Data supplement to EPA 840-B-23001

Development and Evaluation of the Beta Streamflow Duration Assessment Method for the Northeast and Southeast

Report EPA 840-R-23003

November 2023 Addendum

The Beta SDAM for the Northeast and Southeast published in April 2023 contained an error in the calculation of *at least intermittent* classifications. This error has now been corrected in the web application. The error caused some sites to be identified as *at least intermittent* that should have been identified as *less than perennial* or *needs more information* instead. The *less than perennial* classification occurs when an *intermittent* or *ephemeral* classification cannot be made with high confidence, but a *perennial* classification can be ruled out. Out of the 522 site visits used to develop the beta SDAM for the Northeast, 32 were classified as *at least intermittent*, but 17 of those should have been identified as *less than perennial*. Out of the 394 site visits in the Southeast used to develop the beta SDAM, 20 were classified as *at least intermittent, but* 7 of those should have been identified as *less than perennial*. No site visits would have changed from *at least intermittent* to *needs more information* in the Northeast or Southeast dataset, but it is a possible outcome of the methods.

The data analysis steps described in the April 2023 report accurately describe the process followed to develop the beta SDAMs for the Northeast and Southeast; however, the results for *at least intermittent* are over-represented, and the occurrences of *less than perennial* and potential for *needs more information* classifications are wholly missing.

Applying the corrected calculation of *at least intermittent*, the final beta method for the Northeast correctly classified 73% of site visits among three classes (*perennial* vs. *intermittent* vs. *ephemeral*), while 90% of site visits were classified correctly between two classes (*ephemeral* vs. *at least intermittent*). Similarly, the corrected final beta method for the Southeast correctly classified 73% of site visits among the three classes, with 90% of site visits classified correctly between two classes.

Data Supplement to EPA-843-B-23001

Development and Evaluation of the Beta Streamflow Duration Assessment Methods for the Northeast and Southeast

April 2023
EPA-843-R-23003

United States
Environmental Protection
Agency

Development and Evaluation of the Beta Streamflow Duration Assessment Methods for the Northeast and Southeast

Data analysis supplement

Shannon Gross
RTI International
Fort Collins, CO 80528

Michele Eddy
RTI International
Research Triangle Park, NC 27709

Ken M. Fritz
Office of Research and Development
Cincinnati, OH 45268

Brian Topping
Office of Wetlands, Oceans, and Watersheds
Washington, DC 20004

Tracie-Lynn Nadeau
Office of Wetlands, Oceans, and Watersheds
Portland, OR 97205

Rachel Fertik Edgerton
Office of Wetlands, Oceans, and Watersheds
Washington, DC 20004

Raphael D. Mazor
Southern California Coastal Water Research Project
Costa Mesa, CA 92626

Kristina Nicholas, ORISE Fellow
Office of Wetlands, Oceans, and Watersheds
Washington, DC 20004

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. This report fulfills EPA QA requirements. The research for the data was conducted under the Office of Water approved Quality Assurance Project Plan "Streamflow Duration Assessment Method (SDAM) development in the Northeast Southeast (NESE)" which was given an ORD ID of J-WECD-0033408-QP-1-0. Funding was provided to Ecosystem Planning and Restoration, RTI, and SCCWRP under contracts EP-C-17-001 and 68HERC21D0008 for data management and analysis, and TRC companies, PG Environmental and Normandeau under EP-C-16-006 and 68HERC22D0002 for data collection.

Disclaimer

Any mention of trade names, manufacturers or products does not imply an endorsement by the United States Government or the U.S. Environmental Protection Agency. EPA and its employees do not endorse any commercial products, services, or enterprises.

Suggested citation:

Gross, S., Eddy M., Fritz, K.M., Nadeau, T.-L., Topping, B., Fertik Edgerton, R., Mazor, R.D., and Nicholas, K. 2023. Development and Evaluation of the Beta Streamflow Duration Assessment Method for the Northeast and Southeast Regions. Document No. EPA-843-R-23003.

Table of Contents

1 Introduction

Streamflow duration assessment methods (SDAMs) are rapid, field-based methods to determine flow duration class at the reach scale. The development of beta SDAMs for the Northeast and Southeast regions (hereafter referred to as the NE and SE) followed the conceptual framework and process steps presented by Fritz and others (2020) to integrate the three key components of an SDAM development study: hydrological data, indicators, and study reaches.

This supplemental document describes the data collection, data analysis, and evaluation steps that resulted in the beta SDAMs for the NE and SE. This document is available to inform public review and comment on the beta method, as well as serving as a companion to the beta SDAMs NE and SE for those that are interested in more background on the development of the methods and the underlying data. For a complete description of the beta SDAMs NE and SE protocol, please see the User Manual (James et al. 2023, https://www.epa.gov/system/files/documents/2023-04/Literature-Review-Beta-SDAM-NE-and-SE.pdf). The data used to develop the beta SDAMs NE and SE can be found here: (https://doi.org/10.23719/1528743). For more information on the collaborative effort between the U.S. Environmental Protection Agency (EPA) and the U.S. Army Corps of Engineers (Corps) to develop regional SDAMs for nationwide coverage, please see: https://www.epa.gov/streamflow-duration-assessment.

1.1 Streamflow Duration Classes

Streamflow duration governs important ecosystem functions (such as support for aquatic life, sediment transport, and biogeochemical processing rates), and streamflow duration classes are often used to guide watershed management decisions, including assessing the applicability of water quality standards. Our definitions of streamflow duration classes follow those used by Nadeau (2015):

- *Ephemeral reaches* flow only in direct response to precipitation. Water typically flows only during and/or shortly after large precipitation events, the streambed is always above the water table, and stormwater runoff is the primary water source.
- *Intermittent reaches* contain sustained flowing water for only part of the year, typically during the wet season, where the streambed may be below the water table or where the snowmelt from surrounding uplands provides sustained flow. The flow may vary greatly with stormwater runoff.
- *Perennial reaches* contain flowing water continuously during a year of normal rainfall, often with the streambed located below the water table for most of the year. Groundwater typically supplies the baseflow for perennial reaches, but the baseflow may also be supplemented by stormwater runoff or snowmelt.

For these definitions, a reach is a section of stream or river along which similar hydrologic conditions exist (e.g., discharge, depth, velocity, or sediment transport dynamics) and consistent drivers of hydrology are evident (e.g., slope, substrate, geomorphology, or

confinement). A channel is an area that is confined by banks and a bed and contains flowing water (continuously or not).

1.2 Overview of the Beta Method for the Northeast Southeast

The beta SDAMs for the NE and SE use a small number of indicators to predict the streamflow duration class of stream reaches. All indicators are measured during a single field visit. The beta SDAMs for the NE and SE result in one of four possible classifications: *ephemeral*, *intermittent*, *perennial*, or *at least intermittent*. The latter category occurs when an *intermittent* or *perennial* classification cannot be made with high confidence, but an *ephemeral* classification can be ruled out.

The tool uses a machine learning model known as random forest (Figure 1). Random forests are comprised of ensembles of decision trees. Random forest models are increasingly common in the environmental sciences because of their superior performance in handling complex relationships among indicators used to predict classifications (Breiman 2001, Cutler et al. 2007, Ellis et al. 2012). Random forests are well-suited handling different datatypes (i.e., Boolean, numeric, ordinal, categorical) that are present in the NE and SE datasets. They are appropriate for high-dimensional datasets and are generally capable of capturing interacting and non-linear relationships well. Because they are trained on different random subsets of the training data, the models are robust and typically perform well on novel data. This approach was previously used to develop regional SDAMs for the Pacific Northwest (PNW; Nadeau et al. 2015, Nadeau 2015), Arid West (AW; Mazor et al. 2021a, Mazor et al 2021b), Western Mountains (WM; Mazor et al. 2021c, Mazor et al. 2022), and Great Plains (GP; James et al. 2022, Eddy et al. 2022). Therefore, the use of random forests in developing the beta SDAMs for the NE and SE was an a priori decision that built on the findings of previous SDAMs.

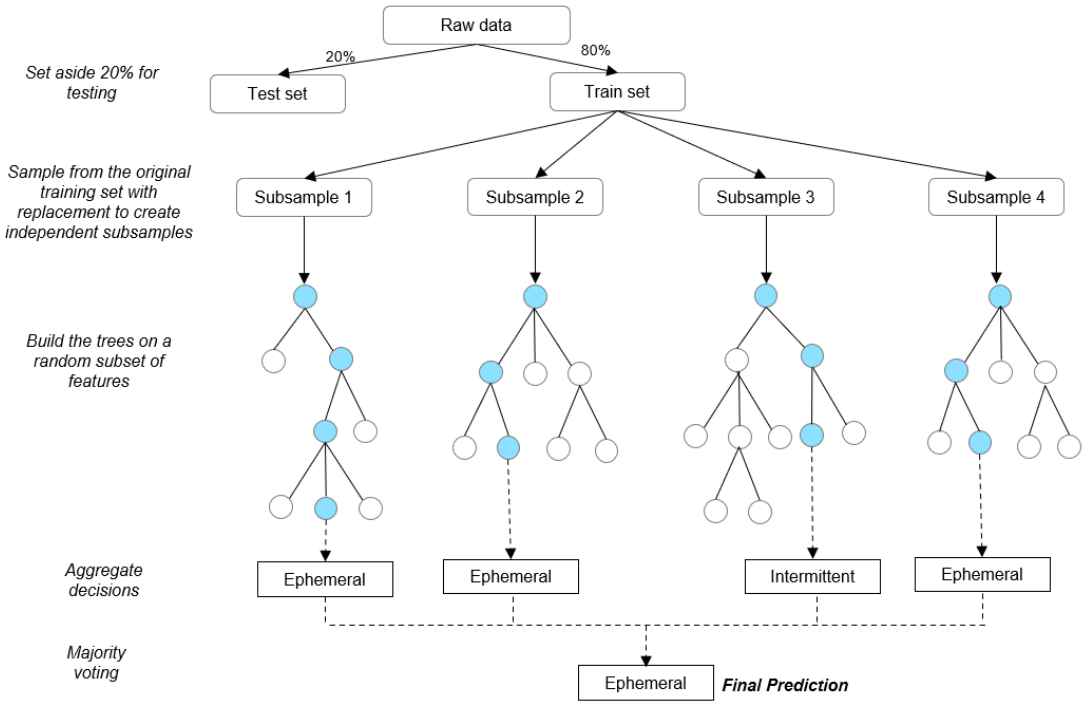


Figure 1. Random forest procedure used to determine a flow classification.

2 Development of the Beta Northeast and Southeast SDAMs

The specific data analysis steps described in this document follow the approach used to develop and evaluate the beta SDAM WM (Mazor et al. 2022) and the beta SDAM GP (Eddy et al. 2022).

2.1 Study Area

The NE and the SE regions (based on Wohl et al. 2016) include states along the Atlantic and Gulf coasts (including Puerto Rico and the U.S. Virgin Islands), extending into the Midwest as far as southeast Missouri (Figure 2). The NE includes all or part of 21 states and is considered those areas dominated by forest-type vegetation where snowmelt contributes at least some flow to streams and rivers during the year. Average yearly precipitation ranges widely across the region, but most areas receive between 40 and 50 inches of precipitation per year, on average. The SE includes all or part of 12 states or territories and is considered those areas characterized by forest-type vegetation that are generally dominated by diverse types of rainfall runoff rather than snowmelt, including tropical storms and hurricanes. Average yearly precipitation also ranges widely across the SE region, but most areas receive between 50 and 60 inches of precipitation per year, on average. Ephemeral and intermittent reaches may be found at any position within a watershed but are more common in smaller headwaters, where flow accumulation is insufficient to sustain longer-duration flows (Fritz et al. 2008). Ephemeral and intermittent reaches may also be more common along the western boundary of the SE region and more southern parts of the

NE, where average yearly precipitation totals are lowest (40 inches or less), and evapotranspiration is relatively high (Hammond et al. 2021).

The NE and SE regions as defined above include many metropolitan areas, including the New York City area (largest by population), Houston, Philadelphia, Baltimore-Washington D.C., Miami, and Atlanta, as well as some of the nation's fastest growing cities, such as Orlando, Raleigh-Durham, and Charlotte. Thus, there are places within the NE and SE where the need for an SDAM in permitting and management programs is, or continues to be, particularly high. In addition, development associated with oil and natural gas, as well as agricultural uses that may require more and/or modified water sources due to climate change, occur across the NE and SE (Vengosh et al. 2014, Perkin et al. 2017). North Carolina developed its own SDAM (NC SDAM) for statewide use in streams of all sizes and flow durations (NCDWQ 2010, Dorney and Russell 2018), which has served as a model for other methods developed in the NE and SE region, such as those developed for Tennessee (TDEC 2020) and parts of Virginia (e.g., James City County 2009). Both the North Carolina and Tennessee methods were developed to comply with Section 401 and state level rules (e.g., riparian buffers in NC), while the James City County method was developed to comply with requirements under the Virginia Chesapeake Bay Preservation Act. Ohio also has an SDAM (OH SDAM) for headwater streams (drainage areas <1.0 sq. mile; OH EPA 2020) which was developed to differentiate among classes of small headwater streams for appropriate characterization for aquatic life use designation.

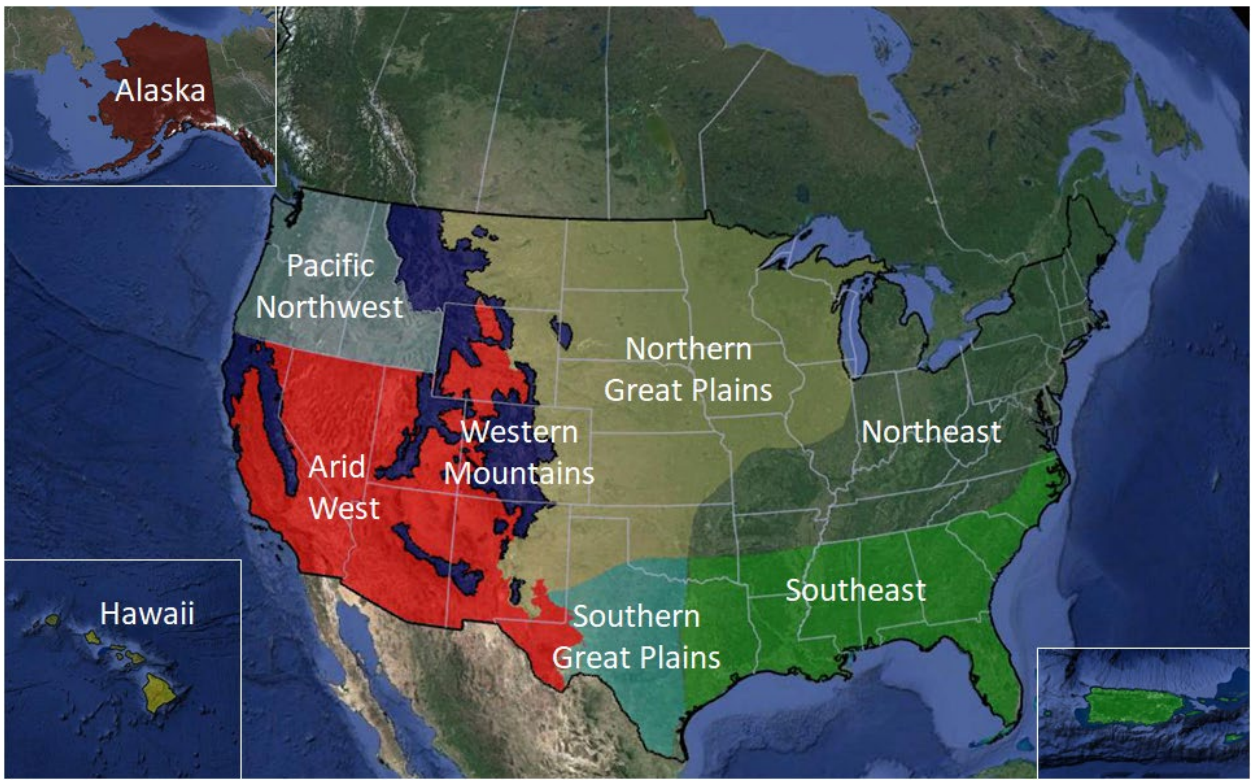


Figure 2. Map of SDAM study regions (based on Wohl et al. 2016). The beta SDAM NESE applies to the Northeast and Southeast regions as shown. Note, U.S. territories in the Caribbean Sea (Puerto Rico and the US Virgin Islands, inset) are not covered by the SE beta method though they were included in sampling.

2.2 Preparation and Candidate Indicators

At the outset of the project, a regional steering committee (RSC) was established consisting of technical staff at Corps Districts and EPA Regional Offices in the NE and SE regions that manage programs where streamflow duration information is often needed (e.g., Clean Water Act programs, including permits and enforcement). RSC members were selected based on their expertise in both scientific and programmatic elements relevant to streamflow duration classification needs. The RSC served several functions in the development process, such as reviewing technical products, facilitating connections with local experts, identifying resources such as sources of hydrologic data, and providing input on the model selection.

Candidate indicators were identified that are supported by the scientific literature (James et al. 2022) or used in previous SDAMs, including the SDAM PNW (Nadeau 2015) and others developed in the NE or SE as mentioned above (e.g., NCDWQ 2010, OH EPA 2020). Following input from the RSC, these candidate indicators were then screened using the criteria described by Fritz and co-authors (2020), including:

<u>Primary criteria</u>

- *Consistency*: Does the indicator consistently discriminate among flow duration classes (e.g., demonstrated in multiple studies)?

- *Repeatability:* Can different practitioners take similar measurements, given sufficient training and standardization?
- *Defensibility:* Does the indicator have a rational mechanistic relationship with flow duration, as either a response or a driver?
- *Rapidness:* Can the indicator be measured during a one-day reach-visit (even if subsequent lab analyses are required)?
- *Objectivity:* Does the indicator rely on objective (often quantitative) measures, as opposed to subjective judgments of practitioners?

Secondary criteria

- *Robustness:* Does human activity complicate indicator measurement or interpretation (e.g., poor water quality may affect the expression of some biological indicators)?
- *Practicality:* Can practitioners realistically sample the indicator with typical capacity, skills, and resources?

Candidate indicators were included in the study (Table 1) if they met all the primary criteria or at least one of the secondary criteria. Desktop geospatial indicators (derived using a geographic information system and applicable spatial datasets) that characterize mechanisms affecting flow duration and have been explored in other flow duration classification tools (e.g., Eng et al. 2016, Jaeger et al. 2019, Mazor et al. 2021c) were also included in the analysis.

Table 1. Candidate indicators evaluated in the present study. Indicators in the Origin column identified with "NC" followed the NC method protocol (NCDWQ 2010), with "NM" followed the NM method protocol (NMED 2011), with "OH" followed the Ohio protocol (OEPA 2020), and with "PNW" followed the PNW protocol (Nadeau 2015); other indicators (OTH) were measured with protocols developed for this study (USEPA 2020) and derived from sources resulting from a literature review completed by James et al. (2022) or recommendations from the RSC. Asterisks (*) indicate hydrologic indicators that are considered direct measures of water presence.

Candidate indicator	Description	Origin
Geomorphic indicators		
Channel continuity	Visual estimate of the continuity of bank and streambed development	NC
Sinuosity	Visual estimate of the curviness of the stream channel	NC
Bankfull width	Width of the channel at bankfull height	PNW
Floodplain channel dimensions	Visual estimate of the extent of channel entrenchment and connectivity to the floodplain	NM
Particle size of stream substrate	Visual estimate of the extent of evidence of substrate sorting within the channel	NC
Slope	Valley slope measured with a handheld clinometer	PNW
In-channel structure/ riffle pool sequence	Visual estimate of the diversity and distinctiveness of riffles, pools, and other flow-based microhabitats	NC
Active or relict floodplain	Visual estimate of floodplain characteristics adjacent to stream channel	NC

Candidate indicator		Description	Origin
	Depositional bars or benches	Visual estimate of the extent of alluvial bars and/or benches present in the channel	NC
	Recent alluvial deposits	Visual estimate of recently deposited alluvium in the channel and on the floodplain	NC
	Headcuts	Visual estimate of the size and number of headcuts in the channel	NC
	Grade control	Visual estimate of the extent and kinds of grade control features in the channel	NC
	Natural valley	Visual estimate of the extent of valley definition (proportion of catchment area sloping to the valley bottom).	NC
	Sediment deposition on plants and debris	Visual estimate of the extent of evidence of sediment deposition on plants and on debris within the floodplain	NC
Hydrologic indicators			
	Surface and subsurface flow*	Estimate of the percent of the reach-length with surface and subsurface flow	PNW
	Isolated pools*	Number of pools in the channel without any connection to flowing surface water	PNW
	Presence of baseflow*	Visual estimate of the extent of surface flow from groundwater discharge in the channel	NC
	Seeps and springs*	Presence/absence of springs or seeps within one-half channel width of the channel	NM
	Hydric soils	Presence/absence of hydric soils within the channel, measured at up to three locations	NC
	Leaf litter	Visual estimate of the extent of the streambed area covered by leaf litter	NC
	Maximum pool depth*	Measurement of deepest pool, in centimeters	OH
	Organic drift lines	Visual estimate of the size and distribution of organic debris accumulations in and along channels.	NC
	Soil moisture and texture*	Extent of soil saturation and texture measured at three locations in the channel	OTH
	Woody jams	Number of woody jams within the channel	OTH
Biological indicators			
	Live or dead algal cover	Visual estimate of the percent of streambed covered by live or dead algal growth	OTH
	Stream shading	Percent shade-providing cover above the streambed measured with a densiometer at three locations	OTH
	Hydrophytic plant species (channel only)	Number of OBL or FACW-rated plants (as listed in Lichvar et al. 2016) growing within the channel	NC
	Hydrophytic plant species	Number of all OBL or FACW-rated plants (as listed in Lichvar et al. 2016) growing within the channel or one half-channel width from the channel	PNW

Candidate indicator		Description	Origin
	Fish	Estimate of the overall abundance of fish (other than non-native mosquitofish) in the channel	NC
	Aquatic invertebrates	Estimate of the overall abundance and richness of aquatic invertebrates within the channel	NC
	Aquatic mollusks	Estimate of the overall abundance and richness of aquatic mollusks within the channel	NC
	Crayfish	Abundance of crayfish and palaeomonid shrimp (Decapoda) within the channel	NC
	Amphibians	Estimate of the overall abundance and richness of amphibians within the channel	NC
	Bryophytes	Visual estimate of the percent of streambed and banks covered by live or dead mosses or liverworts	OTH
	Differences in vegetation (riparian corridor)	Visual estimate of the distinctiveness of vegetation in the riparian corridor compared to surrounding upland vegetation	NM
	Absence of upland rooted plants in the streambed	Visual estimate of the extent of upland rooted plants growing within the streambed	NC
	Fibrous roots in streambed	Visual estimate of the extent and distribution of non-woody, small diameter roots of water-intolerant plants in the streambed	NC
	Fibrous roots in streambed	Visual estimate of the extent and distribution of non-woody, small diameter roots of water-intolerant plants in the streambed	NC
	Presence of iron-oxidizing fungi or bacteria	Presence of oily sheens indicative of iron-oxidizing fungi or bacteria within the assessment reach	NC
Geospatial indicators			
	Elevation	Elevation above mean sea level	OTH
	Drainage area	Drainage area measured using USGS StreamStats or National Mapper	OH
	Long-term normal precipitation and temperature	30-y normal mean annual and monthly precipitation, and 30-y normal mean, maximum, and minimum annual temperature (PRISM climate data; Hart and Bell 2015).	OTH
	Long-term mean snow persistence (1 January to 3 July)	Snow persistence (Hammond et al. 2017)	OTH
	Region	Northeast or Southeast	OTH
	Stream Order	Strahler stream order from USGS StreamStats synthetic network (first, second, or greater than second order)	NC

2.3 Candidate Reach Identification and Data Collection

The two objectives in selecting candidate reaches for this study were as follows: first, to include a sufficient number of reaches in each streamflow duration class to characterize variability in indicator measurements; and second, to select reaches representing the range of key natural and disturbance gradients within the NE and SE to support applicability of the method across anticipated conditions (Figure 3).

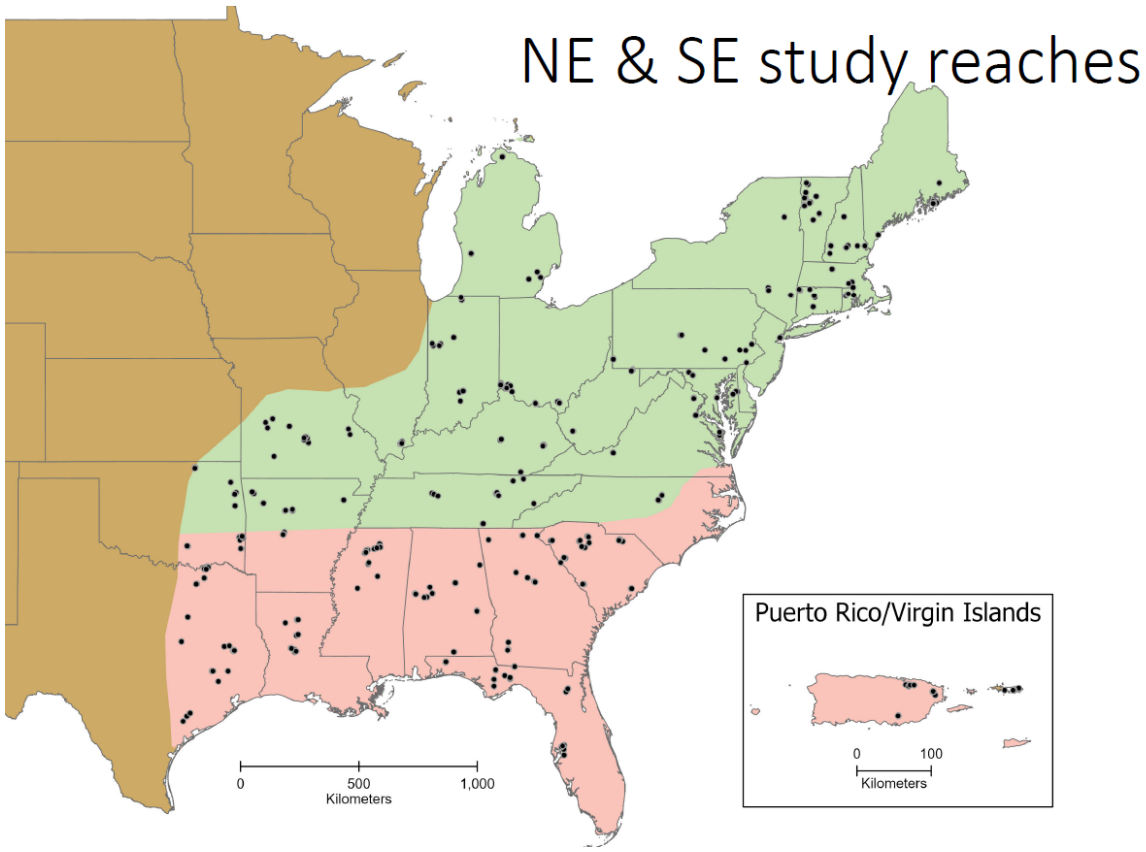


Figure 3. The Northeast and Southeast regions; study reaches sampled in support of the beta SDAMs for the NE and SE.

To screen reaches for use in method development, first a list of 8,712 candidate study reaches (5,494 in NE and 3,218 in SE) were compiled based on existing hydrologic data records (e.g., U.S. Geological Survey (USGS) stream gages, water presence loggers, wildlife cameras, field photos), published studies, and interviews with local experts familiar with the specific reach's hydrology. Most of these reaches (4,507 in NE and 2,591 in SE) were derived from the database of stream gages operated by the USGS and 4,330 (96%) in NE and 2,225 (86%) in SE were perennial. Actual streamflow duration class was determined by applying the flowchart in Figure 4, which was informed by existing definitions (Hedman and Osterkamp 1982, Hewlett 1982). Consequently, other sources were required to identify candidate ephemeral and intermittent reaches. Another 1,614 candidate study reaches (987 in NE and 627 in SE) were identified from

published studies or consultation with local experts. Whenever possible, multiple sources of hydrologic information were used to confirm actual streamflow classifications. In the resulting set of candidate reaches, 5% were determined to be ephemeral, 13% were intermittent, and 81% were perennial.

Reaches were prioritized for study inclusion based on being accessible (e.g., on public property or with landowner permission), being wadable, and the number and type of data sources available to determine actual streamflow duration classification. Reaches where streamflow duration class could be determined based on multiple data sources (e.g., water presence loggers and expert knowledge) were categorized as "preferred" for study inclusion. Reaches classified based solely on interpretation of USGS stream gage data without consultation of a local expert were categorized as "USGS gage" reaches. Reaches classified through local expertise alone were categorized as "acceptable" and included in the study to fill gaps in study regions where an insufficient number of "preferred" and "USGS gage" reaches classified as intermittent or ephemeral could be identified.

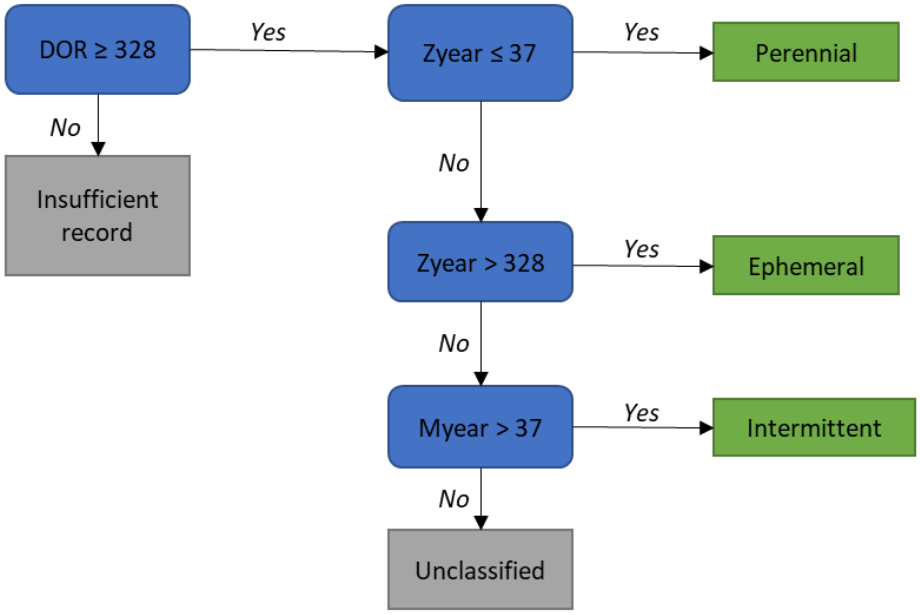


Figure 4. Flowchart used to determine actual streamflow duration class of reaches based on continuous measures of water presence (e.g., USGS stream gages). DOR: days of record. Zyear: Average number of dry days per year. Myear: Average length of longest continuous wet period per year, in days. For USGS gages, at least 20 years of data were analyzed whenever possible (Kelso and Fritz 2021).

Of the 8,712 candidate reaches, 388 (202 in NE and 186 in SE) study reaches were sampled from October 2020 to October 2022. These study reaches were parsed into 'instrumented' and

'single-visit' reaches[1]. Instrumented reaches (117 in NE and 120 in SE) were visited multiple times (up to four), and each had at least one Stream Temperature, Intermittence, and Conductance (STIC; Chapin et al. 2014) logger deployed, with all instrumented reaches having duplicate data loggers installed by the second field visit. Instrumented reaches generally had fewer existing lines of evidence to determine actual streamflow duration classification before sampling; therefore, post-sampling reach classifications were reviewed in light of the STIC logger data and hydrology indicator data that were direct measures of water presence collected during each visit. For further details on STIC data loggers and their verification/calibration, deployment, and data retrieval, see Schumacher and Fritz (2019). Single-visit reaches (85 in NE and 66 in SE) were visited once (with a 10% resample) and did not have loggers deployed. Because actual streamflow duration classification of most single-visit reaches was determined using existing data collected by USGS and others, these reaches generally had multiple direct flow duration data sources. Ultimately, due to data loss from STIC loggers and other factors, actual streamflow duration class at 33 reaches (30 instrumented and three single-visit reaches) could not be determined with confidence and were excluded from analysis used to develop the beta SDAMs for the NE and SE. In addition, 19 sites from Puerto Rico and U.S. Virgin Islands were excluded from the SE dataset for reasons described below. Of the remaining 336 study reaches used to develop the beta SDAMs for the NE and SE, 71 were ephemeral, 150 were intermittent, and 115 were perennial (Table 2).

Table 2. Distribution of reaches used to develop the beta SDAMs for the NE and SE. Instrumented reaches were visited up to four times and had Stream Temperature, Intermittence, and Conductance loggers installed and single-visit reaches were visited once (rarely, twice) and did not have loggers installed.

	Single-Visit		Instrumented			
Class	**Gaged**	**Preferred**	**Gaged**	**Preferred**	**Acceptable**	**Total**
Ephemeral	*3*	*31*	*0*	*10*	*27*	*71*
Northeast	3	15	0	7	13	38
Southeast	0	16	0	4	13	33
Intermittent	*14*	*44*	*4*	*20*	*68*	*150*
Northeast	10	26	2	9	39	86
Southeast	4	18	2	11	29	64
Perennial	*20*	*24*	*6*	*16*	*49*	*115*
Northeast	12	17	5	8	24	66
Southeast	8	7	1	8	25	49

During each field visit to a study reach, the suite of candidate indicators (Table 3) was measured following the development protocol (USEPA 2019). This compilation of indicators from a single field visit constitutes one reach sample (or observation) in terms of the analyses described within this data analysis supplement. Surrounding land use may affect or disturb streamflow

[1] These reaches were termed 'baseline' and 'validation', respectively, in prior beta SDAMs but have been renamed for clarity.

duration indicators without substantially shifting flow duration at reaches (e.g., changes in water quality). Up to two predominant land use categories within a 100-m radius of each study reach were noted on each field visit. If “urban” or “agriculture” were the identified land use category the sample was considered disturbed; otherwise, the sample was considered not disturbed for comparisons of beta NE and SE SDAM performance.

2.4 Data analysis

2.4.1 Metric calculation

Candidate indicator data were used to create 96 candidate metrics, of which 42 were biological, 14 were geomorphological, five were hydrologic (indirectly measured water presence), and 35 were geospatial (Table 3). Note, additional metrics were developed and introduced during the model refinement steps and are discussed below in Section 2.4.4.

As in the development of previous SDAMs, direct measures of water were excluded from the analysis. Metrics that directly measure water (e.g., soil moisture, number of isolated pools, water in channel) can greatly increase performance. However, such metrics introduce circularity (because water presence was used to confirm and update actual streamflow duration classes in the development data set) and may degrade the ability of the SDAM to perform well during atypical conditions, such as drought. Although such metrics might provide valuable supporting information in an assessment, including it in the SDAM could introduce circularity and reduce acceptance of the tool (Mazor et al. 2021b). Therefore, only indirect measures of water presence (e.g., hydric soils and number of woody jams) were included in the development of a beta method for the NE and SE regions.

Table 3. Candidate metrics evaluated for the development of the beta SDAMs for the NE and SE. Please see Appendix A for full definitions of candidate metrics. Abbreviations in candidate metric names include – EPT: Ephemeroptera, Plecoptera, and Trichoptera insect orders. GOLD: Gastropoda, Oligochaeta, and Diptera invertebrate groups. OCH: Odonata, Coleoptera, and Heteroptera insect orders. For Type the following categories apply – Ord: Ordinal metrics. Cat: Categorical metrics. Bin: Binary metrics. Con: Continuous metrics. The following fields provide the screening criteria – PctDom: Percent of reach samples with the most common value (typically zero). Min: minimum value. Max = maximum value. Range: Maximum possible value minus minimum possible value for the candidate metric. PvIvE: F-statistic from a comparison of mean values at perennial, intermittent, and ephemeral reaches. EvALI: Absolute t-statistic from a comparison of mean values at ephemeral and at least intermittent reaches. PvNP: Absolute t-statistic from a comparison of mean values at perennial and non-perennial reaches. PvIWet: Absolute t-statistic from a comparison of mean values at flowing intermittent and perennial reaches. EvIdry: Absolute t-statistic from a comparison of mean values at non-flowing intermittent and ephemeral reaches. rf_MDA: Metric importance from a random forest model, measured as mean decrease in accuracy. Screened: Indicates if the metric passed or failed screening criteria in Table 4. NA: Not applicable.

Candidate metrics	Group	Type	PctDom	Min	Max	Range	PvIvE	EvALI	PvNP	PvIwet	EvIdry	rf_MDA	Screened
MeanSnowPersistence_01	GIS	Con	1%	0.0	44.9	44.9	16.70	7.48	4.28	2.68	3.54	0.01	Pass
MeanSnowPersistence_05	GIS	Con	1%	0.1	45.1	45.0	16.61	7.62	4.21	2.56	3.66	0.01	Pass
MeanSnowPersistence_10	GIS	Con	1%	0.2	45.4	45.2	15.68	7.33	4.10	2.45	3.58	0.01	Pass
ppt	GIS	Con	1%	772.7	1717.3	944.6	4.00	1.48	2.77	3.18	1.85	0.01	Pass
ppt.m01	GIS	Con	1%	32.3	154.4	122.1	1.96	1.00	1.30	0.73	1.99	0.01	Fail
ppt.m02	GIS	Con	1%	33.0	147.8	114.8	1.40	1.90	0.42	0.41	1.95	0.01	Fail
ppt.m03	GIS	Con	1%	45.0	158.5	113.5	2.66	2.44	1.05	0.73	2.35	0.01	Pass
ppt.m04	GIS	Con	2%	57.6	138.1	80.5	9.85	0.85	4.40	4.05	0.70	0.01	Pass
ppt.m05	GIS	Con	1%	60.0	180.3	120.4	27.48	2.56	7.64	5.86	0.78	0.02	Pass
ppt.m06	GIS	Con	1%	69.0	197.1	128.2	22.91	2.94	7.20	6.77	3.52	0.03	Pass
ppt.m07	GIS	Con	1%	45.7	206.2	160.5	0.75	1.13	0.61	1.21	2.34	0.01	Pass
ppt.m08	GIS	Con	1%	51.7	232.4	180.7	1.58	0.31	1.88	0.24	2.55	0.01	Pass
ppt.m09	GIS	Con	1%	68.9	178.0	109.1	3.87	2.47	0.32	2.04	0.09	0.01	Pass
ppt.m10	GIS	Con	1%	56.9	150.4	93.4	3.14	1.67	1.16	2.07	0.49	0.01	Pass
ppt.m11	GIS	Con	1%	46.4	162.3	116.0	1.28	0.05	1.47	1.75	0.11	0.01	Fail
ppt.m12	GIS	Con	1%	47.2	166.2	119.0	2.96	2.15	2.08	1.67	1.31	0.01	Pass
temp.m01	GIS	Con	2%	-9.6	15.8	25.4	4.67	3.08	2.36	1.95	3.30	0.01	Pass
temp.m02	GIS	Con	1%	-7.9	17.3	25.3	5.50	3.27	2.66	2.19	3.26	0.01	Pass
temp.m03	GIS	Con	2%	-3.3	19.4	22.7	7.63	3.79	3.22	2.55	3.29	0.01	Pass
temp.m04	GIS	Con	2%	3.9	21.8	17.9	8.08	3.91	3.34	2.64	3.27	0.01	Pass
temp.m05	GIS	Con	2%	10.5	25.4	14.9	6.67	3.38	3.09	2.61	3.20	0.01	Pass
temp.m06	GIS	Con	2%	15.5	27.9	12.4	5.74	3.18	2.84	2.32	2.84	0.01	Pass
temp.m07	GIS	Con	3%	17.6	28.9	11.3	4.92	2.58	2.82	2.30	2.25	0.01	Pass
temp.m08	GIS	Con	2%	16.8	29.3	12.5	6.24	2.75	3.25	2.72	2.37	0.01	Pass
temp.m09	GIS	Con	2%	12.8	27.4	14.6	5.16	2.85	2.78	2.40	2.93	0.01	Pass

Candidate metrics	Group	Type	PctDom	Min	Max	Range	PvIvE	EvALI	PvNP	PvIwet	EvIdry	rf_MDA	Screened
temp.m10	GIS	Con	2%	6.5	24.4	17.9	5.16	2.80	2.82	2.52	3.06	0.01	Pass
temp.m11	GIS	Con	2%	0.6	20.5	19.9	4.26	2.74	2.41	2.17	3.20	0.01	Pass
temp.m12	GIS	Con	2%	-6.1	17.0	23.2	2.69	2.37	1.71	1.61	3.13	0.01	Pass
tmax	GIS	Con	2%	10.6	28.3	17.6	5.89	3.08	2.93	2.43	2.96	0.01	Pass
tmean	GIS	Con	2%	4.8	22.7	17.9	5.51	3.12	2.77	2.32	3.10	0.01	Pass
tmin	GIS	Con	2%	-1.0	17.5	18.5	4.96	3.11	2.54	2.17	3.22	0.01	Pass
ActiveFloodplain_score	Geomorph	Ord	29%	0	3	3	31.06	8.15	5.07	0.64	3.20	0.00	Pass
AlluvialDep_score	Geomorph	Ord	46%	0	3	3	32.76	9.16	6.10	1.04	2.10	0.00	Pass
BankWidthMean	Geomorph	Con	3%	0.2	52.3	52.1	46.75	11.19	7.33	4.61	2.80	0.01	Pass
ChannelDimensions_score	Geomorph	Ord	55%	0	3	3	6.55	0.35	3.06	3.98	0.45	0.00	Pass
Continuity_score	Geomorph	Ord	54%	0	3	3	64.56	8.39	9.05	2.42	3.45	0.00	Pass
Depositional_score	Geomorph	Ord	30%	0	3	3	64.74	10.66	9.52	3.24	2.98	0.00	Pass
fp_entrenchmentratio_mean	Geomorph	Con	36%	1.0	2.5	1.5	8.32	0.98	3.91	4.16	0.39	0.00	Pass
GradeControl_score	Geomorph	Ord	28%	0	1.5	1.5	3.14	0.44	2.11	1.48	0.31	0.00	Pass
Headcut_score	Geomorph	Ord	72%	0	3	3	23.95	6.09	4.41	0.58	1.69	0.00	Pass
NaturalValley_score	Geomorph	Ord	39%	0	1.5	1.5	12.61	3.98	1.55	2.34	2.88	0.00	Pass
RifflePoolSeq_score	Geomorph	Ord	35%	0	3	3	26.28	4.97	6.77	1.17	0.57	0.00	Pass
Sinuosity_score	Geomorph	Ord	55%	0	3	3	18.86	6.20	3.37	0.15	3.41	0.00	Pass
Slope	Geomorph	Con	24%	0	46	46	33.30	6.58	6.45	1.73	2.23	0.01	Pass
SubstrateSorting_score	Geomorph	Ord	42%	0	3	3	52.45	7.07	9.31	3.53	2.92	0.01	Pass
HydricSoils_score	H20 (Indirect)	Bin	74%	0	3	3	26.06	6.35	2.45	1.60	5.13	0.00	Pass
LeafLitter_score	H20 (Indirect)	Ord	29%	0	1.5	1.5	84.33	12.47	9.85	2.84	4.08	0.00	Pass
ODL_score	H20 (Indirect)	Con	38%	0	1.5	1.5	27.02	7.01	5.34	1.16	3.69	0.00	Pass
SedimentOnPlantsDebris_score	H20 (Indirect)	Ord	33%	0	1.5	1.5	23.89	8.32	3.59	0.65	4.34	0.00	Pass
WoodyJams_number	H20 (Indirect)	Ord	82%	0	11	11	0.59	0.95	0.68	0.04	0.21	0.00	Fail
DRNAREA_mi2	GIS	Con	5%	0.0	289.0	289.0	17.01	6.88	4.30	3.16	2.20	0.03	Pass
Elev_m	GIS	Con	3%	10	887	877	7.20	2.57	3.60	2.14	0.07	0.01	Pass
REGION	GIS	Cat	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
StreamOrder	GIS	Ord	56%	1	3	2	81.55	17.65	9.41	3.84	5.95	0.01	Pass
alglive_cover_score	Bio	Ord	63%	0	4	4	40.57	11.91	6.58	1.64	3.76	0.00	Pass
alglivedead_cover_score	Bio	Ord	63%	0	4	4	40.00	12.08	6.41	1.39	3.81	0.00	Pass
Amphib_abundance	Bio	Con	74%	0	42	42	12.17	7.40	3.66	1.96	2.99	0.00	Pass
Amphib_richness	Bio	Con	97%	1	3	2	4.87	4.79	2.42	1.45	2.49	0.00	Fail

Candidate metrics	Group	Type	PctDom	Min	Max	Range	PvIvE	EvALI	PvNP	PvIwet	EvIdry	rf_MDA	Screened
Crayfish_abundance	Bio	Con	79%	0	10	10	7.49	4.52	2.77	0.90	2.44	0.00	Pass
fishabund_score2	Bio	Ord	78%	0	1.5	1.5	41.93	14.60	6.87	2.23	3.14	0.00	Pass
ironox_bfscore	Bio	Ord	85%	0	3	3	25.14	8.92	5.63	3.95	3.87	0.00	Pass
EPT_abundance	Bio	Con	48%	0	72	72	88.57	13.61	10.57	6.29	1.78	0.01	Pass
EPT_taxa	Bio	Con	48%	0	12	12	122.06	16.26	12.66	6.99	2.77	0.01	Pass
GOLD_abundance	Bio	Ord	46%	0	36	36	37.69	12.02	6.51	2.15	3.59	0.00	Pass
Mollusk_abundance	Bio	Con	89%	0	16	16	10.76	6.12	3.67	1.87	1.64	0.00	Pass
Mollusk_taxa	Bio	Con	89%	0	5	5	9.25	5.50	3.57	1.32	1.54	0.00	Pass
Noninsect_abundance	Bio	Con	45%	0	46	46	19.07	7.70	2.02	1.98	5.13	0.00	Pass
Noninsect_taxa	Bio	Con	45%	0	8	8	30.72	9.83	4.18	0.71	5.25	0.00	Pass
OCH_abundance	Bio	Con	60%	0	25	25	20.77	8.51	4.82	2.56	2.60	0.00	Pass
perennial_NC_abundance	Bio	Con	46%	0	64	64	106.33	14.58	11.33	7.64	1.98	0.02	Pass
perennial_NC_live_abundance	Bio	Con	47%	0	58	58	106.94	14.42	11.38	7.77	1.94	0.02	Pass
perennial_NC_taxa	Bio	Con	46%	0	13	13	169.20	18.63	14.69	9.26	2.95	0.03	Pass
perennial_PNW_abundance	Bio	Con	62%	0	34	34	72.93	11.76	9.13	7.15	2.90	0.01	Pass
perennial_PNW_live_abundance	Bio	Con	62%	0	33	33	76.51	11.66	9.34	7.53	2.90	0.01	Pass
perennial_PNW_taxa	Bio	Con	62%	0	8	8	115.77	14.90	11.82	8.47	2.85	0.01	Pass
Richness	Bio	Con	25%	0	20	20	181.13	20.01	14.95	7.04	5.86	0.01	Pass
TolRelAbund	Bio	Con	39%	0	1	1	24.87	6.44	0.96	4.85	4.98	0.01	Pass
TotalAbundance	Bio	Con	25%	0	105	105	101.84	15.74	10.73	4.54	5.14	0.01	Pass
DifferencesInVegetation_score	Bio	Ord	39%	0	3	3	49.65	11.27	6.95	0.71	2.96	0.00	Pass
FibrousRootedPlants_score	Bio	Ord	55%	0	3	3	30.83	4.60	8.07	3.31	0.99	0.00	Pass
hydrophytes_inchannel	Bio	Con	70%	0	12	12	11.42	7.47	2.40	0.28	2.12	0.00	Pass
hydrophytes_present	Bio	Con	35%	0	16	16	55.31	12.91	7.72	3.32	3.65	0.00	Pass
hydrophytes_present_noflag	Bio	Con	37%	0	16	16	51.20	12.22	7.48	3.32	3.20	0.00	Pass
liverwort_cover_score	Bio	Ord	91%	0	3	3	13.16	5.69	4.22	3.24	2.01	0.00	Pass
moss_cover_score	Bio	Ord	94%	0	3	3	5.32	4.83	2.24	0.54	0.77	0.00	Pass
OBL_inchannel	Bio	Con	83%	0	8	8	14.67	8.88	3.59	0.81	2.52	0.00	Pass
OBL_present	Bio	Con	74%	0	10	10	21.62	10.42	4.24	1.15	3.28	0.00	Pass
OBL_present_noflag	Bio	Con	76%	0	10	10	18.65	9.45	4.04	1.23	2.66	0.00	Pass
PctShading	Bio	Con	41%	0	1	1	6.25	5.01	1.42	0.83	1.24	0.00	Pass
UplandRootedPlants_score	Bio	Ord	58%	0	3	3	124.21	10.01	15.81	5.96	2.70	0.02	Pass

2.4.2 Metric Screening

Metric screening was performed across the entire NE and SE datasets. As an initial data exploration step, the relationships between actual streamflow duration class (hereafter "flow class") and indicators by ordinating all 90 candidate metrics for all samples in the dataset in a nonmetric multidimensional scaling using Gowers' distance (Gower 1971) were visualized. Convex hulls were drawn around each flow class to help visualize their distributions in ordination space. The ordination of all candidate metrics for NE and SE samples showed extensive overlap of intermittent, ephemeral, and perennial reaches, indicating the challenge of separating samples by flow class (Figure 5). This was the case for both the Northeastern and Southeastern regions, as well as the U.S. Caribbean (Figure 6 and Figure 7).

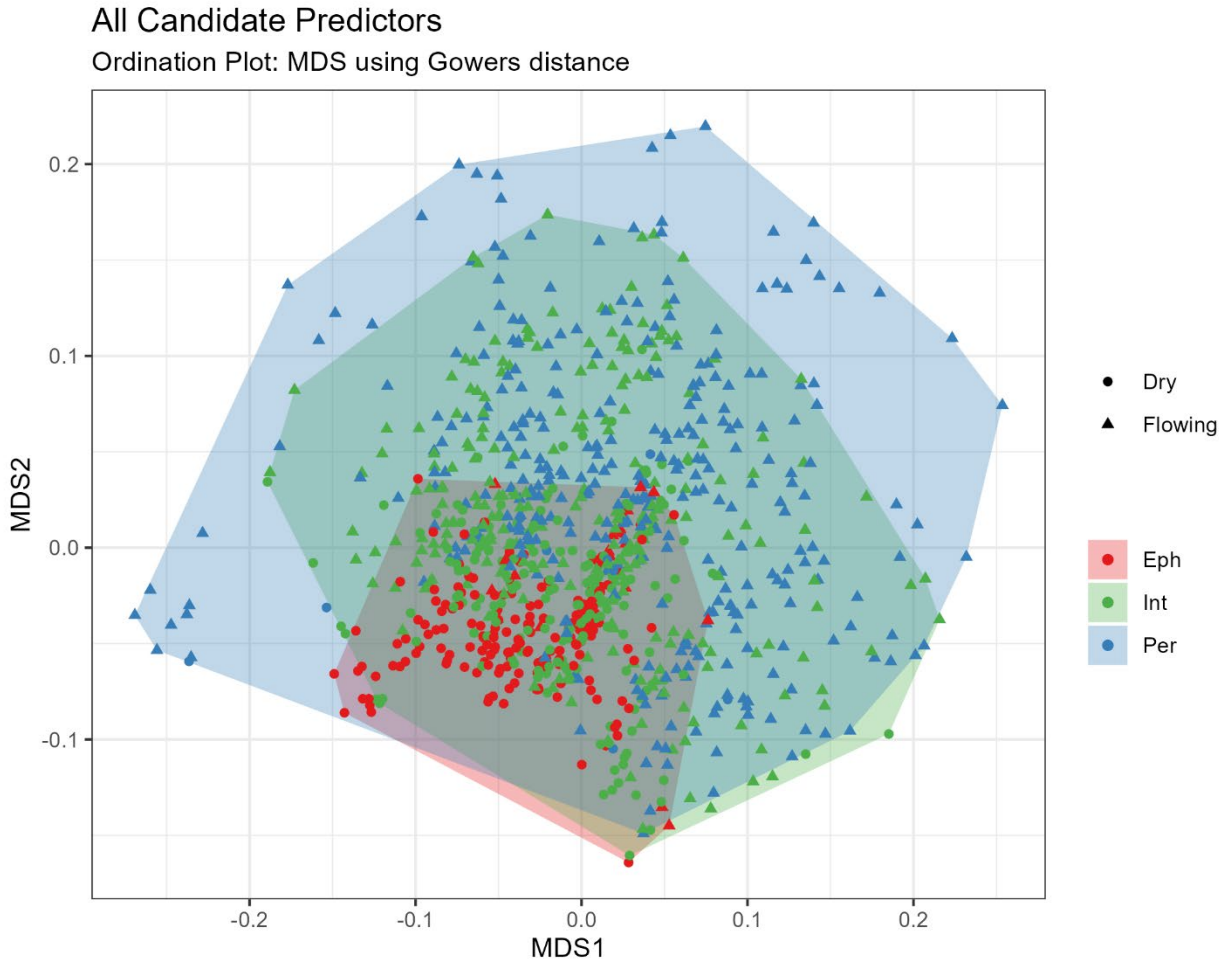


Figure 5. Beta SDAM candidate metric ordination. Ordination plot shows candidate metrics from all regions (from Northeast, Southeast, and U.S. Caribbean). Overlapping red, green, and blue hulls indicate the challenge of separating by flow class.

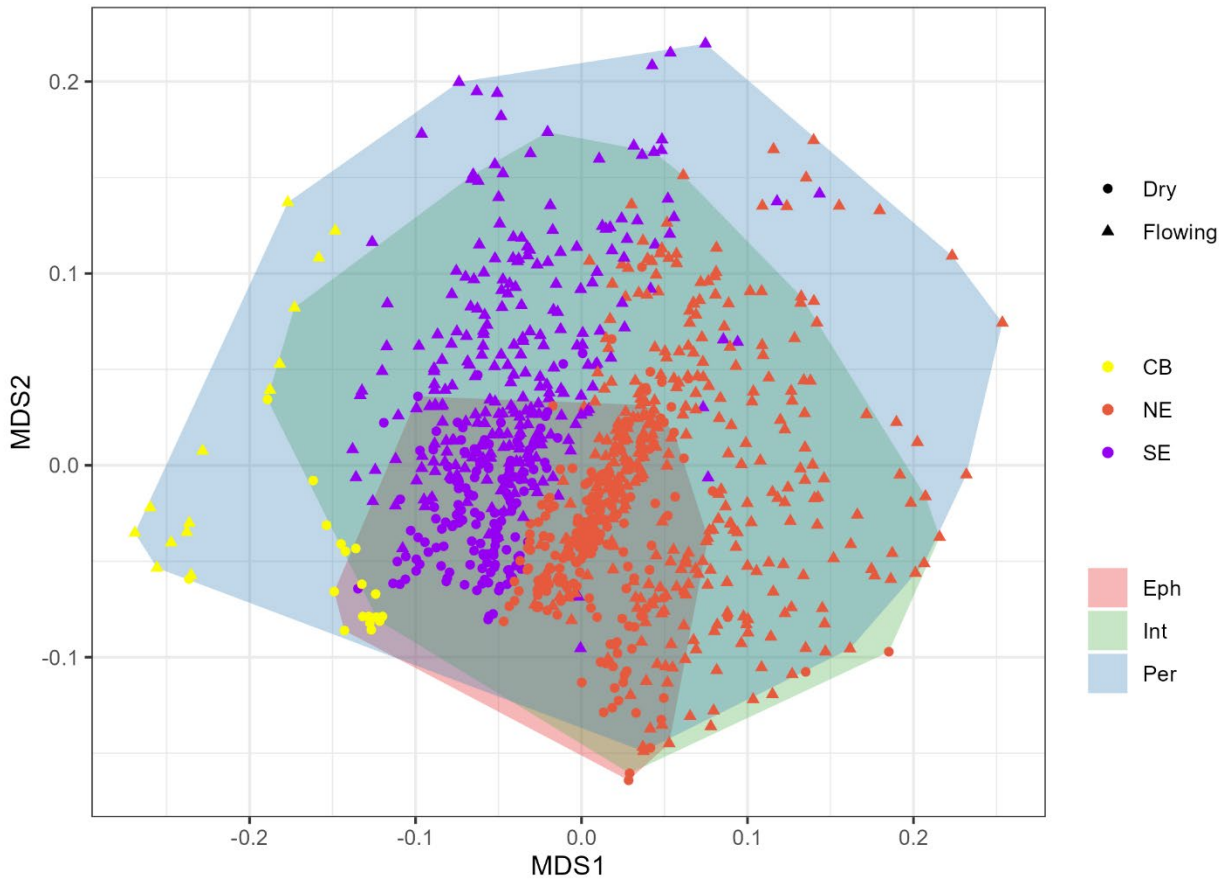


Figure 6. Same as Figure 5; points are colored by Region (Northeast = orange, Southeast = purple, Caribbean = yellow).

Figure 6 and Figure 7 show the same ordination plot as in Figure 5 but with points colored by region. Clustering of points may indicate the presence of regional similarities, likely driven in large part by geospatial metrics. The subsequent steps in SDAM calibration excluded the U.S. Caribbean samples because of 1) the degree of separation between the U.S. Caribbean samples and most of the SE samples in the ordination (Figure 6) and 2) the limited number of samples available to develop a U.S. Caribbean-specific SDAM at this time. Spread of points across overlapping Ephemeral (red), Intermittent (green), and Perennial (blue) hulls indicate the challenge of distinguishing flow class within either region.

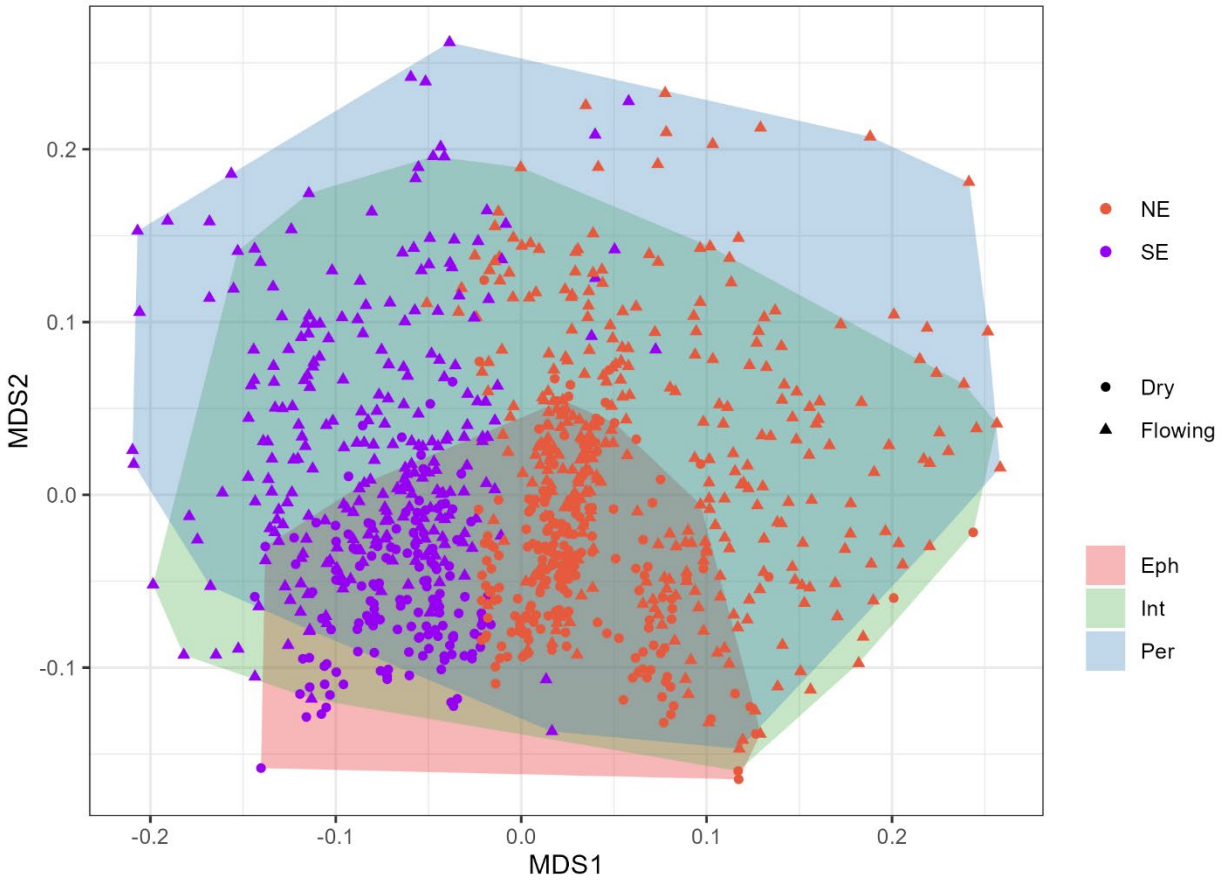


Figure 7. Same ordination plot as Figure 6 but with Caribbean data removed; points are colored by Region (Northeast = orange; Southeast = purple).

After the initial data exploration, candidate metrics were evaluated using criteria for inclusion in the beta SDAMs for the NE and SE (Table 4) characterizing distribution of data and responsiveness:

- Distribution statistic criterion: calculated as percent dominance of the most common value (which was typically zero); all metrics had to meet this criterion.
- Criteria measuring the responsiveness of metrics (i.e., ability to discriminate across flow classes) included:
    - A set of statistical comparisons of mean values at different subsets of reaches (e.g., t-statistic from a comparison of metric values at perennial and non-perennial reaches), as has been used in other studies (Hawkins et al. 2010, Cao and Hawkins 2011, Mazor et al. 2016).
    - A responsiveness statistic based on metric importance (specifically, mean decrease in accuracy) from a random forest model to predict flow class from all candidate metrics; the model was calibrated using the default option from the randomForest function in the *randomForest* package in R (Liaw and Wiener 2002).

Candidate metrics had to meet at least one responsiveness criterion, in addition to the distribution criterion, to be considered in further analyses. A total of 89 of 90 candidate metrics were considered as screened metrics (REGION was not included in screening). Five metrics failed: Amphib_richness, ppt.m01, ppt.m02, ppt.m011, and WoodyJams_number. These metrics failed because they had a Percent Dominance (PctDom) score greater than 95% and/or because they did not meet at least one of the responsiveness criteria. Note that this evaluation was carried out using the training dataset described in the next section and that U.S. Caribbean data were not included.

Table 4. Metric screening criteria. Metrics had to meet the distribution criterion and at least one responsiveness criterion to be considered screened for further analysis.

Criterion		Definition
Distribution criterion		
% dominance of most common value	<95%	Frequency of most common value (typically, zero) in the development data set
Responsiveness criteria		
PvIvE	F>2	F-statistic in a comparison of values at perennial versus intermittent versus ephemeral reaches
EvALI	t>2	t-statistic in a comparison of values at ephemeral versus at least intermittent reaches
PvNP	t>2	t-statistic in a comparison of values at perennial versus non-perennial reaches
PvIwet	t>2	t-statistic in a comparison of values at perennial versus flowing intermittent reaches
EvIdry	t>2	t-statistic in a comparison of values at ephemeral versus dry intermittent reaches
rf_MDA	Top quartile	Mean decrease accuracy (MDA) in a random forest model to predict perennial, intermittent, or ephemeral streamflow duration class

2.4.3 Data Preparation

Prior to method development, a portion of the data was withheld for use in final model testing. Samples from 20% of the study reaches, balanced by Class and Region, were withheld into a "test" dataset. These samples were used to inform the final model(s) selection and refinement, by evaluating the model(s) on novel reaches. Samples from the remaining 80% of the reaches were used to develop (or "train") the model and are referred to hereafter as the training dataset.

U.S. Caribbean data were not included in the beta model development because there were not enough samples to adequately inform the modeling process. However, a separate section of this data supplement (2.4.8) is dedicated to evaluating how well the final beta model performs in Puerto Rico and the U.S. Virgin Islands using the U.S. Caribbean data as a testing set. As explained in more detail below, the classification accuracy of separate models for NE and SE versus that of a global (or unstratified) model for NE and SE was assessed during the model-

building process to determine whether separate models for the NE and SE would provide more accurate classifications than a single, unstratified model that would be used across both regions.

2.4.3.1 Repeat reach visits

Of the 336 reaches included in the NE and SE datasets, each was visited between one and four times, yielding a total of 916 samples. Figure 8 shows the distribution of repeat reach visits.

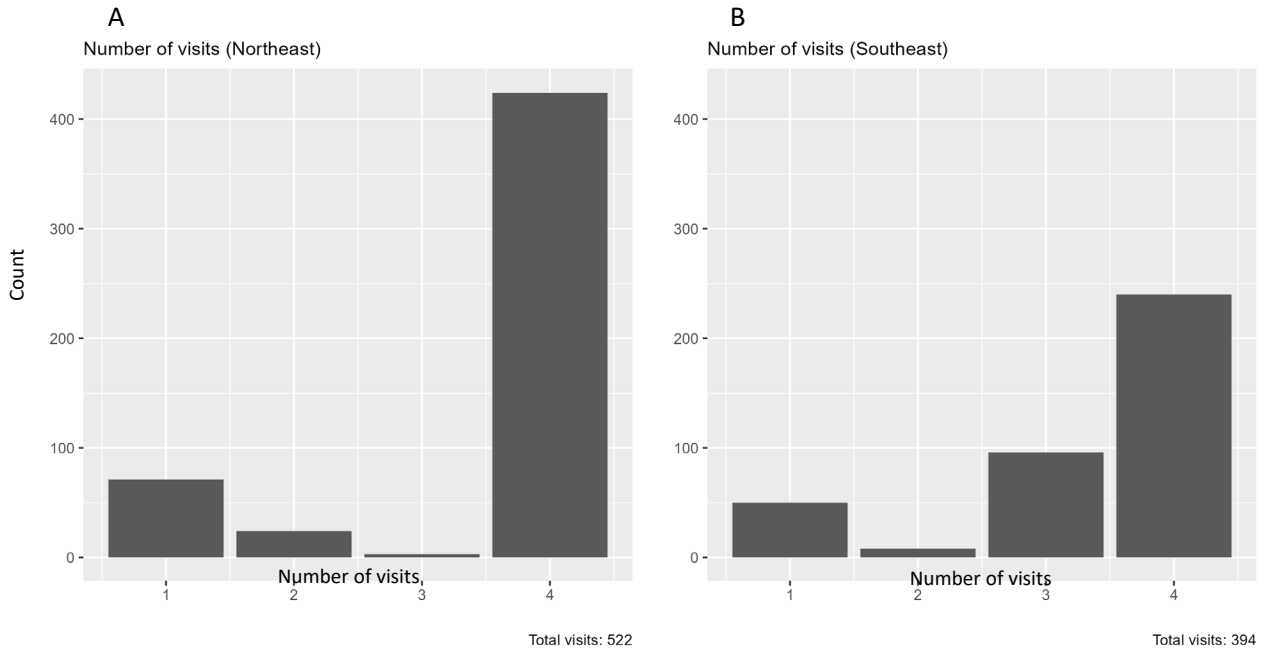


Figure 8. Distribution of number of visits across the (A) Northeast (190 reaches) and (B) Southeast (146 reaches).

To minimize bias, oversampling was performed on the training dataset (Figure 9). Oversampling is a common preprocessing step that serves to give under-represented classes more visibility in the data (Mohammed et al. 2020).

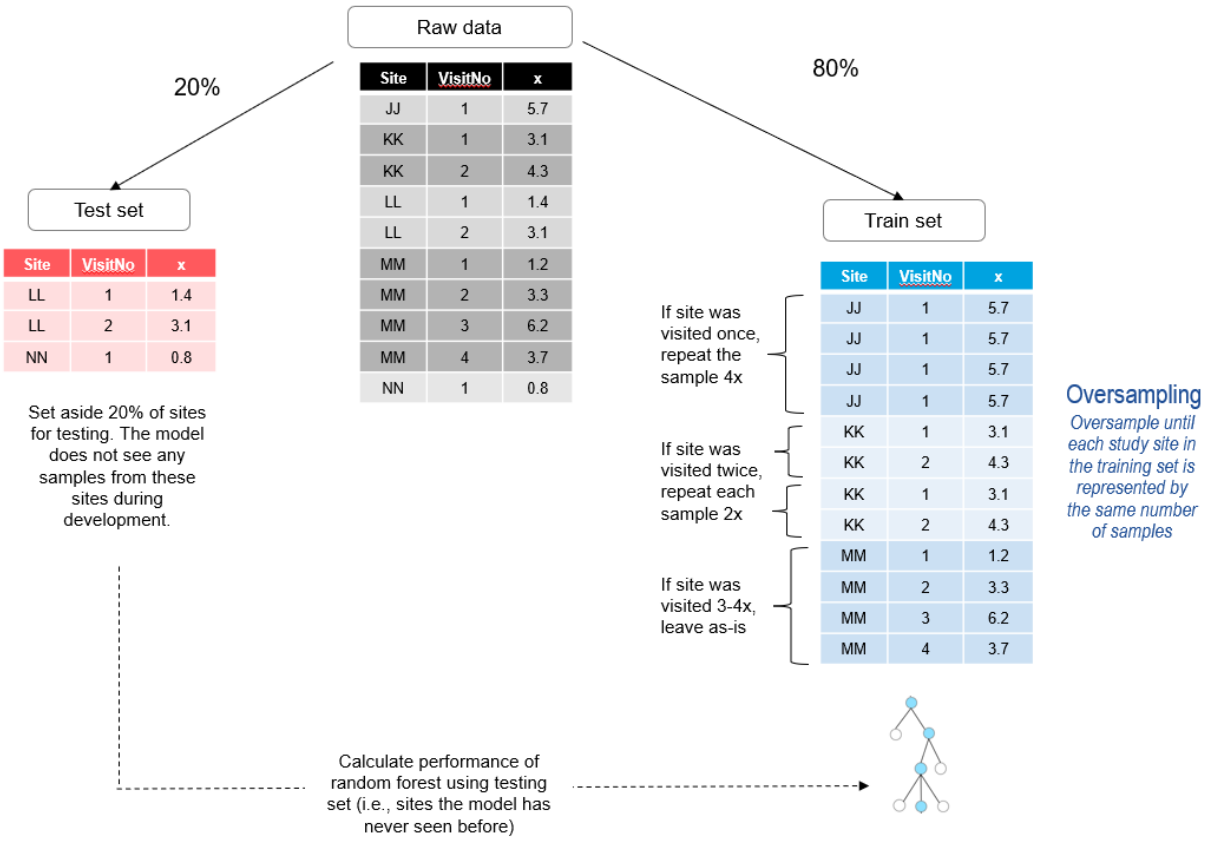


Figure 9. Oversampling process used for training dataset. x is a hypothetical candidate indicator.

Oversampling was performed on the training dataset only (no manipulations were conducted on the test dataset) and included the following steps:

- If a reach was sampled one time, each sample was repeated four times.
- If a reach was sampled twice, each sample was repeated two times.
- If a reach was sampled three or four times, the samples were left as-is.

The result of the oversampling process was that each study reach had three or four samples used in the analysis process for method development and the distribution of flow duration classes was preserved from the original training dataset to the oversampled training dataset, which also matched well to the distribution of flow duration classes within the testing dataset (Figure 10). Therefore, the augmented (oversampled) training data with 1060 samples were used in the next step of the method development analysis process to select screened metrics.

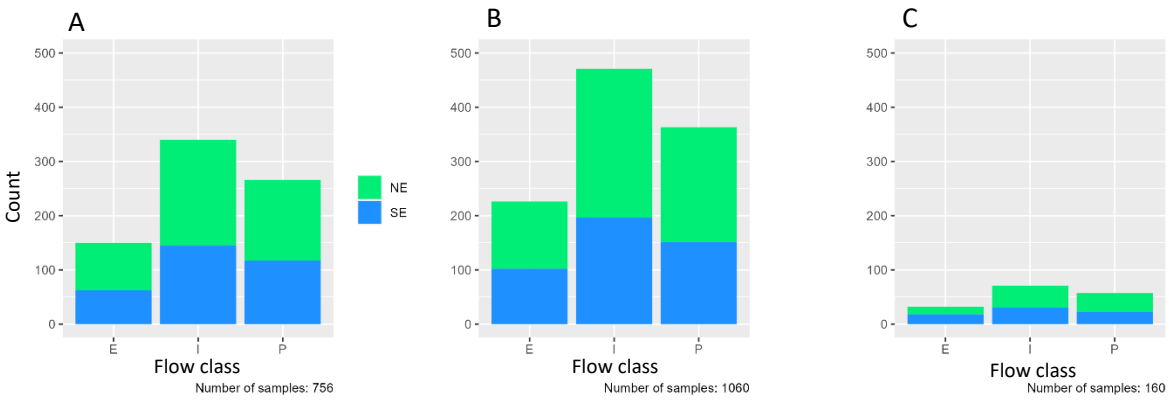


Figure 10. Distribution of ephemeral (E), intermittent (I), and perennial (P) classes in the (A) training dataset before oversampling, (B) training dataset after oversampling, and the (C) testing dataset (not oversampled). Shown for each bar is the number of samples for a streamflow duration class colored by region. A balanced distribution between classes is important to mitigate bias and improve model accuracy. Note that no oversampling is performed on the testing data.

2.4.4 Metric selection

The 94 screened metrics were reduced to a final set of 89 metrics for the NE and SE beta SDAMs based on their importance in random forest models using the Recursive Feature Elimination (RFE) function in the R *caret* package (Kuhn 2020). Briefly, RFE is a form of stepwise selection where complex models (i.e., those based on many metrics) are calibrated and simpler models are considered incrementally by eliminating the least important metrics. Here, the most complex model was first considered. Then, the five least important metrics were eliminated based on their relative performance in the random forest model. This process was iterated until a 20-metric model was identified, after which only one metric was eliminated in each successive step. The best-performing model (i.e., highest accuracy in predicting true streamflow duration class) was identified. Then, the simplest model (i.e., the one with the fewest metrics) with accuracy within 1% of the model with the best accuracy was selected to identify the final set of metrics. If the simplest model selected by this approach had more than 20 metrics, the 20-metric model was selected. For this analysis, accuracy on the training dataset was measured with Cohen's Kappa statistic—a measure of accuracy that accounts for uneven distribution among the three streamflow duration classes. Note that the Kappa statistic varies from 0 to 1, where 0 equals agreement equivalent to chance and 1 equates to perfect agreement between the predicted and true classification. Due to the use of random forest models, the Out-of-Bag (OOB) error rate is provided. This means that the prediction error measure for the model is computed through bootstrap or bagging, where subsampling with replacement creates a set of training samples for the model to learn from and the OOB error is the mean prediction error on each training sample for all flow classes (James et al. 2013).

At the beginning of the analysis, it was unclear whether greater accuracy would be achieved having one model for the combined NE and SE (Unstratified model) or two models separated by region (Stratified models). Thus, the preliminary modeling process was applied to all three types of models.

In addition, it was unclear how many geospatial (GIS) metrics should be included in the model-building process. There are advantages and disadvantages to including geospatial metrics in an SDAM. GIS metrics may improve SDAM performance but require GIS analysis in the application of the resulting method. Furthermore, GIS metrics tend to dominate during the RFE selection, resulting in models that are almost entirely comprised of geospatial metrics. See Mazor et al. (2021b) for a discussion of the implications of including geospatial metrics in SDAMs. Initial investigation of stratified and unstratified models allowing all GIS metrics showed that the average monthly precipitation metric for June (e.g., ppt.m06) was consistently among the most important metrics in addition to drainage area (Figure 11). However, allowing RFE to select any number of GIS metrics led to entirely GIS or GIS-dominated models. An attempt was made to limit the number of GIS indicators selected by aggregating the average monthly precipitation and temperature metrics into averages over three consecutive months (seasons) centered around the most consistently important month across all models, June ppt.m06, and adding the adjacent months to create new aggregate GIS metrics for precipitation and temperature (Table 5). The other GIS metrics (Elev_m, StreamOrder, DRNAREA_mi2, REGION, and the three MeanSnowPersistance metrics) were not changed. In addition, the annual summary climate metrics tmax, tmin, tmean, and ppt which were never or rarely selected in any of the models and so were dropped. This screening reduced the number of GIS metrics available for selection by RFE from 35 to 15 metrics. Ultimately, three new experiments were performed using all the screened candidate field metrics and the shorter list of new aggregated GIS metrics: allow RFE to select any number of GIS metrics (all GIS); prevent the model from selecting any GIS metrics (no GIS); and allow the model to include exactly two GIS metrics (2 GIS). The 2 GIS experiment was seen as a compromise between two extremes, where the highest-performing geospatial metrics could be included without letting the models become GIS-dominated.

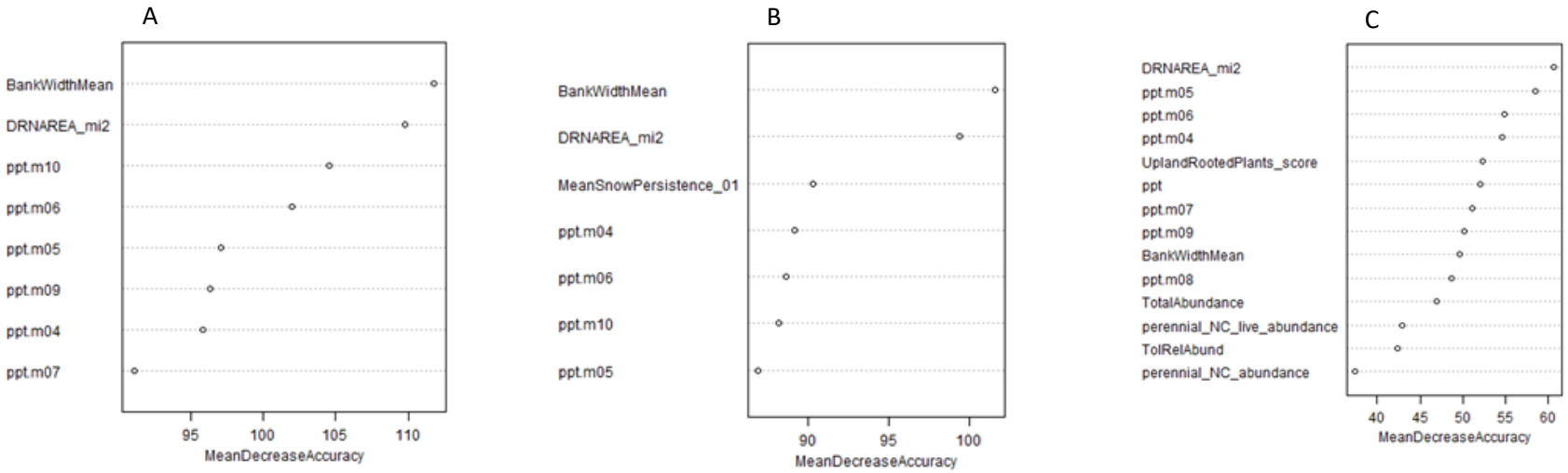


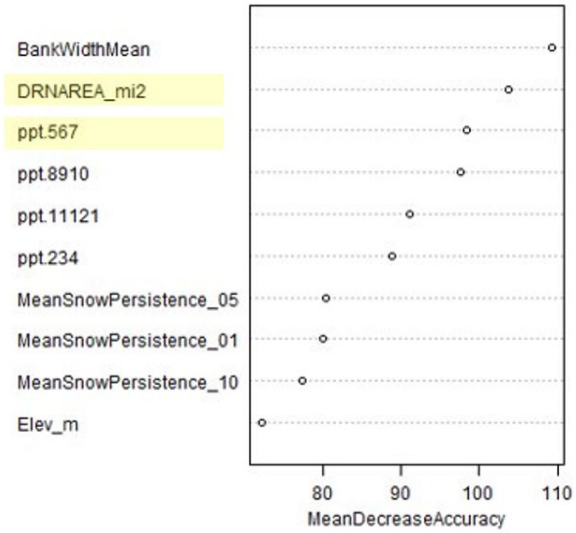
Figure 11. Metrics selected by Recursive Feature Elimination for the (A) unstratified (B) Northeast only, and (C) Southeast only models when allowing all candidate GIS metrics.

Table 5. Aggregation of GIS metrics

Original GIS metrics	New (aggregate) GIS metrics
ppt.m02	ppt.234
ppt.m03	
ppt.m04	
ppt.m05	ppt.567
ppt.m06	
ppt.m07	
ppt.m08	ppt.8910
ppt.m09	
ppt.m10	
ppt.m11	ppt.11121
ppt.m12	
ppt.m01	
ppt	Dropped from consideration (rely on seasonal precipitation instead)
temp.m02	temp.234
temp.m03	
temp.m04	
temp.m05	temp.567
temp.m06	
temp.m07	
temp.m08	temp.8910
temp.m09	
temp.m10	
temp.m11	temp.11121
temp.m12	
temp.m01	
tmax	Dropped from consideration (rely on seasonal temperature instead)
tmean	
tmin	

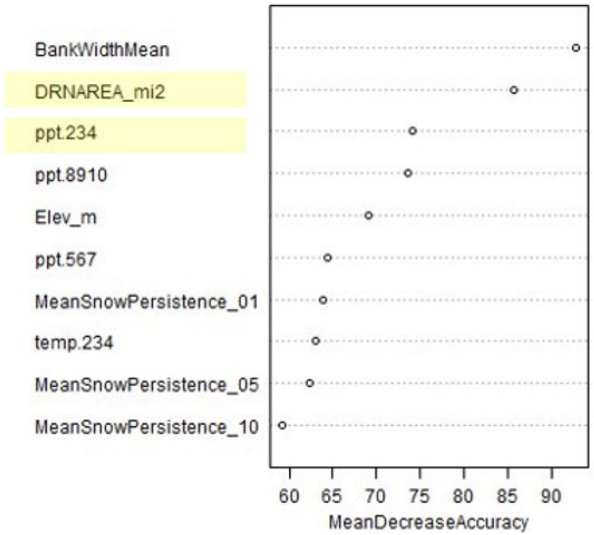
Even with fewer GIS metrics available for selection following aggregation, the RFE algorithm still created models that were largely GIS-dominated (Figure 12). Thus, the two most important GIS metrics for each model were chosen to proceed in the analysis with the remaining GIS metrics eliminated from consideration. Here, "importance" was determined mainly using Mean Decrease in Accuracy, which is the relative loss in predictive performance when the particular metric is omitted from the model. However, the Mean Decrease in Gini Index, or how important the metric is in splitting between different streamflow duration classes, was also applied to confirm the selection of the two most important GIS metrics. Gini importance measures how well a potential decision tree split separates between the flow classes. Ranking the Gini index provides insight on which metrics may be most relevant to the models by indicating the "quality" of the split, or how much information was gained by using the metric to discriminate between flow classes.

A



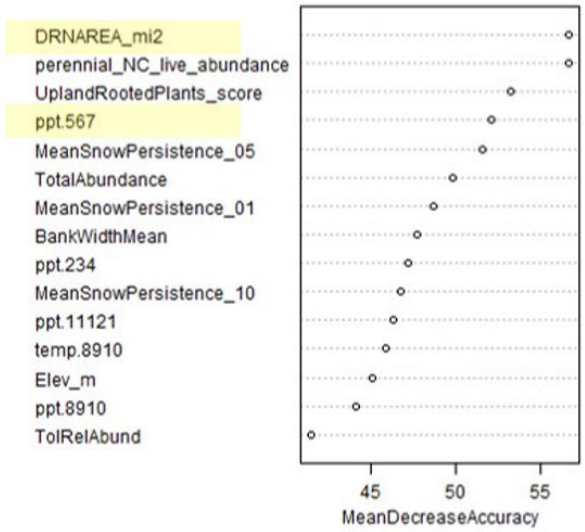
NE % Correct PvIvE (test): 58.9%
NE % Correct EvALI (test): 85.6%
SE % Correct PvIvE (test): 32.9%
SE % Correct EvALI (test): 65.7%

B



% Correct PvIvE (test): 67.8%
% Correct EvALI (test): 83.3%

C



% Correct PvIvE (test): 55.7%
% Correct EvALI (test): 74.3%

Figure 12. Metrics selected by Recursive Feature Elimination for the (A) unstratified, (B) Northeast only, and (C) Southeast only models when using aggregated GIS metrics. Models were still GIS-dominated; therefore, analysis proceeded using only the top two performing GIS metrics in each model (highlighted in yellow).

The modeling process (including RFE) produced nine models using the training dataset:

1. Unstratified Model (all GIS): a single model for the entire NE and SE with no limit on the number of GIS metrics that may be selected by RFE
2. Unstratified Model (no GIS): a single model for the entire NE and SE region that excludes GIS metrics (e.g., temp, precip) from consideration
3. Unstratified Model (2 GIS): a single model for the entire NE and SE, where only the two highest-performing geospatial metrics may be selected

4-5. Stratified Models (all GIS): separate models for the NE and SE regions, with no limit on the number of GIS metrics that may be selected by RFE

6-7. Stratified Models (no GIS): separate models for the NE and SE regions that exclude GIS metrics (e.g., temp, precip) from consideration

8-9. Stratified Models (2 GIS): separate models for the NE and SE regions, where only the two highest-performing GIS metrics within a region may be selected

The nine models were compared to determine the degree of improved performance by regional stratification versus a combined model and by the number of GIS metrics (all, two most important, none) that were included. Model design characteristics and optimal number of metrics selected by RFE are shown in Table 6, and the selected metrics for each model are shown in Figure 13.

Biological metrics, particularly those based on aquatic invertebrates, were among the most frequently selected metrics across model sets (Figure 13). Among non-biological metrics, mean bankfull width was the only frequently selected geomorphological metric.

Table 6. Design characteristics of the nine models. Unstratified models include Northeast (NE) and Southeast (SE) training data. Models stratified by region are models using only NE or SE training data. All GIS: no limitation on number of selected geospatial metrics. No GIS: excluded all geospatial metrics. 2 GIS: maximum of two selected geospatial metrics. # samples: number of samples used in model training and testing. # metrics: number of metrics eligible and selected in best models. RFE OOB error rate: Out-Of-Bag (OOB) error rate of the best model produced by recursive feature elimination (RFE).

Model set	# samples (training - original)	# samples (training - augmented)	# samples (testing)	# metrics eligible	# metrics selected	# GIS metrics selected	Kappa (training)	RFE OOB error rate
Unstratified models								
NESE (all GIS)	756	1060	160	69	10	9	0.99	0.19
NESE (no GIS)	756	1060	160	54	14	N/A	0.71	19.06
NESE (2 GIS)	756	1060	160	56	10	2	0.79	13.77
Models stratified by region								
NE (all GIS)	432	611	90	69	10	9	0.99	0.16
NE (no GIS)	432	611	90	54	13	N/A	0.70	19.64
NE (2 GIS)	432	611	90	56	10	2	0.82	9.17
SE (all GIS)	324	449	70	69	15	10	0.92	3.79
SE (no GIS)	324	449	70	54	12	N/A	0.75	16.93
SE (2 GIS)	324	449	70	56	10	2	0.82	12.25

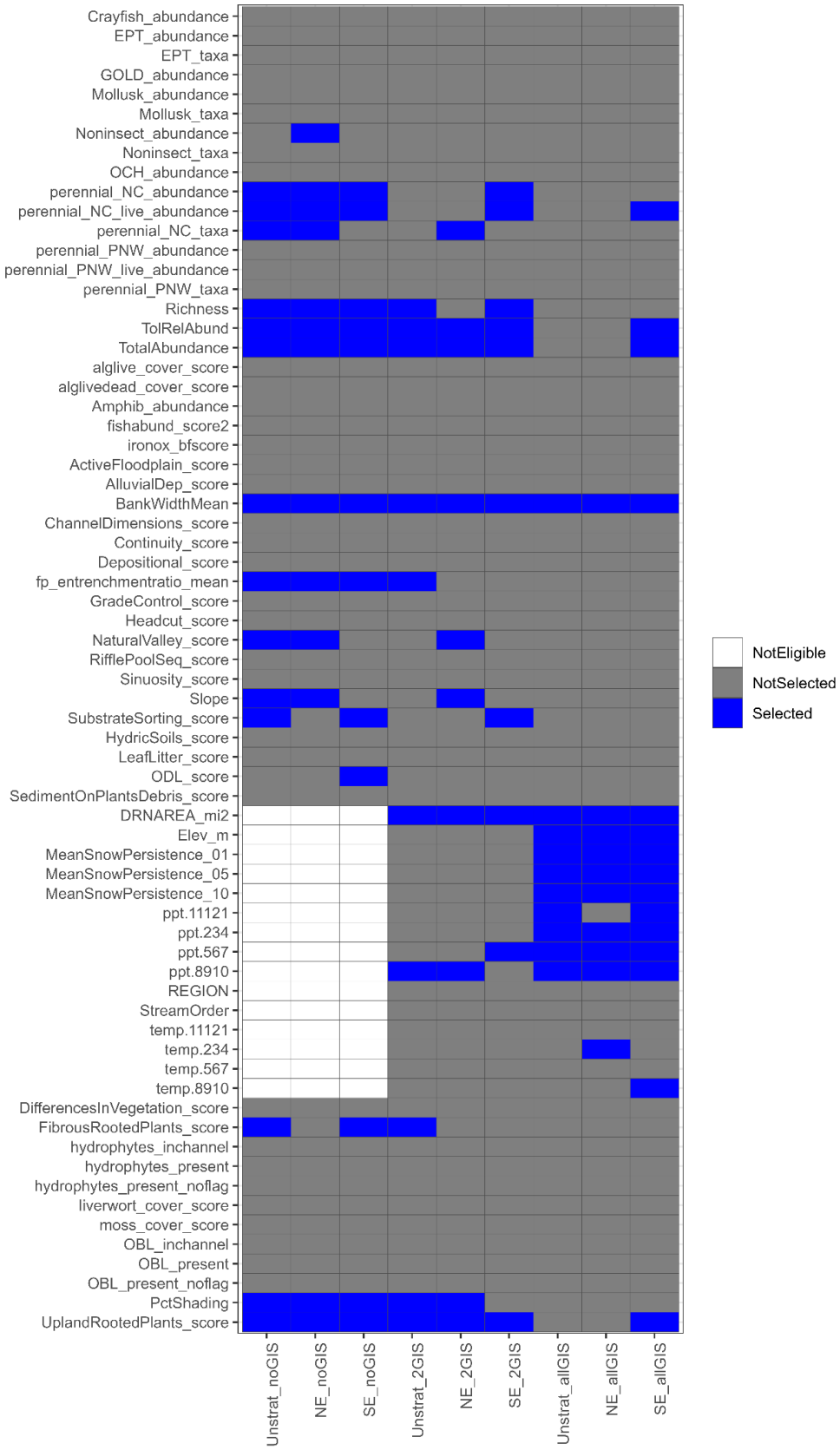


Figure 13. Screened metrics (left) selected by Recursive Feature Elimination for each model (bottom, see Table 6). White tiles indicate that a screened metric was ineligible for selection in that model set (e.g., Elev_m was ineligible for models that did not allow GIS metrics). Y-axis labels refer to screened metrics described in Table 3 and Appendix A.

2.4.4.1 Preliminary model calibration and performance assessment

Random forest models were fit for each of the nine models using the randomForest function in the *randomForest* package in R (Liaw and Wiener 2002) using default parameters, except that the number of trees was set to 1500 instead of the default 500.

Model performance evaluation focused on two aspects: accuracy and repeatability (Table 7 and Figure 14). Accuracy was assessed by calculating the same comparisons used to evaluate metric responsiveness during the metric screening phase (e.g., ephemeral versus at least intermittent reaches [EvALI], perennial versus wet intermittent reaches [PvIwet], etc.; Appendix A). Accuracy of a model's ability to correctly distinguish among flow classes was assessed on both the training and testing datasets independently. Training and testing measures of accuracy were compared to see if models validated poorly (training dataset accuracy substantially higher than testing dataset accuracy). Poorly validated models may be overfitting for the training reaches and thus may not be generally predictive of streamflow duration classification. The performance of unstratified models was evaluated by examining results for reaches within each region separately.

Repeatability, or precision, was assessed using data from the 215 reaches that were resampled (Figure 8) and calculated as the percent of reaches where model classifications from repeated samples at the same reach were consistent (regardless of classification accuracy). Due to the limited amount of data, precision was only assessed for the entire NE and SE and not within each stratum (Table 7).

Table 7. Performance evaluation of the nine Random Forest model sets developed for the NE and SE (see Table 6 for descriptions). PvIvE: Percent of reach samples classified correctly as perennial, intermittent, or ephemeral. EvALI: Percent of reach samples classified correctly as ephemeral or at least intermittent. PvNP: Percent of reach samples classified correctly as perennial or non-perennial. PvIwet: Percent of flowing reach samples classified correctly as perennial or intermittent. IvEdry: Percent of dry reach samples correctly classified as intermittent or ephemeral. Precision: percent of reaches classified consistently across visits. Train: Result for training data. Test: Result for testing data.

		Accuracy										
		PvIvE		EvALI		PvNP		PvIwet		IvEdry		
Model set	Scope	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Precision
Unstrat (all GIS)	East	100%	57%	100%	81%	100%	73%	100%	57%	100%	55%	97%
Unstrat (no GIS)	East	81%	64%	92%	89%	88%	84%	81%	60%	81%	71%	81%
Unstrat (2 GIS)	East	86%	66%	94%	91%	91%	75%	86%	61%	86%	74%	82%
Unstrat (all GIS)	NE	100%	59%	100%	84%	100%	74%	100%	63%	100%	52%	99%
Unstrat (no GIS)	NE	79%	52%	90%	88%	88%	62%	80%	47%	77%	61%	82%
Unstrat (2 GIS)	NE	86%	56%	93%	89%	92%	66%	87%	53%	84%	61%	80%
Unstrat (all GIS)	SE	100%	54%	100%	77%	100%	71%	100%	50%	100%	60%	94%
Unstrat (no GIS)	SE	84%	79%	95%	90%	89%	89%	83%	75%	86%	83%	80%
Unstrat (2 GIS)	SE	87%	80%	96%	93%	91%	87%	86%	73%	89%	90%	84%
Regional (all GIS)	NE	100%	62%	100%	86%	100%	77%	100%	64%	100%	58%	97%
Regional (no GIS)	NE	80%	59%	91%	91%	88%	68%	81%	54%	79%	68%	88%
Regional (2 GIS)	NE	91%	63%	96%	91%	95%	72%	92%	59%	89%	71%	86%
Regional (all GIS)	SE	96%	51%	98%	71%	97%	77%	97%	60%	94%	40%	86%
Regional (no GIS)	SE	83%	81%	95%	93%	88%	89%	82%	80%	84%	83%	87%
Regional (2 GIS)	SE	88%	69%	96%	89%	91%	79%	87%	63%	89%	77%	90%
Regional (all GIS)	East	98%	58%	99%	79%	99%	77%	98%	62%	98%	50%	95%
Regional (no GIS)	East	69%	82%	92%	93%	77%	88%	66%	82%	75%	81%	88%
Regional (2 GIS)	East	89%	66%	96%	90%	93%	75%	90%	61%	89%	73%	87%

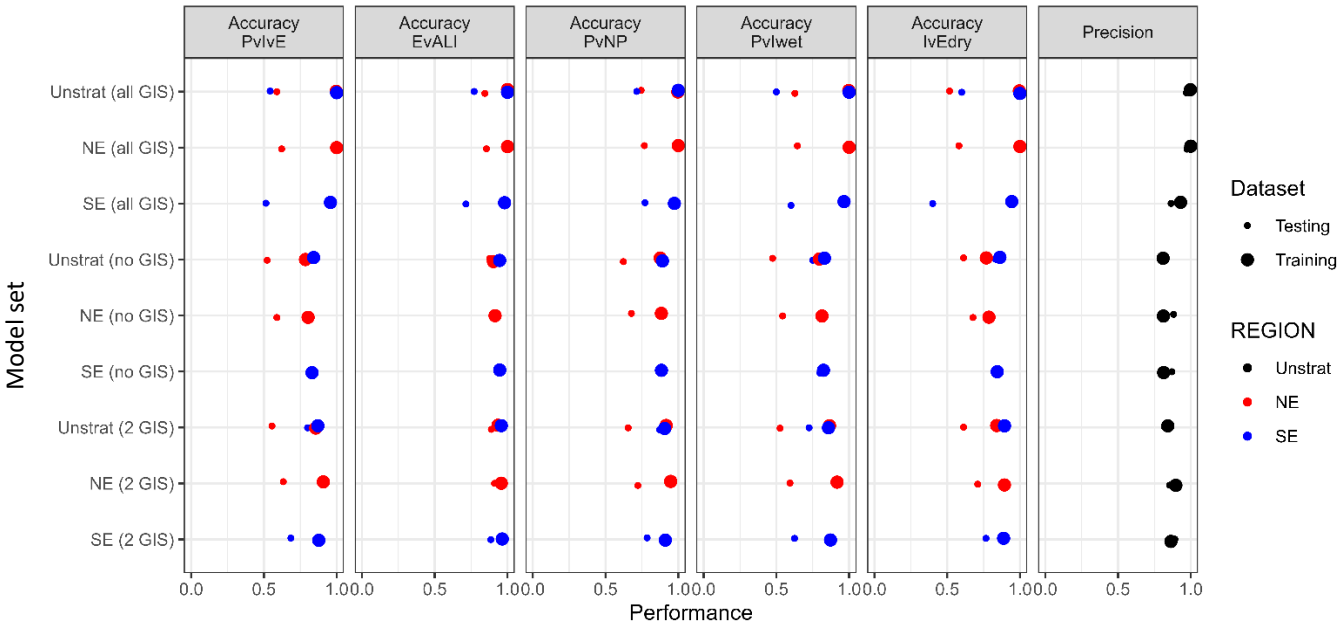


Figure 14. Performance of the nine Random Forest model sets developed for Northeast and Southeast regions (see Table 6 for descriptions). PvIvE: Proportion of reach samples classified correctly as perennial, intermittent, or ephemeral. EvALI: Proportion of reach samples classified correctly as ephemeral or at least intermittent. PvNP: Proportion of reach samples classified correctly as perennial or non-perennial. PvIwet: Proportion of flowing reach samples classified correctly as perennial or intermittent. IvEdry: Proportion of dry reach samples correctly classified as intermittent or ephemeral. Precision: proportion of reaches classified consistently across visits.

2.4.4.2 Comparison of Model Options

Performance of the nine preliminary models (described in Section 2.4.4) were evaluated and compared in detail using confusion matrices (Figures 15, 16, and 17). Rather than simply tallying correct versus incorrect predictions, confusion matrices provide further information by summarizing how the correct and incorrect predictions are spread across the three streamflow duration classes. In each confusion matrix, the X-axis lists the actual streamflow class and the Y-axis lists the predicted class. Thus, the main diagonal (blue) highlights correct predictions, while the other cells indicate incorrect predictions.

Confusion matrices using the training dataset are expected to be the most accurate because the models were developed using those data. Confusion matrices using the testing datasets are expected to be less accurate, because these represent novel reaches that were not used to build the model. Final model selection is informed by performance on the test dataset, which is more indicative of a model's ability to perform on novel data.

Figure 15. Detailed performance of the three Unstratified Models (all GIS), (no GIS), and (2 GIS) (see Table 6 for descriptions). The top row provides confusion matrices for the training data and the second row shows confusion matrices on the testing data. Shading of boxes in matrices describe the proportion of samples in each dataset. The third row lists the metrics chosen via Recursive Feature Elimination for each model (with metrics at the top being ranked as more important in terms of contribution to model accuracy). The bottom row lists the accuracy metrics PvIvE and EvALI on the test data.

Figure 16. Detailed performance of the three Northeast Models (all GIS), (no GIS), and (2 GIS) (see Table 6 for descriptions). The top row provides confusion matrices for the training data and the second row shows confusion matrices for the testing data. Shading of boxes in matrices describe the proportion of samples in each dataset. The third row lists the metrics selected by Recursive Feature Elimination for each model (with metrics at the top having greater importance in terms of contribution to model accuracy). The bottom row lists the accuracy metrics PvIvE and EvALI on the test data.

Figure 17. Detailed performance of the three Southeast Models (all GIS), (no GIS), and (2 GIS) (see Table 6 for descriptions). The top row provides confusion matrices for the training data and the second row shows confusion matrices on the testing data. Shading of boxes in matrices describe the proportion of samples in each dataset. The third row lists the metrics chosen via Recursive Feature Elimination for each model (with metrics at the top being ranked as more important in terms of contribution to model accuracy). The bottom row lists the accuracy metrics PvIvE and EvALI on the test data.

2.4.4.3 Selection of the final model

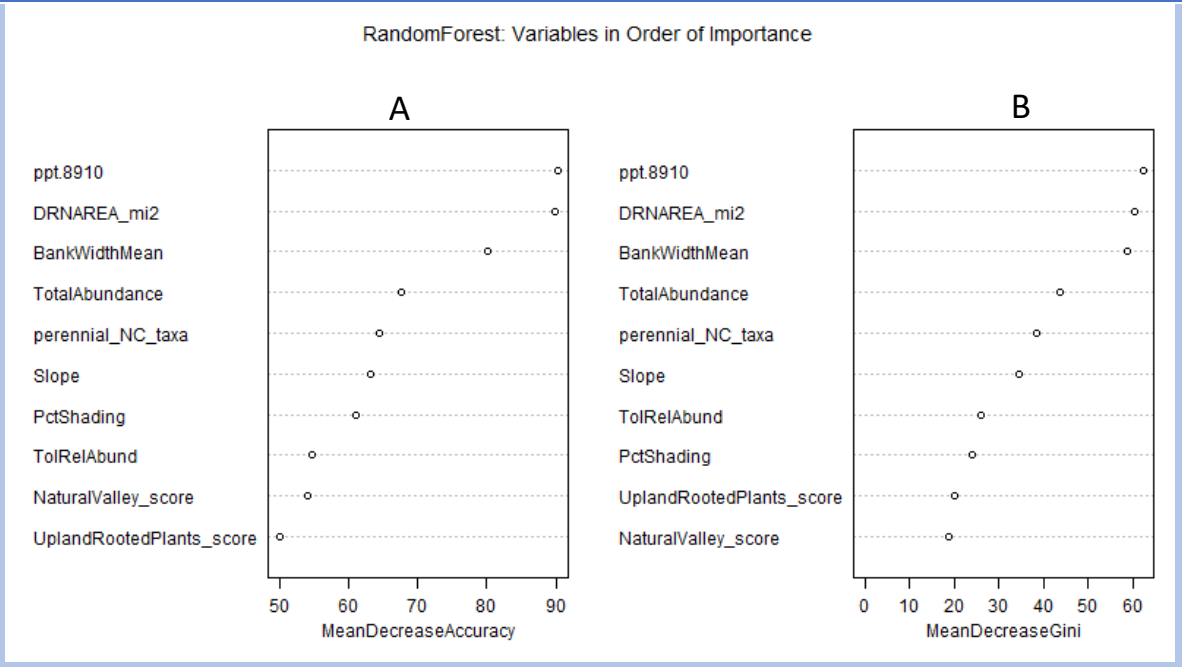
As expected, for all nine model versions performance was highest on the training datasets used to develop the models. Selecting a model whose performance did not vary greatly between the training and testing data was optimal. For example, the Unstratified (all GIS) model was 99.8% accurate in predicting all three flow classes on the training data but only 56.9% accurate on the testing data, indicating that the model was likely overfitting to the training data and unable to perform well on novel reaches (Table 7 and Figure 14).

When the models were permitted to select any number of geospatial metrics (all GIS), the RFE selection process was generally dominated by GIS metrics that were most likely to overfit to the training data. Given this result, and in consultation with the RSC, model selection was constrained to models that had only a limited number of GIS metrics. Based on performance of the stratified versus unstratified models, as well as practical considerations, regionally separated models (i.e., one NE and one SE model) using the 2 GIS version as the base model for each region were selected. These base models were refined further (as explained in the following sections) for improved performance and use as beta SDAMs.

2.4.4.4 NE and SE base model descriptions

The NE (2 GIS) and SE (2 GIS) base models each contained 10 metrics selected via RFE. The metrics are shown in Figure 18 by their order of importance. Here, importance to the random forest model is evaluated in two ways: (A) through mean decrease in accuracy and (B) through mean decrease in Gini Index, which is a measure of node impurity, or how important the metric is in discriminating between different flow duration classes.

NE Base Model (2 GIS) – Metrics Selected



SE Base Model (2 GIS) – Metrics Selected

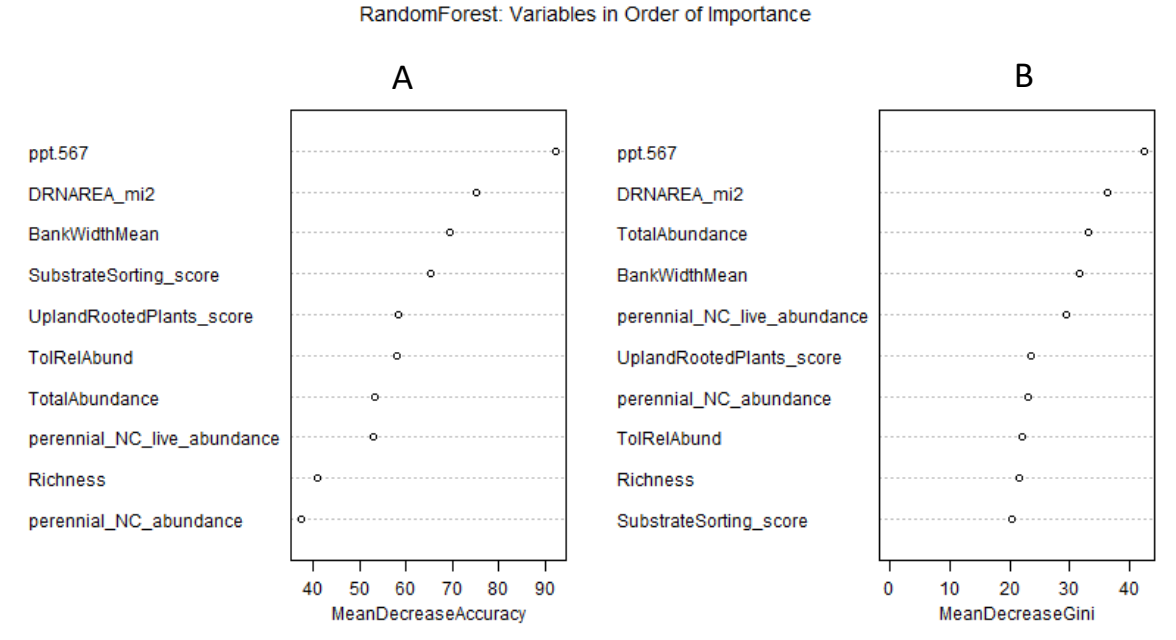


Figure 18. Metrics included in the NE and SE base models, by decreasing order of importance. (A) Mean Decrease in Accuracy is the relative loss in predictive performance when the particular metric is omitted from the model. (B) Mean Decrease in Gini: Gini Index is a measure of node impurity, or how important the metric is in discriminating between different streamflow duration classes.

To evaluate the overall performance of each base model, confusion matrices were created for both training and testing datasets (Figure 19). The highest number of misclassifications in the testing datasets were perennial reaches misclassified as intermittent in the Northeast (n=17) and intermittent reaches misclassified as perennial in the Southeast (n=9). No perennial reach samples were misclassified as ephemeral in either testing datasets; only one ephemeral reach sample was misclassified as perennial in the SE testing dataset.

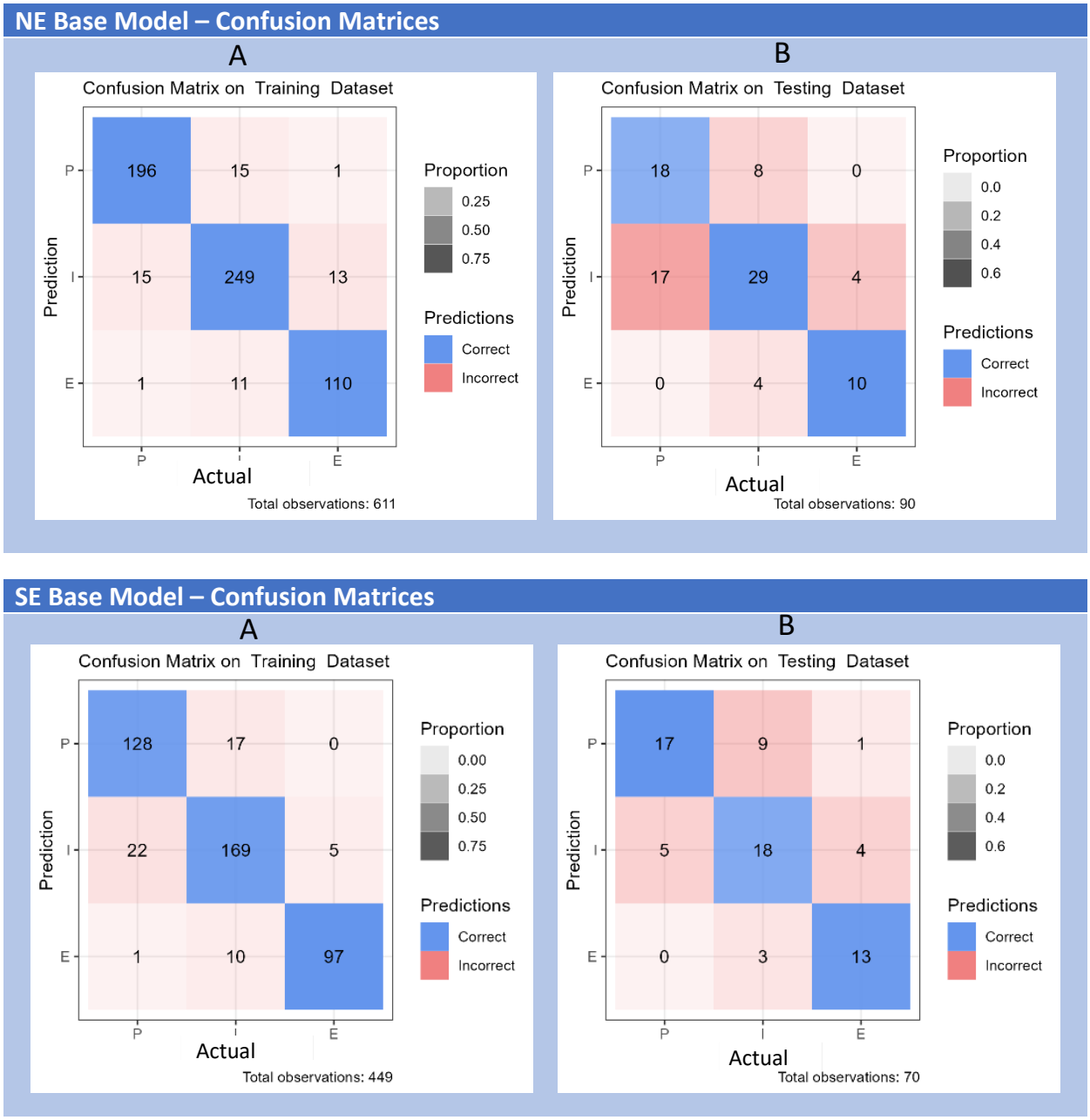


Figure 19. Confusion matrices of the (A) training and (B) testing dataset on the unstratified (2 GIS) model. The training datasets contained a total of 611 (NE) and 449 (SE) samples and the testing datasets contained 90 (NE) and 70 (SE) samples. X-axis shows actual flow duration class and Y-axis shows predicted flow duration class. Blue diagonal indicates correct predictions. P = perennial, I = intermittent, and E = ephemeral. Shading of boxes in matrices describe the proportion of samples in each dataset.

2.4.5 Simplification of the base models

Following selection of the base models for each region, the next step was to simplify each model. Simplification facilitates SDAM field implementation while maintaining or improving model performance. Simplification proceeded in three steps:

1. Refinement of metrics
2. Increased confidence required for classifications
3. Addition of single indicators of *at least intermittent* flow

2.4.5.1 Refinement of metrics

The metric selection process described above identified an optimal set of metrics to use in each SDAM, but it did so without considering difficulties in measuring each metric or effort required to measure all selected metrics. For example, RFE may have selected a metric based on the total number of aquatic invertebrates, even if there was little or no value for model performance provided once more than 40 individuals were recorded. That is, SDAM users might be able to cease counting aquatic invertebrates once 40 individuals were recorded. Refinement of metrics aims to increase SDAM application efficiency and facilitate method use and transparency. Increasing SDAM application efficiency also helps ensure that an SDAM can be applied during a single site visit.

Some metrics were eliminated because they were closely related to another metric in the selected model (i.e., they described similar stream characteristics, such as perennial_NC_abundance and perennial_NC_live_abundance). Metrics that were more time-consuming to measure were replaced if a simpler alternative was available and continuous metrics were converted to binary or ordinal metrics based on visual interpretation of their distributions; a procedure known as binning. (Binary and ordinal metrics are typically more rapid to measure and easier to standardize than continuous metrics.) Accuracy and repeatability measures were re-evaluated to ensure that overall model performance was not substantially diminished by the refinements.

The suite of metrics comprising a selected regional model was iteratively refined while monitoring model accuracy and repeatability. In each iteration, one or more metrics were either eliminated, binned, or otherwise simplified. The impact of each iterative refinement on performance was assessed, and the highest performing refined regional model was selected. Performance was assessed in terms of three accuracy measures: PvIvE (i.e., proportion of reach samples classified correctly as perennial, intermittent, or ephemeral), EvALI (i.e., proportion of reach samples classified correctly as ephemeral or at least intermittent), and IvEdry (i.e., proportion of reach samples classified correctly as intermittent or ephemeral when the reach did not have surface flow).

More than 50 iterative refinements were performed on each of the base models. To illustrate the consequence of these refinements, a subset of key refinements is presented in Figures 20 and 21. For example, a refinement made between Version 0 and Version 1 in the SE model

(Figure 20) was the replacement of perennial_NC_abundance, perennial_NC_live_abundance, TolRelAdbud, and Richness with BMI_score. Other metrics created during the iterative process of manual metric refinement (Figures 20 and 21) are described in Table 8.

Additional combinations of refined and unrefined metrics were attempted during the iterative metric refinement process but are not shown in Figures 20 and 21 for brevity.

Table 8. Metrics Created During the Iterative Process of Manual Metric Refinement

Metric	Model		Description	Revision Version	
	NE	SE		NE	SE
DRNAREA_0.5bin	X		The original continuous metric DRNAREA_mi2 was transformed by binning it into two discrete groups: less than 0.5 mi^2, greater than or equal to 0.5 mi^2	1,3,4,5,6, 7, 8,10	
DRNAREA_0.1bin		X	Transformed DRNAREA_mi2 into two groups: less than 0.1 mi^2, greater than or equal to 0.1 mi^2		8,9
BMI_score	X	X	Original ordinal scoring of benthic macro-invertebrates based on the benthic macro-invertebrate metric used in the North Carolina Method (NCDWQ 2010), see Appendix for more detailed description	2,3,8, 9,10	1, 2, 3, 4, 5, 6, 7, 10
BMI_score_alt1	X		Same scoring as BMI_score with simplified tolerant taxa list	4	
BMI_score_alt2	X		Simplified scoring of BMI_score with original tolerant taxa list	5	
BMI_score_alt3	X	X	Simplified scoring with simplified tolerant taxa list	6	8
BMI_score_alt4	X	X	Simplified scoring of BMI_score without considering relative abundance of tolerant taxa	7	9
PctShad_20_60	X		The original continuous metric Percent Shading was binned into three discrete groups: PctShading less than 0.2, PctShading between 0.2 and 0.6, and PctShaing greater than or equal to 0.6	10	
Slope_10bin	X		Slope less than 10, Slope greater than or equal to 10	9	
Slope_7bin	X		The original Slope metric was separated into two groups: Slope less than 7, Slope greater than or equal to 7	10	
TotalAbundance_0.5bin		X	The original TotalAbundance metric was separated into two groups: less than 0.5, TotalAbundance greater than or equal to 0.5		3
TA_0_10_32_plus		X	TotalAbundance separated into four groups: TotalAbundance equals zero, TotalAbundance between 1 and 10, TotalAbundance between 11 and 32, TotalAbundance 33 or greater		8, 9
TA_0_4_10_plus		X	TotalAbundance separated into four groups: TotalAbundance equals zero, TotalAbundance between 1 and 4, TotalAbundance between 5 and 10, TotalAbundance 11 or greater 10		10
Richness_1bin		X	Richness less than 1, Richness greater than or equal to 1		4
BankWidth_1bin		X	BankWidthMean less than 1, BankWidthMean greater than or equal to 1		6
BankWidth_1.3bin		X	BankWidthMean less than 1.3, BankWidthMean greater than or equal to 1.3		7

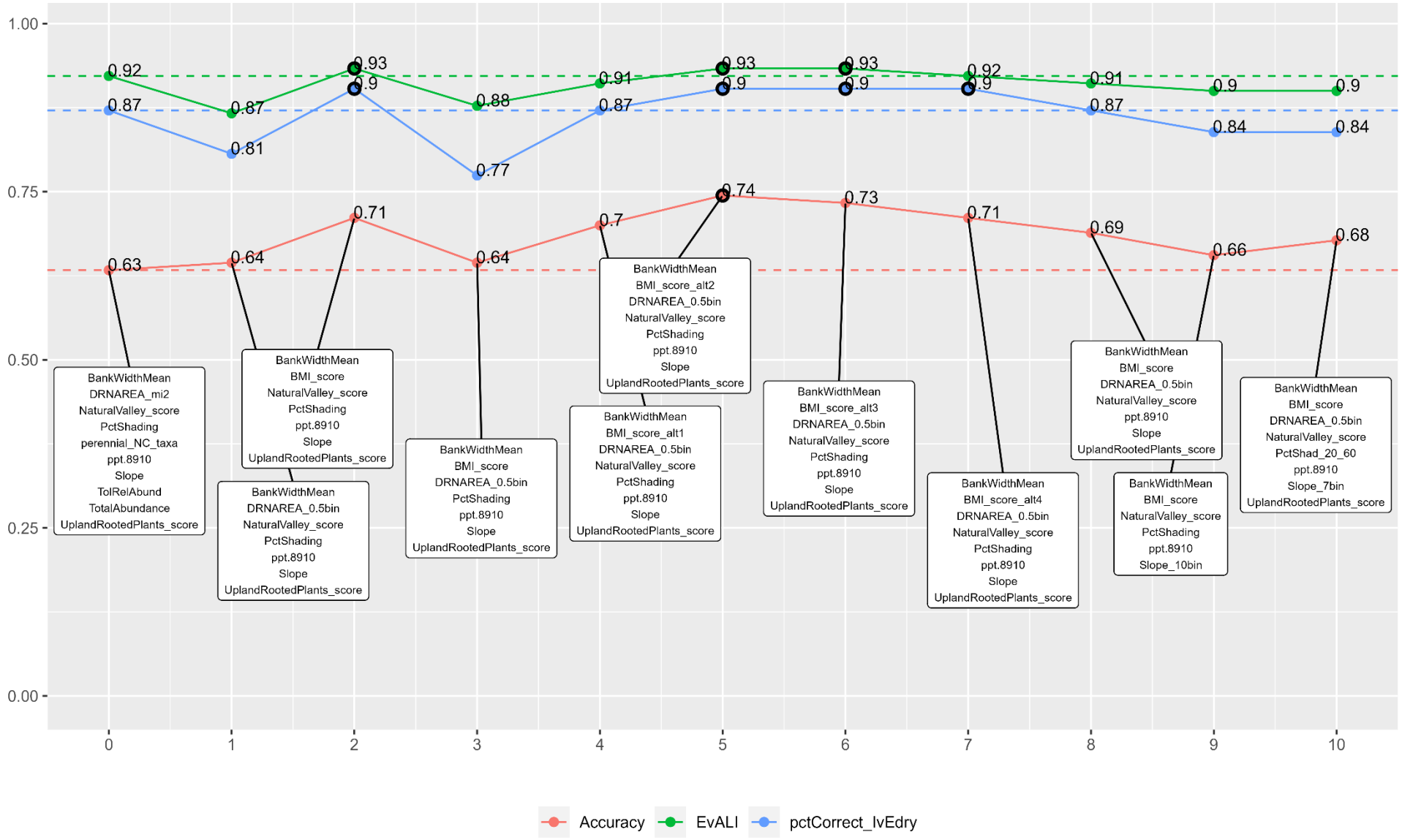


Figure 20. Ten model refinement versions of the NE base model (refinement version 0). Each refinement description is relative to the description for NE base model description. Black circles indicate the highest Accuracy (PvIvE), EvALI, and IvEdry scores. Dashed lines show performance of the NE base model.

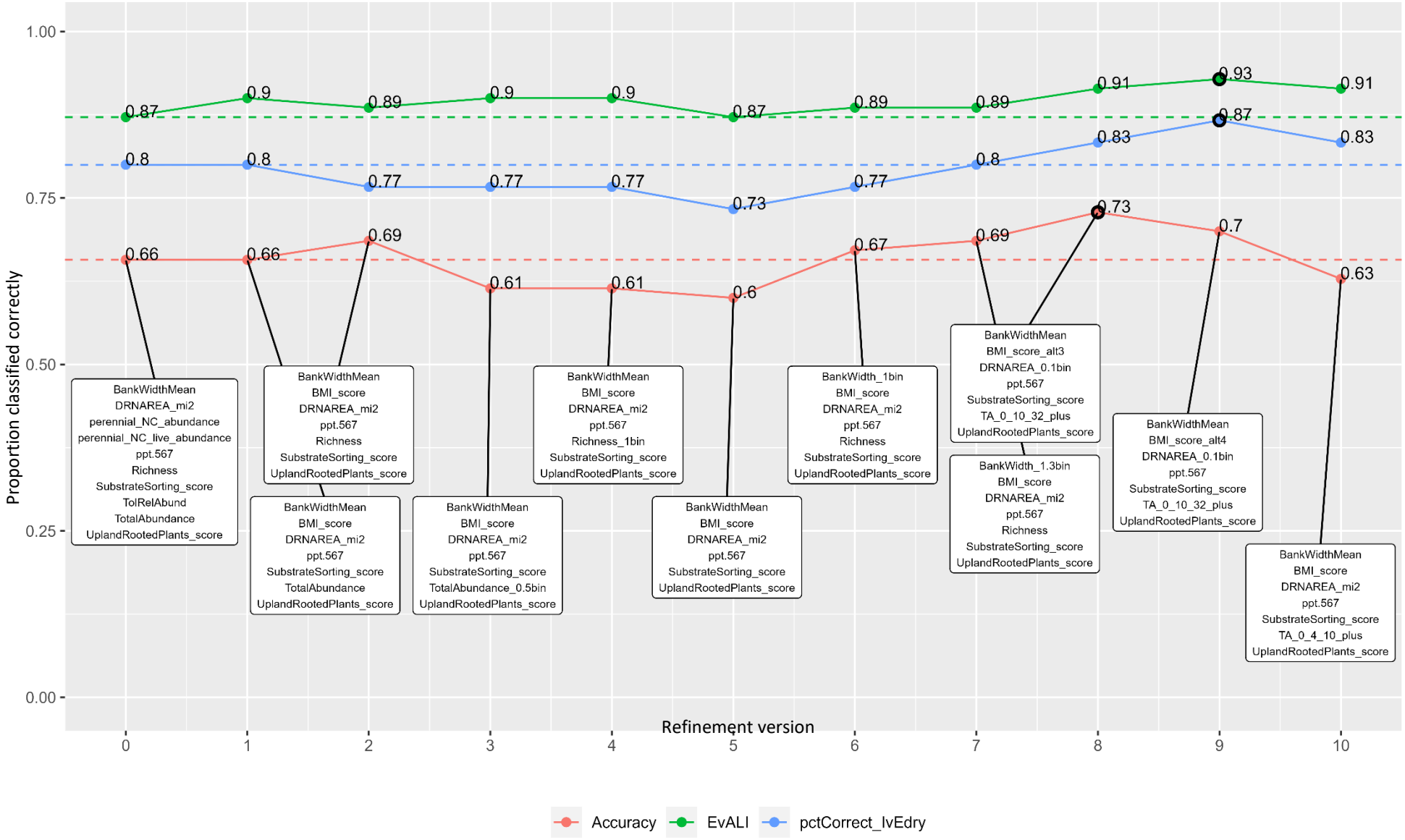


Figure 21. Ten model refinement versions of the SE base model (refinement version 0). Each refinement description is relative to the SE base model description. Black circles indicate the highest Accuracy (PvIvE), EvALI, and IvEdry scores. Dashed lines show performance of the SE base model.

As shown by the performance lines in Figures 20 and 21, some refinements improved the performance of the NE and SE base models. This may be due to binning providing more informative model splits that are less prone to overfitting on the training dataset. Additionally, the refinements were able to replace or modify more time-consuming metrics with simpler alternatives without sacrificing performance.

2.4.6 NE Final beta model selection

Consultation with the RSC resulted in selection of the Version 7 refinement of the NE base model for the Northeast beta model. The Version 7 refinement differs from the NE base model as follows:

- **BMI_score_alt4** (abundance and diversity only): replaced perennial_NC_taxa, TolRelAbund, and Total Abundance.
- **Drainage Area** (<0.5 or greater than 0.5 sq mile): originally a continuous metric ranging from 0.002–396 mi^2, drainage area was binned into discrete groups (less than 0.5 mi^2 and greater than or equal to 0.5 mi^2). These discrete groups were based on visual interpretation of the metric distributions across ephemeral, intermittent, and perennial classes, and through trial-and-error testing.
- **BankWidthMean**: no change
- **Natural Valley score**: no change
- **Percent shading**: no change
- **Average precipitation in August – October**: no change
- **Slope**: no change
- **Upland rooted plants score**: no change

Performance of the final NE refined model (Figure 22) is similar to that of the NE base model (Figure 19).

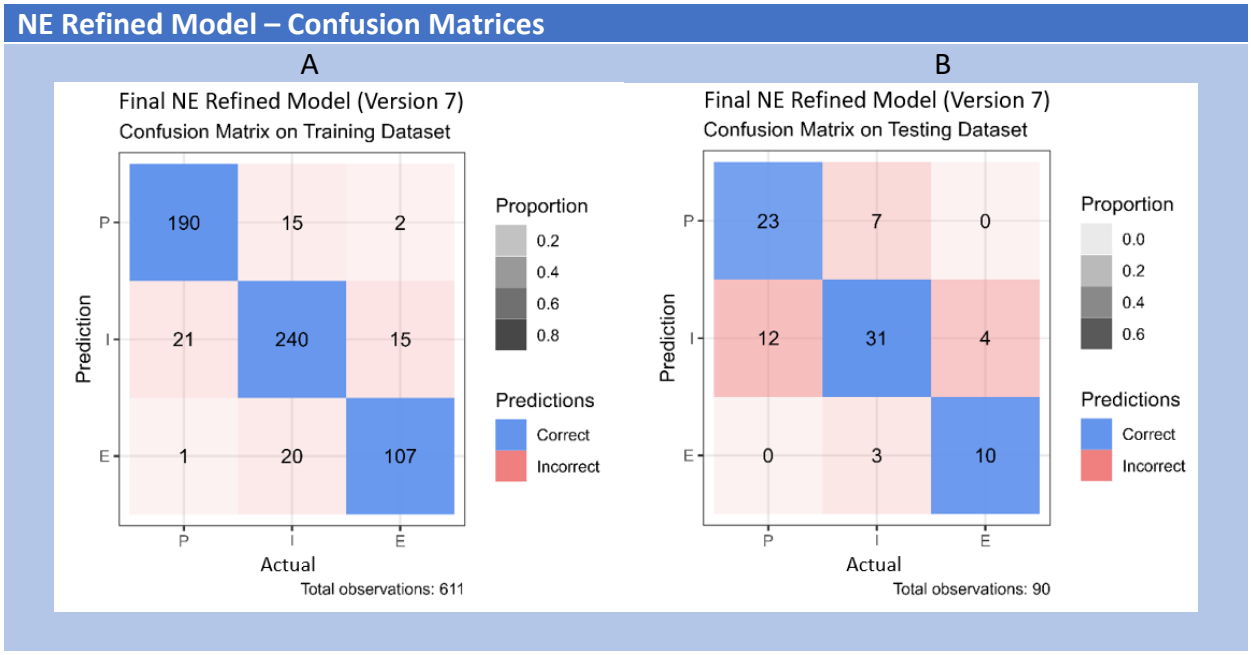


Figure 22. Performance of the final NE refined model based on the (A) training and (B) testing datasets. X-axis shows actual flow duration class and Y-axis shows predicted flow duration class. Blue diagonal indicates correct predictions. P = perennial, I = intermittent, and E = ephemeral. Shading of boxes in matrices describe the proportion of reach samples in each dataset.

Using the NE refined model, one sample (of four visits to site VANE9099_B) in the training dataset continued to incorrectly predict *ephemeral* when the actual classification was *perennial.* In addition, two samples (visits to sites OKNE9440_B and TNNE9109_B) in the training dataset incorrectly predicted *perennial* when the actual classification was *ephemeral*. The sites with these ephemeral-perennial and perennial-ephemeral misclassifications are shown in Table 9. These three misclassified sites are evident in the top-right and bottom-left corners of the confusion matrix shown in panel A of Figure 22. No such misclassifications between perennial and ephemeral classes were evident in the testing dataset (panel B of Figure 22).

Table 9. Perennial-ephemeral and ephemeral-perennial misclassifications for the final refined NE model

Reach Code	State	Region	Dataset	Actual	Predicted
OKNE9440_B	OK	NE	Training	E	P
TNNE9109_B	TN	NE	Training	E	P
VANE9099_B	VA	NE	Training	P	E

No incorrect predictions between *ephemera*l and *perennial* occurred using the final NE refined model on the testing dataset.

2.4.7 SE Final beta model selection

After consultation with the PDT and RSC, the final model selected for the Southeast was the Version 9 refinement of the SE base model. The Version 9 refinement differs from the SE base model as follows:

- **BMI_score_alt4** (abundance and diversity only): replaced perennial_NC_taxa, perennial_NC_live_taxa, and TolRelAbund.
- **Drainage Area** (<0.1 or greater than 0.1 mi^2): originally a continuous metric ranging from 0.00283–579 mi^2, drainage area was binned into discrete groups (less than 0.1 mi^2 and greater than or equal to 0.1 mi^2). These discrete groups were based on visual interpretation of the metric distributions across ephemeral, intermittent, and perennial classes, and through trial-and-error testing.
- **Total abundance of BMI**: Originally a continuous metric whose counts ranged from 0 to 105 in the SE region, Total Abundance was binned into the following discrete groupings:
    - Total Abundance <1
    - Total Abundance between 1 and 10
    - Total Abundance between 11 and 32
    - Total Abundance >33

  These discrete groups were based on visual interpretation of the metric distributions across ephemeral, intermittent, and perennial classes, and through trial-and-error testing.
- **BankWidthMean**: no change
- **Average precipitation in May – July**: no change
- **Substrate sorting score**: no change
- **Upland rooted plants score**: no change

Performance of the final SE refined model (Figure 23) is similar to that of the SE base model (Figure 19).

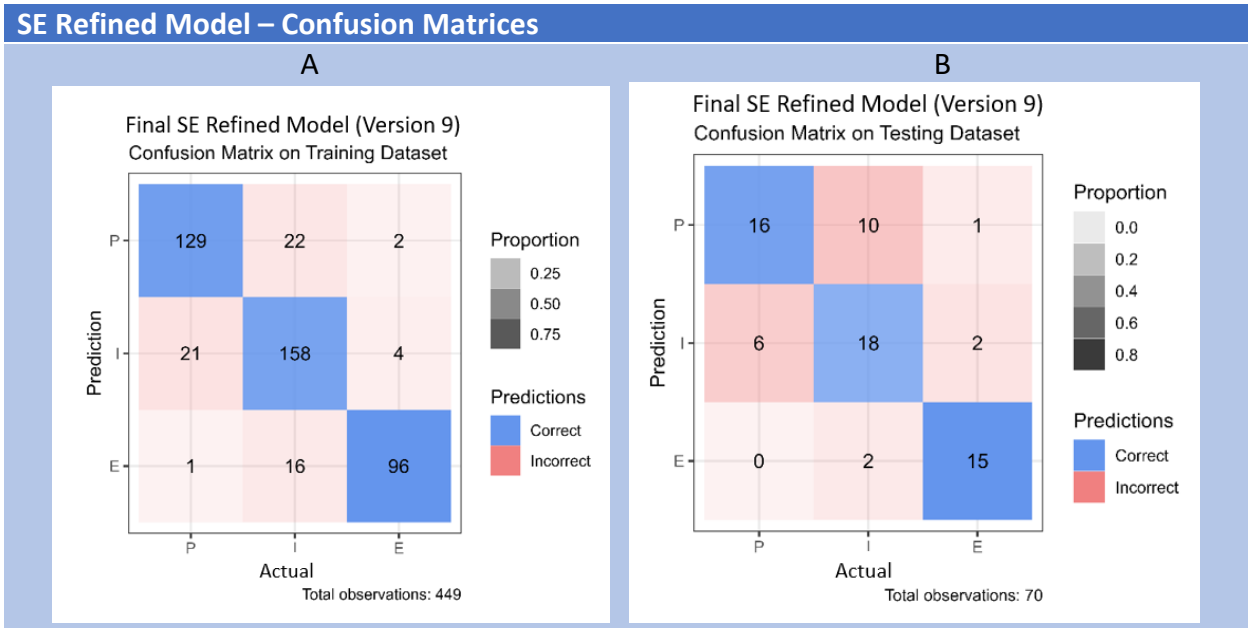


Figure 23. Performance of the selected final refined SE model based on the (A) training and (B) testing datasets. X-axis shows actual flow duration class and Y-axis shows predicted flow duration class. Blue diagonal indicates correct predictions. P = perennial, I = intermittent, and E = ephemeral. Shading of boxes in matrices describe the proportion of reach samples in each dataset.

Using the final refined SE model, one sample (site ALSE8675_B) in the training dataset continued to incorrectly predict *ephemeral* when the actual classification was *perennial* (as highlighted in the top-right and bottom-left corners of the confusion matrices in Figure 23). In addition, two reaches in the training dataset incorrectly predicted *perennial* when the actual classification was *ephemeral*. This misclassification also occurred in one of the samples in the SE testing dataset. The sites with these ephemeral-perennial and perennial-ephemeral misclassifications are shown in Table 10.

Table 10. Perennial-ephemeral and ephemeral-perennial misclassifications for the final refined SE model

Reach Code	State	Region	Dataset	Actual	Predicted
ALSE8675_B	AL	SE	Training	P	E
FLSE9509_B	FL	SE	Training	E	P
TXSE9496_BO	TX	SE	Training	E	P
TXSE9467_V	TX	SE	Testing	E	P

2.4.7.1 Increased confidence required for classifications

Random forest models created for classification traditionally make assignments based on the class that receives the highest number of votes by each tree in the "forest." Thus, in a three-way decision (ephemeral, intermittent, or perennial), the class with the most votes could receive much less than a majority of all votes—as low as 34%. Given concern that such low-confidence classifications may not provide sufficient defensibility for some management

decisions, approaches to distinguish between high- and low-confidence classifications were explored.

Increasing the minimum number of votes required to make a confident classification from 33% to 100% by increments of 1% was explored to understand the effect on classification. When the selected refined model was applied to a novel test reach and a single class received a sufficient percent of votes, then the reach was classified accordingly. If none met the minimum but the combined percent of votes for intermittent and perennial classes exceeded the minimum, then the reach was classified as *at least intermittent*. In all other cases, the reach was classified as *need more information*. This decision framework reflects that distinguishing between ephemeral and at least intermittent reaches is a high priority use of the beta SDAMs for the NE and SE. The percent of reaches under each of the five possible classifications with increasing minimum vote agreement thresholds was calculated.

The minimum proportion threshold of 0.5 was set for flow classification, which is the same threshold as used in the AW, WM, and GP beta methods. At the minimum required proportion of votes of 0.5 (or 50%) in the final refined NE model, only 4.3% of reach samples in the training dataset (1.8% of reach samples in the test dataset) were classified as *at least intermittent*, and none were classified as *need more information* (Figure 24A, B). At a minimum required proportion of votes of 0.5 in the final refined SE model, 2.0% of reach samples in the training dataset (8.6% of reach samples in the test dataset) were classified as *at least intermittent*, and none were classified as *need more information* (Figure 24C, D). Classifications of *at least intermittent* first appear with a minimum proportion of 0.39 (NE) and 0.35 (SE) in the training datasets and 0.43 (NE) and 0.36 (SE) in the testing datasets, whereas classifications of *need more information* appear at 0.51 in both models. Although it cannot be ruled out, it is unlikely that the beta SDAMs for the NE and SE will result in a classification of *need more information.*

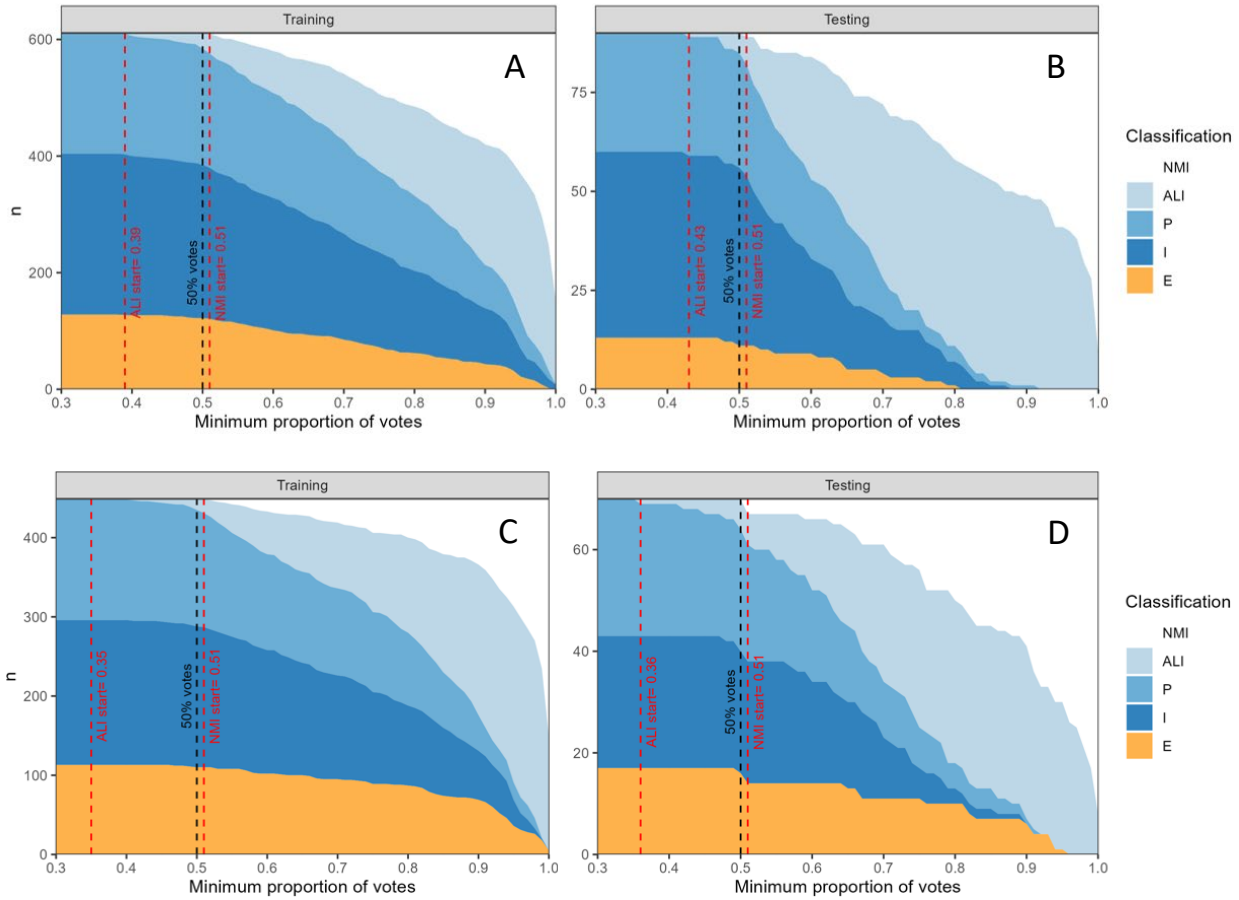


Figure 24. Influence of the minimum proportion of votes required to make a classification on n (the number of reaches in each class) for the (A) final refined Northeast model training data, (B) final refined Northeast model testing data, (C) final refined Southeast model training data, and (D) final refined Southeast model testing data. NMI: Need more information. ALI: At least intermittent. P: Perennial. I: Intermittent. E: Ephemeral. The vertical black line represents a minimum proportion of required votes of 0.5. The two red lines represent the proportion of votes that first result in classification of ALI (the lower line) or NMI (the upper line) for the datasets.

Note, after updating the voting threshold to a requirement of 50% for classification, there were some changes to the perennial-ephemeral and ephemeral-perennial misclassifications. Three of these errors, all of which occurred in the training dataset, were changed from their original prediction to *At Least Intermittent*. For the case of reach perennial ALSE8675_B, this caused an originally incorrect *Ephemeral* prediction of to be updated to a correct prediction of *At Least Intermittent* (highlighted in green in Table 10, below). For the other two reaches (OKNE9440_B and FLSE9509_B) whose predictions were changed by the updated voting threshold, their originally incorrect *Perennial* predictions were changed to *At Least Intermittent* (highlighted in yellow in Table 11, below).

Table 11. Updated "Big Error" classifications after increased vote threshold (50%).

Reach Code	State	Region	Dataset	Actual Class	Original Prediction (Majority Voting)	Updated Prediction (50% Threshold)
OKNE9440_B	OK	NE	Training	E	P	ALI
TNNE9109_B	TN	NE	Training	E	P	P
VANE9099_B	VA	NE	Training	P	E	E
ALSE8675_B	AL	SE	Training	P	E	ALI
FLSE9509_B	FL	SE	Training	E	P	ALI
TXSE9496_BO	TX	SE	Training	E	P	P
TXSE9467_V	TX	SE	Testing	E	P	P

2.4.7.2 Evaluation of single indicators of at least intermittent flow

Single indicators can supersede a model classification of *ephemeral* to change it to the classification of *at least intermittent*. Single indicators can provide technical benefits (i.e., improved accuracy) as well as non-technical benefits, such as rapidity of determining flow duration and greater acceptance of the SDAM, given existing public understanding of, for example, the role of streamflow duration in supporting biological organisms. Single indicators are also used in some other SDAMs (e.g., Nadeau et al. 2015, Dorney and Russell 2018, Mazor et al. 2021a); for instance, the presence of fish, iron-oxidizing bacteria, hydric soils, and/or aquatic vertebrates (amphibians and reptiles), among others.

Single indicators used in previous SDAMs were evaluated. The number of instances where inclusion of a prior single indicator would correct a misclassification (i.e., the reach was truly intermittent or perennial) and would introduce a misclassification/mistake (i.e., the reach was truly ephemeral) was quantified. All single indicators investigated had minimal impact on performance or introduced more errors than were corrected (Figure 25). Based on these results, the RSC did not recommend including any of the evaluated single indicators in the beta SDAMs NE and SE.

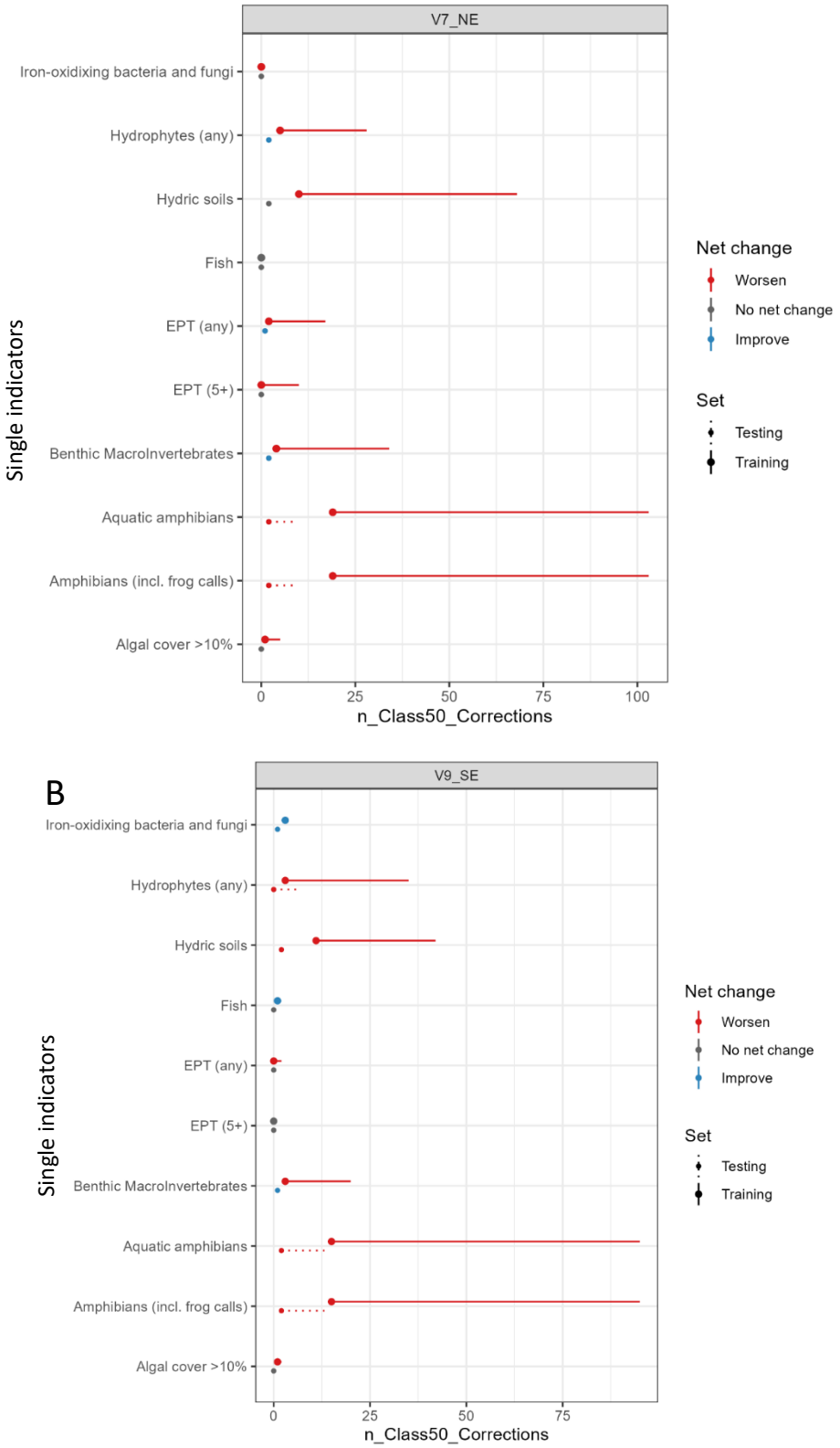


Figure 25. Influence of single indicators on (A) final NE refined model performance and (B) final SE refined model performance.

2.4.8 Performance of the Beta SDAMs NE and SE

Performance of the selected refined models after assigning a minimum proportion voting threshold of 50% for the beta SDAMs NE and SE are summarized in Table 12. The overall classification accuracy among the three classes (*perennial*, *intermittent*, *ephemeral*) for the NE model was 99% in the training dataset (and 72.2% in the testing dataset), but this accuracy increased to 99.5% in the training dataset (and 92.2% in the testing dataset) when only *ephemeral* versus at *least intermittent* classifications were considered (i.e., both blue and green cells in Table 12 were treated as correct). The overall classification accuracy among the three classes for the SE model was 98% in the training dataset (and 70% in the testing dataset), but this accuracy increased to 98% in the training dataset (and 91.4% in the testing dataset) when only *ephemeral* versus at *least intermittent* classifications were considered.

Table 12. Classifications of the final version of the beta SDAM NE and SE. Blue cells indicate correct classifications of perennial, intermittent, at least intermittent, and ephemeral reaches, whereas green cells indicate correct classifications of ephemeral versus at least intermittent. Green numbers represent the reach visits with matching actual and predicted classes and red numbers are reach visits with non-matching actual and predicted classes.

NE Model: Actual streamflow duration class (augmented data not included)

Predicted Class	Ephemeral (Training)	Ephemeral (Testing)	Intermittent (Training)	Intermittent (Testing)	Perennial (Training)	Perennial (Testing)
Ephemeral	87	9	1	2	0	0
Intermittent	0	4	194	29	0	12
ALI	1	1	0	4	2	0
Perennial	0	0	0	6	147	23

Total NSE samples: 522

SE Model: Actual streamflow duration class (augmented data not included)

Predicted Class	Ephemeral (Training)	Ephemeral (Testing)	Intermittent (Training)	Intermittent (Testing)	Perennial (Training)	Perennial (Testing)
Ephemeral	61	14	4	2	0	0
Intermittent	0	2	141	16	1	6
ALI	0	1	0	4	0	1
Perennial	1	1	0	8	116	15

Total SE samples: 394

2.5 Disturbed Sites

Using the LandUse indicator to identify reaches that were disturbed (LandUse = urban or agriculture, alone or in combination with any other land use category) and not disturbed (LandUse does not include urban or agriculture) at the time of the site visit, there were 37 individual reaches identified as disturbed during at least one site visit with a total of 55 disturbed samples (before augmentation) in the Northeast Region. There were 45 disturbed

samples in the training dataset and 10 in the testing dataset. These tallies focus on the samples of the original dataset before augmentation (n = 522 NE).

Among the samples identified as disturbed by human activity in the Northeast testing dataset (n=10), accuracy among all classes was 60%, which improved to 90% when only *ephemeral* versus *at least intermittent* classifications were considered. For samples in the Northeast testing dataset that were not disturbed (n=80), the accuracy values of the disturbed sites were 74% PvIvE and 93% EvALI.

For the Southeast dataset, the LandUse indicator flagged 16 individual reaches identified as disturbed during at least one site visit with a total of 28 disturbed samples (before augmentation). There were 26 disturbed samples in the training dataset and two in the testing dataset. These tallies focus on the samples of the original dataset before augmentation (n = 394 SE).

Among the samples identified as disturbed by human activity in the Southeast testing dataset, (n=2) accuracy among all classes was 50%, which did not change when only *ephemeral* versus *at least intermittent* classifications were considered. For samples in the Southeast testing dataset that were not disturbed (n=68), the accuracy values were 72% PvIvE and 93% EvALI.

3 Performance of beta SDAMs NE and SE against other methods

For reference, a comparison of the results using nearby SDAMs (the Ohio, North Carolina, and beta Great Plains methods) are included in this section (Table 13). To apply the beta GP method, a metric called "Strata" was created for the NE and SE datasets so that the GP Strata was "Northern" for the NE region and "Southern" for the SE region. The following analysis does not include oversampling for the training datasets; for sites in the NE the Northern GP was selected and for sites in the SE the Southern GP was selected when applying the beta SDAM for the GP.

Table 13. Comparing performance of NE and SE beta SDAMs to the OH, NC, and beta GP SDAMs

			OH	NC		Beta GP		Beta NE or SE	
Dataset	**Region**	**# samples**	**% EvALI Correct**	**% PvIvE Correct**	**% EvALI Correct**	**% PvIvE Correct**	**% EvALI Correct**	**% PvIvE Correct**	**% EvALI Correct**
Testing	SE	70	90%	67%	94%	80%	93%	71%	91%
Testing	NE	90	86%	51%	90%	53%	90%	72%	92%
Testing	CB	33	79%	64%	88%	55%	82%	42%	73%

3.1 Performance of the beta SDAM SE using the U.S. Caribbean data

While there were not enough samples (19 reaches, 33 site visits) to develop a U.S. Caribbean-specific model, the performance of the final refined SE model using the U.S. Caribbean data as a novel dataset was assessed (Figure 26).

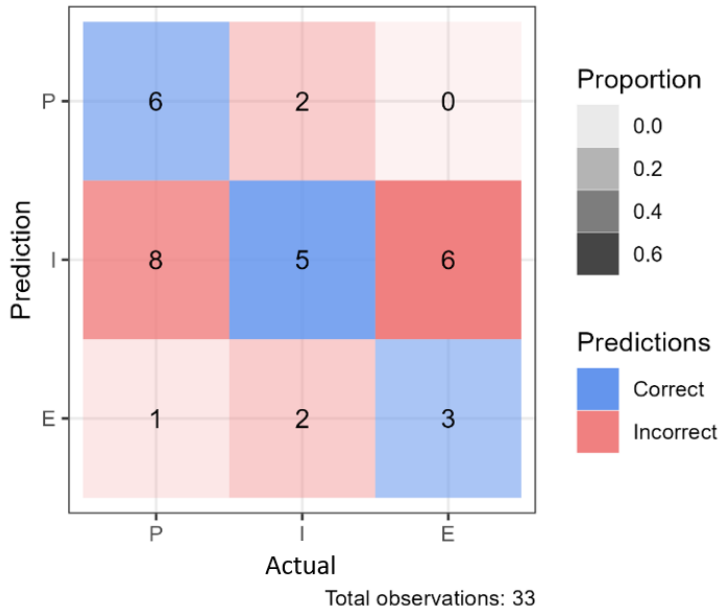


Figure 26. Confusion matrix showing performance of U.S. Caribbean data (33 samples from 19 reaches) using the final refined SE model.

The accuracy of the refined SE model when applied to the data collected at sites in the U.S. Caribbean was 42% correct PvIvE (increased to 73% correct for EvALI). The following "Big Error" was noted for the Caribbean dataset:

Reach Code	State	Region	Dataset	Actual	Predicted
PRSE9540_B	PR	CB	Testing	P	E

4 Data and code availability

All data used to develop the method and R code used in analysis are available at the following repository: https://doi.org/10.23719/1528743.

5 Next steps

The beta SDAMs for the NE and SE are being made available for one year for public review and comment while the additional data at the study sites collected in 2022 and 2023 are processed, after which final methods will be developed and released to replace the beta methods.

6 Acknowledgements

The development of this method and supporting materials was guided by a national and regional steering committee (RSC) consisting of representatives of federal regulatory agencies in the Northeast Southeast of the U.S.: Erica Sachs and Raymond Putnam (USEPA – Region1), Stephanie Tougas, Marco Finocchiaro, Stephanie Andreescu, and Nancy Rodriguez (USEPA –

Region 2), Christine Mazzarella and Greg Pond (USEPA – Region 3), Eric Somerville (USEPA – Region 4), Melanie Burdick (USEPA – Region 5), Chad LaMontagne and Kamren Metzger (USACE St. Louis District), Roger Allan and Damon McDermott (USACE Memphis District), Aric Payne, Mark McIntosh, and William Sinclair (USACE Nashville District), Patti Grace-Jarrett, Russell Retherford, and Sam Werner (USACE Louisville District), Joseph Shelnutt and Sabrina Miller (USACE Fort Worth District), Robert Hoffmann (USACE Tulsa District), Peter Krakowiak (USACE Buffalo District), Gabrielle C. L. David (USACE—Engineer Research and Development Center, Cold Regions Research and Engineering Laboratory), Rose Kwok (USEPA—Headquarters), Tunis McElwain (USACE—Headquarters), and Matt Wilson (USACE—Headquarters).

We thank Abel Santana, Robert Butler, Duy Nguyen, Kristine Gesulga, Kenneth McCune, Adriana LeCompte-Santiago, Will Saulnier and Anne Holt for assistance with data management and web application development. We thank Megan Annis, Jackson Bates, Joe Bertherman, Emma Duguay, Brian Emlaw, Zak Erickson, Hannah Erickson, Heidi Fisher, Kate Forsmark, Richard Judge, Kort Kirkeby, Alec Lambert, Libby Lee, Claire Leedy, Bryan Lees, Mindi Lundberg, Abe Margo, Buck Meyer, Margaret O'Brien, Addison Ochs, Jake Okun, Jack Poole, Morgan Proko, Chris Roche, Olivia Shaw, Chelsey Sherwood, Craig Smith, Ali Sutphin, Alex Swain, James Treacy, Charlie Waddell, and Jeff Weaver for assistance with data collection. Stephanie Andreescu, Raul Gutierrez, Tamara Heartsill-Scalley, Nolan Hahn, Jeffery Lapp, Todd Lutte, Robert Montgomerie, Sofia Olivero Lora, Kathryn Quesnell, Jose Soto, and Cynthia Van Der Wiele assisted with plant identification.

Numerous researchers and land managers with local expertise assisted with the selection of study reaches to calibrate the method: Susie Adams, Laurie Alexander, Dan Allen, Carla Atkinson, Brent Aulenbach, Debbie Arnwine, Scott Bailey, Joe Bartlett, Mary Becker, Taylor Bell, Sean Beyke, Emery Boose, Rick Chormann, Joshua Clemmons, Matt Cohen, Shannon Curtis, Daniel Dauwalter, Daragh Deegan, Janet Dewey, John Dorney, Jon Duncan, Bob Easter, Mike Fargione, Jacob Ferguson, Brock Freyer, Bill Gawley, Cynthia Gilmour, Natalie Griffith, Kevin Grimsley, Brandon Hall, Steve Hamilton, Russell Hardee, Andy Harrison, Blaine Hastings, Tamara Heartsill, Katy Hofmeister, Darrin Hunt, Jeremiah Jackson, Rhett Jackson, Allan James, Carrie Jenson, Nate Jones, Tom Jordan, Jeanine Lackey, Bryan Lees, Mike Lott, Joshua Keeley, Sean Kelly, Vicky Kelly, Dan Marion, Jason Martin, Gustavo Martinez, Bruce Means, Carl Neilson, Jules NeSmith, Jami Nettles, C. Nicholas, Greg Pond, Kai Rains, Carlos Ramos-Scharron, Jamie Robb, Mary Rocky, Carlos Rodriguez, Randy Sarver, Kim Sash, Kristen Selikoff, Stephanie Siemke, Knight Silas Cox, Chelsea Smith, Doug Smith, Eric Snyder, Tedmund Soileau, Matthew Stahman, Emily Stephan, Carl Trettin, Ross Vander Vorste, Robert Voss, Peter Wampler, Glenn Wilson, Brandon Yates, Shawyn Yeamans, and Margaret Zimmer.

7 References

Breiman, L. 2001. Random forests. Mach. Learn 45: 5–32.

Cao, Y., and C. P. Hawkins. 2011. The comparability of bioassessments: a review of conceptual and methodological issues. Journal of the North American Benthological Society 30: 680–701.

Chapin, T. P., A. S. Todd, and M. P. Zeigler. 2014. Robust, low-cost data loggers for stream temperature, flow intermittency, and relative conductivity monitoring. Water Resources Research 50: 6542–6548.

Cutler, D.R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88: 2783–2792. https://doi.org/10.1890/07-0539.1.

Dorney, J., and P. Russell. 2018. North Carolina Division of Water Quality methodology for identification of intermittent and perennial streams and their origins. Pages 273–279 in J. Dorney, R. Savage, R. W. Tiner, and P. Adamus (eds.), Wetland and Stream Rapid Assessments. Elsevier, San Diego, CA.

Eddy M., S. Gross, K. M. Fritz, T. -L. Nadeau, B. Topping, R. Fertik Edgerton, and J. Kelso. 2022. Development and Evaluation of the Beta Streamflow Duration Assessment Method for the Great Plains. Document No. EPA-840-R-22003.

Ellis, N., S. J. Smith, and C. R. Pitcher. 2012. Gradient forests: calculating importance gradients on physical predictors. Ecology 93: 156–168. https://doi.org/10.1890/11-0252.1.

Eng, K., D. M. Wolock, and M. D. Dettinger. 2016. Sensitivity of intermittent streams to climate variations in the USA. River Research Applications 32: 885–895.

Fritz, K. M., B. R. Johnson, and D. M. Walters. 2008. Physical indicators of hydrologic permanence in forested headwater streams. Journal of the North American Benthological Society 27: 690-704.

Fritz, K. M., T. -L. Nadeau, J. E. Kelso, W. S. Beck, R. D. Mazor, R. A. Harrington, and B. J. Topping. 2020. Classifying streamflow duration: The scientific basis and an operational framework for method development. Water 12: 2545.

Gower, J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27: 857-874.

Hammond, J. C., M. Zimmer, M. Shanafield, K. Kaiser, S. E. Godsey, M. C. Mimms, S. C. Zipper, R. M. Burrows, S. K. Kampf, W. Dodds, C. N. Jones, C. A. Krabbenhoft, K. S. Boersma, T. Datry, J. D. Olden, G. H. Allen, A. N. Price, K. Costigan, R. Hale, A. S. Ward, and D. C. Allen. 2021. Spatial patterns and drivers of nonperennial flow regimes in the contiguous United States. Geophysical Research Letters 48: e2020GL090794.

Hart, E., and K. Bell. 2015. Prism: Access Data From The Oregon State Prism Climate Project.

Hawkins, C. P., Y. Cao, and B. Roper. 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology 55: 1066–1085.

Hedman, E. R., and W. R. Osterkamp. 1982. Stream Flow Characteristics Related to Channel Geometry of Streams in Western United States. USGS Water-Supply Paper 2193, Washington, DC. p. 17. DOI:10.3133/wsp2193.

Hewlett, J. D. 1982. Principles of Forest Hydrology; University of Georgia Press: Athens, GA, USA, p. 192.

Jaeger, K.L., R. Sando, R. R. McShane, J. B. Dunham, D. P. Hockman-Wert, K. E. Kaiser, K. Hafen, J. C. Risley, and K. W. Blasch. 2019. Probability of streamflow permanence model (PROSPER): a spatially continuous model of annual streamflow permanence throughout the Pacific Northwest. Journal of Hydrology X 2: 1000005.

James, A., K. McCune, and R. Mazor. 2022. Review of Flow Duration Methods and Indicators of Flow Duration in the Scientific Literature, Northeast Southeast of the United States. Document No. EPA-840-B-22006. 56 pp. (Available from: https://www.epa.gov/system/files/documents/2023-04/Literature-Review-Beta-SDAM-NE-and-SE.pdf)

James, A., Nadeau, T.-L., Fritz, K.M., Topping, B., Fertik Edgerton, R., Kelso, J., Mazor, R., and Nicholas, K. 2023. User Manual for Beta Streamflow Duration Assessment Methods for the Northeast and Southeast of the United States. Version 1.0. Document No. EPA-843-B-23001.

James City County, VA. 2009. Perennial Stream Protocol Guidance Manual. (Available from: https://jamescitycountyva.gov/DocumentCenter/View/2158/JCC-Perennial-Stream-Protocol-Manual-PDF?bidId=)

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An Introduction to Statistical Learning. Springer, NY. 440 pp.

Kelso, J. E., and K. M. Fritz. 2021. Standard Operating Procedure: Processing Data and Classifying Streamflow Duration Using Continuous Hydrologic Data. EPA Report J-WECD-ECB-SOP-4425-0. Environmental Protection Agency, Cincinnati, OH. 25 pp.

Kuhn, M. 2020. caret: Classification and Regression Training. (Available from: https://cran.r-project.org/web/packages/caret/caret.pdf)

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2: 18–22.

Mazor, R. D., A. C. Rehn, P. R. Ode, M. Engeln, K. C. Schiff, E. D. Stein, D. J. Gillett, D. B. Herbst, and C. P. Hawkins. 2016. Bioassessment in complex environments: designing an index for consistent meaning in different settings. Freshwater Science 35: 249–271.

Mazor, R. D., B. J. Topping, T.-L. Nadeau, K. M. Fritz, J. E. Kelso, R. A. Harrington, W. S. Beck, K. McCune, H. Lowman, A. Aaron, R. Leidy, J. T. Robb, and G. C. L. David. 2021a. User Manual for a Beta Streamflow Duration Assessment Method for the Arid West of the United States. Version 1.0. Document No. EPA 800-K-21001. 83 pp. (Available from: https://www.epa.gov/sites/production/files/2021-03/documents/user_manual_beta_sdam_aw.pdf)

Mazor, R. D., B. J. Topping, T.-L. Nadeau, K. M. Fritz, J. E. Kelso, R. A. Harrington, W. S. Beck, K. S. McCune, A. O. Allen, R. Leidy, J. T. Robb, and G. C. L. David. 2021b. Implementing an operational framework to develop a streamflow duration assessment method: A case study from the Arid West United States. Water 13: 3310.

Mazor, R. D., B. J. Topping, T.-L. Nadeau, K. M. Fritz, J. E. Kelso, R. A. Harrington, W. S. Beck, K. McCune, A. Allen, R. Leidy, J. T. Robb, G. C. L. David, and L. Tanner. 2021c. User Manual for a Beta Streamflow Duration Assessment Method for the Western Mountains of the United States. Version 1.0. Document No. EPA-840-B-21008. 116 pp. (Available from: https://www.epa.gov/system/files/documents/2021-12/beta-sdam-for-the-wm-user-manual.pdf)

Mazor, R. D., K. M. Fritz, B. Topping, T. -L. Nadeau, and J. Kelso. 2022. Development and Evaluation of the Beta Streamflow Duration Assessment Method for the Western Mountains – Data Supplement. Document No. EPA 840-R-22002. 38 pp. (Available from: https://www.epa.gov/system/files/documents/2022-05/WM%20Data%20supplement_5-4-22%20FINAL.pdf)

Mohammed, R., J. Rawashdeh, and M. Abdullah. 2020. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. Pages 243-248 in Proceedings of the 11th International Conference on Information and Communication Systems. Irbid, Jordan 7-9 April 2020.

Nadeau, T.-L. 2015. Streamflow Duration Assessment Method for the Pacific Northwest. EPA 910-K-14-001, U.S. Environmental Protection Agency. 36 pp. (Available from: https://www.epa.gov/sites/default/files/2016-01/documents/streamflow_duration_assessment_method_pacific_northwest_2015.pdf)

Nadeau, T.-L., S. G. Leibowitz, P. J. Wigington, J. L. Ebersole, K. M. Fritz, R. A. Coulombe, R. L. Comeleo, and K. A. Blocksom. 2015. Validation of rapid assessment methods to determine streamflow duration classes in the Pacific Northwest, USA. Environmental Management 56: 34–53.

NC Division of Water Quality (NCDWQ). 2010. Methodology for Identification of Intermittent and Perennial Streams and their Origins (Version 4.11). North Carolina Department of Environment and Natural Resources. (Available from: http://portal.ncdenr.org/web/wq/swp/ws/401/waterresources/streamdeterminations)

New Mexico Environment Department (NMED). 2011. Hydrology Protocol for the Determination of Uses Supported by Ephemeral, Intermittent, and Perennial Waters. Surface Water Quality Bureau, New Mexico Environment Department, Albuquerque, NM. 35 pp. (Available from: https://www.env.nm.gov/surface-water-quality/wp-content/uploads/sites/25/2019/11/WQMP-CPP-Appendix-C-Hydrology-Protocol-20201023-APPROVED.pdf)

Ohio EPA. 2020. Field Methods for Evaluating Primary Headwater Streams in Ohio (Version 4.1*)*. Ohio EPA, Division of Surface Water. (Available from https://epa.ohio.gov/static/Portals/35/credibledata/PHWHManual_2020_Ver_4_1_May_2020_Final.pdf)

Omernik, J.M. 1995. Ecoregions: a framework for managing ecosystems. The George Wright Forum 12: 35–50.

Perkin, J. S., K. B. Gido, J. A. Falke, K. D. Fausch, H. Crockett, E. R. Johnson, and J. Sanderson. 2017. Groundwater declines are linked to changes in Northeast Southeast stream fish assemblages. Proceedings of the National Academy of Sciences USA 114: 7373–7378.

Schumacher, C., and K. M. Fritz. 2019. Standard Operating Procedure: Verifying/Calibrating, Deploying, Retrieving Stream Temperature, Intermittency, and Conductivity (STIC) Data Loggers, and Downloading and Converting Data. EPA Report J-WECD-ECB-SOP-1016-02. Environmental Protection Agency, Cincinnati, OH. 13 pp.

Tennessee Department of Environment and Conservation (TDEC). 2020. Guidance for Making Hydrologic Determinations, version 1.5. Division of Water Pollution Control. (Available from: https://www.tn.gov/content/dam/tn/environment/water/policy-and-guidance/dwr-nr-g-03-hydrologic-determinations%E2%80%9304012020.pdf)

United States Environmental Protection Agency (USEPA). 2019. Flow duration protocol version 2.1. 38 pp.

Vengosh, A., R. B. Jackson, N. Warner, T. H. Darrah, and A. Kondash. 2014. A critical review of the risks to water resources from shale gas development and hydraulic fracturing in the United States. Environmental Science & Technology 48: 8334–8348.Wohl, E., M. K. Mersel, A. O. Allen, K. M. Fritz, S. L. Kichefski, R. W. Lichvar, T.-L. Nadeau, B. J. Topping, P. H. Trier, and F. B. Vanderbilt. 2016. Synthesizing the Scientific Foundation for Ordinary High Water Mark Delineation in Fluvial Systems. Wetlands Regulatory Assistance Program ERDC/CCREL SR-16-5, U.S. Army Corps of Engineers Engineer Research and Development Center. 217 pp. (Available from: https://apps.dtic.mil/sti/pdfs/AD1025116.pdf)

Wolock, D. M. 2003. Base-flow index grid for the conterminous United States: U.S. Geological Survey Open-File Report 03-263, digital dataset. (Available from: https://water.usgs.gov/lookup/getspatial?bfi48grd)

8 Appendix A: Glossary of Terms Used

Streamflow Class	Description
Ephemeral reaches	Flow only in direct response to precipitation. Water typically flows only during and/or shortly after large precipitation events, the streambed is always above the water table, and stormwater runoff is the primary water source.
Intermittent reaches	Contain sustained flowing water for only part of the year, typically during the wet season, where the streambed may be below the water table or where the snowmelt from surrounding uplands provides sustained flow. The flow may vary greatly with stormwater runoff.
Perennial reaches	Contain flowing water continuously during a year of normal rainfall, often with the streambed located below the water table for most of the year. Groundwater typically supplies the baseflow for perennial reaches, but the baseflow may also be supplemented by stormwater runoff or snowmelt.
At Least Intermittent (ALI)	Contain more than ephemeral flow but cannot be determined with high confidence if it is intermittent or perennial

Performance Measure	Description
PvIvE	Overall measure of accuracy. Ability of model to correctly classify between Perennial versus Intermittent versus Ephemeral. Calculated as the percent of reach-visits classified correctly (weighted by the number of visits per reach).
EvALI	Ability of model to correctly classify between Ephemeral and At Least Intermittent (I or P). Calculated as the percent of reach-visits classified correctly (weighted by the number of visits per reach).
IvEdry	Ability of the model to correctly classify Ephemeral vs Intermittent when sites are dry. Calculated as the percent of dry reach-visits classified correctly.
Precision	For reaches that have multiple visits, are they consistently predicted correctly? Calculated as the proportion of visits within a reach with the most frequent classification, averaged across reaches.

Dataset	Description
Training	A subset of 80% of the total reaches that was used for model development. This subset was randomly selected, stratifying by region (i.e., NE vs SE) and actual streamflow duration class (i.e., perennial, intermittent, and ephemeral).
Testing	A subset of 20% of the total reaches that was used for model testing and is independent from the training reaches. This subset was randomly selected, stratifying by region (i.e., NE vs SE) actual streamflow duration class (i.e., perennial, intermittent, and ephemeral).

Note: Data are divided by reach so that all visits at a single reach are included either in training or testing

Candidate Metric	Description	Type
ActiveFloodplain_score (NC)	Scoring based on visual estimate of floodplain characteristics adjacent to stream channel. Higher scores indicate greater evidence and continuity of adjacent floodplain.	Geom
alglive_cover_score	Visual estimate of live algal cover on the streambed within the study reach	Bio (algae)
alglivedead_cover_score	Visual estimate of the percent of streambed covered by live or dead algal growth	Bio (algae)

Candidate Metric	Description	Type
AlluvialDep_score (NC)	Scoring based on visual estimate of recently deposited alluvium in the channel and on the floodplain. Higher scores indicate greater amounts of fresh alluvium observed.	Geom
Amphib_abundance	Abundance of aquatic life stages of amphibians in the channel	Bio (verts)
Amphib_richness	Richness of amphibians with aquatic life stages in the channel	Bio (verts)
BankWidthMean	Mean bankfull width (m)	Geom
BMI_score (NC)	Original ordinal scoring of benthic macroinvertebrates based on abundance, richness and relative abundance of tolerant taxa (original list). See table below for detailed description.	Bio (aquatic inverts)
BMI_score_alt1	Simplified alternative 1 for ordinal scoring of benthic macroinvertebrates based on abundance, richness, and relative abundance of tolerant taxa (simplified list of tolerant taxa). See table below for detailed description.	Bio (aquatic inverts)
BMI_score_alt2	Simplified alternative 2 for ordinal scoring of benthic macroinvertebrates based on total abundance and richness of non-tolerant taxa (simplified list of tolerant taxa). See table below for detailed description.	Bio (aquatic inverts)
BMI_score_alt3	Simplified alternative 3 for ordinal scoring of benthic macroinvertebrates based on total abundance and richness of non-tolerant taxa (original list). See table below for detailed description.	Bio (aquatic inverts)
BMI_score_alt4	Simplified alternative 4 for ordinal scoring of benthic macroinvertebrates based on total abundance and richness. See table below for detailed description.	Bio (aquatic inverts)
ChannelDimensions_score (NM)	Scoring based on measured or visual estimate of the extent of channel entrenchment and connectivity to the floodplain. Higher scores are less confined (less incised) channels.	Geom
Continuity_score (NC)	Scoring based on visual estimate of the continuity of bank and streambed development. Higher scores indicate greater degree of channel development and continuity of bed and banks.	Geom
Crayfish_abundance	Abundance of crayfish and palaeomonid shrimp (Decapoda) within the channel	Bio (aquatic inverts)
Depositional_score (NC)	Scoring based on visual estimate of the extent of alluvial bars and/or benches present in the channel. Higher scores indicate greater prevalence of the features.	Geom
DifferencesInVegetation_score (NM)	Differences in vegetation between the riparian corridor and adjacent uplands score. Higher scores indicate a more distinct riparian corridor.	Bio (veg)
DRNAREA_mi2	Drainage area (mi^2) measured using USGS StreamStats or National Mapper	GIS
Elev_m	Watershed elevation (m) retrieved from StreamCat database	GIS
EPT_abundance	Abundance of mayflies, stoneflies, or caddisflies (i.e., Ephemeroptera, Plecoptera, Trichoptera, EPT)	Bio (aquatic inverts)
EPT_taxa	Number of EPT families	Bio (aquatic inverts)
FibrousRootedPlants_score (NC)	Scores based on visual estimate of the extent and distribution of non-woody, small diameter roots of water-intolerant plants in the streambed. Higher scores indicate lower density and coverage of roots in streambed.	Bio (veg)

Candidate Metric	Description	Type
fishabund_score2	Scoring of fish abundance (except non-native mosquitofish) in the channel	Bio (verts)
fp_entrenchmentratio_mean	Mean entrenchment ratio (capped at 2.5)	Geom
GOLD_abundance	Abundance of Gastropoda, Oligochaeta, and Diptera (GOLD) taxa	Bio (aquatic inverts)
GradeControl_score	Scores based on visual estimate of the extent and kinds of grade control features in the channel. Higher scores indicate larger size, more prevalence, and/or permanence of structures acting as grade controls in the channel.	Geom
Headcut_score	Scores based on visual estimate of the size and number of headcuts in the channel. Higher scores indicate presence and greater vertical drop in bed associate with headcut(s) in the channel	Geom
HydricSoils_score (NC)	Presence/absence of hydric soils within the study reach	Hydro
Hydrophytes_inchannel	Number of hydrophytic plant species (FACW or OBL) observed within the study reach channel	Bio (veg)
hydrophytes_present	Number of hydrophytic plant species (FACW or OBL) observed within the study reach channel and 1/2 channel width of the stream on either bank	Bio (veg)
hydrophytes_present_noflag	Number of hydrophytic plant species (FACW or OBL) observed within the study reach channel and 1/2 channel width of the stream on either bank (excluding taxa with unusual distributions flagged by the field crew)	Bio (veg)
iofb_score (NC)	Scores based on the visual estimate of the abundance of iron-oxidizing bacteria and fungi. Higher scores indicate greater abundance of iron oxidizing bacteria and fungi.	Bio (other)
LeafLitter_score (NC)	Scores based on the visual extent of the streambed covered by leaf litter. Higher scores indicate greater proportion of the streambed covered by leaves.	Hydro
liverwort_cover_score	Liverwort cover on the streambed. Higher scores indicate higher liverwort cover on streambed.	Bio (veg)
MeanSnowPersistence_01	Mean snow persistence (% of time between 1 Jan to 3 July) within a 1-km radius of the reach	GIS
MeanSnowPersistence_05	Mean snow persistence (% of time between 1 Jan to 3 July) within a 5-km radius of the reach	GIS
MeanSnowPersistence_10	Mean snow persistence (% of time between 1 Jan to 3 July) within a 10-km radius of the reach	GIS
Mollusk_abundance	Abundance of aquatic mollusks in the channel	Bio (aquatic inverts)
Mollusk_taxa	Richness of aquatic mollusk families in the channel	Bio (aquatic inverts)
moss_cover_score	Moss cover on the streambed. Higher scores indicate higher moss cover on streambed.	Bio (other)
NaturalValley_score (NC)	Scores based on the visual estimate of the extent of valley definition (proportion of catchment area sloping to the valley bottom). Higher scores indicate greater proportion of the catchment area slopes to the valley bottom or channel.	Geom
Noninsect_abundance	Abundance of non-insect aquatic invertebrate taxa in the channel	Bio (aquatic inverts)

Candidate Metric	Description	Type
Noninsect_taxa	Richness of non-insect aquatic invertebrate taxa in the channel	Bio (aquatic inverts)
OBL_inchannel	Number of OBL hydrophytic plant species observed within the study reach channel	Bio (veg)
OBL_present	Number of OBL hydrophytic plant species observed within the study reach channel and 1/2 channel width of the stream on either bank	Bio (veg)
OBL_present_noflag	Number of OBL hydrophytic plant species observed within the study reach channel and 1/2 channel width of the stream on either bank (excluding taxa with unusual distributions flagged by the field crew)	Bio (veg)
OCH_abundance	Abundance of aquatic Odonata, Coleoptera, and Heteroptera (OCH) in the channel	Bio (aquatic inverts)
ODL_score	Scores based on the visual estimate of the size and distribution of organic debris accumulations in and along channels. Higher scores indicate larger and more extensive accumulations.	Hydro
PctShading	Percent shading on the streambed	Bio (veg)
perennial_NC_abundance (NC)	Abundance of NC Method perennial invertebrate indicator taxa	Bio (aquatic inverts)
perennial_NC_live_abundance (NC)	Abundance of NC Method perennial invertebrate indicator taxa (living specimens only)	Bio (aquatic inverts)
perennial_NC_taxa (NC)	Number of NC Method perennial invertebrate indicator taxa	Bio (aquatic inverts)
perennial_PNW_abundance (PNW)	Abundance of PNW SDAM perennial indicator invertebrate taxa	Bio (aquatic inverts)
perennial_PNW_live_abundance (PNW)	Abundance of PNW SDAM perennial indicator invertebrate taxa (living specimens only)	Bio (aquatic inverts)
perennial_PNW_taxa	Number of PNW SDAM perennial indicator taxa	Bio (aquatic inverts)
ppt	30-y normal mean annual	GIS
ppt.11121	Average of 30-year normal mean monthly precipitation for November, December, and January	GIS
ppt.234	Average of 30-year normal mean monthly precipitation for February, March, and April	GIS
ppt.567	Average of 30-year normal mean monthly precipitation for May, June, and July	GIS
ppt.8910	Average of 30-year normal mean monthly precipitation for August, September, and October	GIS
ppt.m01	30-year normal mean January precipitation	GIS
ppt.m02	30-year normal mean February precipitation	GIS
ppt.m03	30-year normal mean March precipitation	GIS
ppt.m04	30-year normal mean April precipitation	GIS
ppt.m05	30-year normal mean May precipitation	GIS
ppt.m06	30-year normal mean June precipitation	GIS
ppt.m07	30-year normal mean July precipitation	GIS
ppt.m08	30-year normal mean August precipitation	GIS
ppt.m09	30-year normal mean September precipitation	GIS

Candidate Metric	Description	Type
ppt.m10	30-year normal mean October precipitation	GIS
ppt.m11	30-year normal mean November precipitation	GIS
ppt.m12	30-year normal mean December precipitation	GIS
REGION	Northeast or Southeast	GIS
Richness	Total richness of aquatic invertebrate families	Bio (aquatic inverts)
RifflePoolSeq_score (NC)	Visual estimate of the diversity and distinctiveness of riffles, pools, and other flow-based microhabitats. Higher scores indicate more distinctive riffles, pools, and other flow habitats with clear transitions within the reach.	Geom
SedimentOnPlantsDebris_score (NC)	Visual estimate of the extent of evidence of sediment deposition on plants and on debris within the floodplain. Higher scores indicate that sediment deposition was more prevalent throughout the reach.	Hydro
Sinuosity_score (NC)	Scored channel sinuosity. Higher scores indicate more sinuous channels.	Geom
Slope	Reach slope as measured with a handheld clinometer	Geom
StreamOrder	Strahler stream order from USGS StreamStats synthetic network	GIS
SubstrateSorting_score (NC)	Visual estimate of the extent of evidence of substrate sorting within the channel. Higher scores indicate greater sorting of substrate within the channel relative to surrounding uplands.	Geom
temp11121	Average of 30-year normal mean monthly air temperature for November, December, and January	GIS
temp234	Average of 30-year normal mean monthly air temperature for February, March, and April	GIS
temp567	Average of 30-year normal mean monthly air temperature for May, June, and July	GIS
temp8910	Average of 30-year normal mean monthly air temperature for August, September, and October	GIS
temp.m01	30-year normal mean January air temperature	GIS
temp.m02	30-year normal mean February air temperature	GIS
temp.m03	30-year normal mean March air temperature	GIS
temp.m04	30-year normal mean April air temperature	GIS
temp.m05	30-year normal mean May air temperature	GIS
temp.m06	30-year normal mean June air temperature	GIS
temp.m07	30-year normal mean July air temperature	GIS
temp.m08	30-year normal mean August air temperature	GIS
temp.m09	30-year normal mean September air temperature	GIS
temp.m10	30-year normal mean October air temperature	GIS
temp.m11	30-year normal mean November air temperature	GIS
temp.m12	30-year normal mean December air temperature	GIS
tmax	Maximum annual temperature (PRISM 30-year normal)	GIS
tmean	Mean annual temperature (PRISM 30-year normal)	GIS
tmin	Minimum annual temperature (PRISM 30-year normal)	GIS

Candidate Metric	Description	Type
TolRelAbund	Relative abundance of tolerant aquatic invertebrate taxa (original list)	Bio
TotalAbundance	Total abundance of aquatic invertebrates	Bio (aquatic inverts)
UplandRootedPlants_score (NC)	Scoring based on visual estimate of the extent of upland rooted plants (FAC, FACU, UPL, NI) growing within the streambed. Higher scores indicate fewer upland plants in the streambed.	Bio (veg)
WoodyJams_number	Number of woody jams present within the study reach channel (or up to 10 m outside of the study reach). Woody jams much completely span the active channel and be in contact with the streambed. Contain at least 3 large pieces (>1 m long and >10 cm diameter). Cause sufficient blockage to disrupt flow of water or sediment under flowing conditions.	Hydro

Detailed descriptions of ordinal benthic macroinvertebrate metric scores based on the metric used in the NC Method (NCDWQ 2010) evaluated for refining models for the Northeast and Southeast regions.

Metric abbrev	**Ordinal scores for benthic macroinvertebrate metrics**			
	0	1	2	3
BMI_score[1]	Total abundance = 0	Total abundance >0	Total abundance ≥4 and total relative abundance of tolerant taxa <90%	Total abundance ≥10 and richness ≥3 and total relative abundance of tolerant taxa <90% OR Richness ≥5 and total relative abundance of tolerant taxa <90%
BMI_score_alt1[2]	Total abundance = 0	Total abundance >0	Total abundance ≥4 and total relative abundance of simplified list of tolerant taxa <90	Total abundance ≥10 and richness ≥3 and total relative abundance of simplified list of tolerant taxa <90% OR Richness ≥5 and total relative abundance of simplified list of tolerant taxa <90%
BMI_score_alt2[2]	Total abundance = 0	Total abundance 1 to 2 individuals of non-tolerant taxa based on the simplified list of tolerant taxa	Total abundance ≥4 and at least 2 non-tolerant taxa present based on the simplified list of tolerant taxa	Total abundance ≥10 and at least 3 non-tolerant taxa present based on the simplified list of tolerant taxa
BMI_score_alt3[1]	Total abundance = 0	Total abundance 1 to 2 individuals of non-tolerant taxa OR only tolerant individuals present	Total abundance ≥4 and at least 2 non-tolerant taxa present	Total abundance ≥10 and at least 3 non-tolerant taxa present
BMI_score_alt4	Total abundance = 0	Total abundance 1 to 3	Total abundance ≥4	Total abundance ≥10 and richness ≥3 OR Richness ≥5

[1] Uses original list of tolerant taxa includes: Annelida, Hydracarina, Turbellaria, Nematoda, Nematomorpha, Physidae, Amphipoda, Isopoda, Chironomidae, Culicidae, Psychodidae, Libellulidae, Coenargionidae, Belastomatidae, Corixidae, and Haliplidae

[2] Uses simplified list of tolerant taxa includes: all non-insects (except Bivalvia and Decapoda), Culicidae, and Chironomidae