and controls on stream network expansion and contraction. *Earth Surface Processes and Landforms*, *39*(6), 790–803. https://doi.org/10.1002/esp.3533

Ha, S., Liu, D., & Mu, L. (2021). Prediction of Yangtze River streamflow based on deep learning neural network with El Niño–Southern Oscillation. *Scientific Reports*, *11*(1), 11738. https://doi.org/10.1038/s41598-021-90964-3

Hayes, M. M., Miller, S. N., & Murphy, M. A. (2014). High-resolution landcover classification using Random Forest. *Remote Sensing Letters*, *5*(2), 112–121. https://doi.org/10.1080/2150704X.2014.882526

Jaeger, K. L., Hafen, K. C., Dunham, J. B., Fritz, K. M., Kampf, S. K., Barnhart, T. B., Kaiser, K. E., Sando, R., Johnson, S. L., McShane, R. R., & Dunn, S. B. (2021). Beyond Streamflow: Call for a National Data Repository of Streamflow Presence for Streams and Rivers in the United States. *Water*, *13*(12), 1627. https://doi.org/10.3390/w13121627

Jaeger, K. L., Sando, R., McShane, R. R., Dunham, J. B., Hockman-Wert, D. P., Kaiser, K. E., Hafen, K., Risley, J. C., & Blasch, K. W. (2019). Probability of Streamflow Permanence Model (PROSPER): A spatially continuous model of annual streamflow permanence throughout the Pacific Northwest. *Journal of Hydrology X*, *2*, 100005. https://doi.org/10.1016/j.hydroa.2018.100005

Jensen, C. K., McGuire, K. J., & Prince, P. S. (2017). Headwater stream length dynamics across four physiographic provinces of the Appalachian Highlands. *Hydrological Processes*, *31*(19), 3350–3363. https://doi.org/10.1002/hyp.11259

Jenson, S. K., & Domingue, J. O. (1988). Extracting topographic structure from digital elevation data for geographic information-system analysis. In *Photogrammetric Engineering and Remote Sensing* (Vol. 54, Issue 11, p. 15931600).

Jiang, D., & Wang, K. (2019). The Role of Satellite-Based Remote Sensing in Improving Simulated Streamflow: A Review. *Water*, *11*(8), 1615. https://doi.org/10.3390/w11081615

Johansen, K., Phinn, S., & Witte, C. (2010). Mapping of riparian zone attributes using discrete return LiDAR, QuickBird and SPOT-5 imagery: Assessing accuracy and costs. *Remote Sensing of Environment*, *114*(11), 2679–2691. https://doi.org/10.1016/j.rse.2010.06.004

Leopold, L. B., Wolman, M. G., Miller, J. P., Wohl, E., & Wohl, E. E. (1964). *Fluvial Processes in Geomorphology*. Courier Dover Publications.

Li, X., Sha, J., & Wang, Z.-L. (2019). Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. *Hydrological Sciences Journal*, *64*(15), 1857–1866. https://doi.org/10.1080/02626667.2019.1680846

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure, 405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Matthews, K., Eddy, M., Jones, P., Southerland, M., Morgan, B., & Rogers, G. (2015). Tennessee Integrated Assessment of Watershed Health. *EPA-841-R-15-002,* *Washington, DC: RTI International, Research Triangle Park, NC, and Versar for US Environmental Protection Agency, Healthy Watersheds Program*, 103.

Michez, A., Piégay, H., Lejeune, P., & Claessens, H. (2017). Multi-temporal monitoring of a regional riparian buffer network (>12,000 km) with LiDAR and photogrammetric point clouds. *Journal of Environmental Management*, *202*(Pt 2), 424–436. https://doi.org/10.1016/j.jenvman.2017.02.034

O'Callaghan, J. F., & Mark, D. M. (1984). The extraction of drainage networks from digital elevation data. *Computer Vision, Graphics, and Image Processing*, *28*(3), 323–344. https://doi.org/10.1016/S0734-189X(84)80011-0

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition* (pp. 154–168). Springer. https://doi.org/10.1007/978-3-642-31537-4_13

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Pate, A. A., Segura, C., & Bladon, K. D. (2020). Streamflow permanence in headwater streams across four geomorphic provinces in Northern California. *Hydrological Processes*, *34*(23), 4487–4504. https://doi.org/10.1002/hyp.13889

Peng, F., Wen, J., Zhang, Y., & Jin, J. (2020). Monthly Streamflow Prediction Based on Random Forest Algorithm and Phase Space Reconstruction Theory. *Journal of*

Physics: Conference Series, *1637*(1), 012091. https://doi.org/10.1088/1742-6596/1637/1/012091

Pham, L. T., Luo, L., & Finley, A. O. (2020). *Evaluation of Random Forest for short-term daily streamflow forecast in rainfall and snowmelt driven watersheds* [Preprint]. Catchment hydrology/Modelling approaches. https://doi.org/10.5194/hess-2020-305

Planchon, O., & Darboux, F. (2002). A fast, simple and versatile algorithm to fill the depressions of digital elevation models. *CATENA*, *46*(2), 159–176. https://doi.org/10.1016/S0341-8162(01)00164-3

PRISM Climate Group at Oregon State University. (2014). https://prism.oregonstate.edu/terms/

Razavi, T., & Coulibaly, P. (2013). Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. *Journal of Hydrologic Engineering*, *18*, 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

Sando, R., & Blasch, K. W. (2015). Predicting alpine headwater stream intermittency: A case study in the northern Rocky Mountains. *Ecohydrology & Hydrobiology*, *15*(2), 68–80. https://doi.org/10.1016/j.ecohyd.2015.04.002

Senay, G. B., Bohms, S., Singh, R. K., Gowda, P. H., Velpuri, N. M., Alemu, H., & Verdin, J. P. (2013). Operational Evapotranspiration Mapping Using Remote Sensing and Weather Datasets: A New Parameterization for the SSEB Approach.

JAWRA Journal of the American Water Resources Association, *49*(3), 577–591. https://doi.org/10.1111/jawr.12057

Seo, S., Kim, Y., Han, H.-J., Son, W. C., Hong, Z.-Y., Sohn, I., Shim, J., & Hwang, C. (2021). Predicting Successes and Failures of Clinical Trials With Outer Product–Based Convolutional Neural Network. *Frontiers in Pharmacology*, *12*, 670670. https://doi.org/10.3389/fphar.2021.670670

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E. H., & Karssenberg, D. (2022). Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. *Computers & Geosciences*, *159*, 105019. https://doi.org/10.1016/j.cageo.2021.105019

Soil Survey Staff. (2022). *Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey.*

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. https://doi.org/10.1021/ci034160g

Tarboton, D. G., Bras, R. L., & Rodriguez-Iturbe, I. (1991). On the extraction of channel networks from digital elevation data. *Hydrological Processes*, *5*(1), 81–100. https://doi.org/10.1002/hyp.3360050107

Turner, D. S., & Richter, H. E. (2011). Wet/Dry Mapping: Using Citizen Scientists to Monitor the Extent of Perennial Surface Flow in Dryland Regions. *Environmental Management*, *47*(3), 497–505. https://doi.org/10.1007/s00267-010-9607-y

Wasser, L., Chasmer, L., Day, R., & Taylor, A. (2015). Quantifying land use effects on forested riparian buffer vegetation structure using LiDAR data. *Ecosphere*, *6*(1), art10. https://doi.org/10.1890/ES14-00204.1

Wasser, L., Day, R., Chasmer, L., & Taylor, A. (2013). Influence of Vegetation Structure on Lidar-derived Canopy Height and Fractional Cover in Forested Riparian Buffers During Leaf-Off and Leaf-On Conditions. *PLOS ONE*, *8*(1), e54776. https://doi.org/10.1371/journal.pone.0054776

Winter, T. C. (2007). The Role of Ground Water in Generating Streamflow in Headwater Areas and in Maintaining Base Flow1. *JAWRA Journal of the American Water Resources Association*, *43*(1), 15–25. https://doi.org/10.1111/j.1752-1688.2007.00003.x

Wohl, E. (2017). The significance of small streams. *Frontiers of Earth Science*, *11*(3), 447–456. https://doi.org/10.1007/s11707-017-0647-y