Environmental Modelling and Software 132 (2020) 104776

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

Position Paper

A clustering program for delineating ecologically homogenous segments within a stream network using catchment area and valley landscape similarity rules☆

Bastian J. Schmidt *, Nicholas E. Jones

Ontario Ministry of Natural Resources and Forestry, Trent University, 2140 East Bank Drive, Peterborough, ON, K9J 1Z8, Canada

ARTICLE INFO

Keywords:
Clustering
Ecology
Stream network
Landscape character
Upstream catchment
Computer program

ABSTRACT

The characteristics of a landscape and drainage size defines the abiotic character of the stream. We advance on an existing clustering method by making it faster, more stable, capable of processing large areas, and adding several new user defined options. The software clusters interconfluence stream reaches into larger homogenous stream segments using a series of rules about network adjacency, upstream catchment area, and similarity of landscape characteristics. Landscape variables and clustering parameters are discussed to provide guidance for helping users achieve their clustering objectives. We provide an example of clustering stream reaches in Ontario, Canada that was used to develop a provincial aquatic classification.

1. Introduction

The physical character of a landscape defines the abiotic character of the stream it drains, or as Hynes (1975) aptly states "the valley rules the stream". In turn, a stream's abiotic character is the template that determines the composition of ecological communities (Southwood, 1977; Poff, 1997). This connection between streams and their landscapes allows us to make inferences about lotic communities using broad-scale digital landscape data (e.g. geology, land cover). This means that costly field data on every stream reach is not required for inventory. The growth of computing power and remote sensing technology has enabled network analysis to rapidly change thinking in stream ecology and opened new avenues of discovery (Jones and Schmidt, 2017a).

However, there are challenges to overcome before meaningful inferences can be made based on the habitat templates because they are not discretely compartmentalized in space and time (Cullum et al., 2016). Instead, streams flow through a complex collection of overlapping gradients of environmental conditions. Ecological classification attempts to impose boundaries at meaningful thresholds along these gradients to create ecologically homogenous habitat units that are useful for management (Omernik, 1987; Melles et al., 2014). The fundamental premise is that ecological variability within the classified units is less than the variability between units and that they support similar biota, which respond similarly to stressors. The resulting classification reduces complexity allowing resource managers to make informed decisions not possible in the past.

A major challenge is determining boundaries that discretize stream networks into homogenous parts. Previous efforts have used expert opinion to define stream units which is fine if the landscape is well known (Baker, 2006; Seelbach et al., 2006). But as noted by Brenden et al. (2008a) individual aquatic ecologists will likely interpret the same maps differently, which may result in different river valley segment boundaries for a river network. Earlier works used broad zones to define longitudinal characteristics (Illies and Botosaneanu, 1963) and combinations of watersheds and terrestrial land classifications (Omernik and Bailey, 1997).

Defining stream units also requires a hierarchical approach to account for the different spatial scales at which the physical processes operate (Frissell et al., 1986; Poole, 2002). A bottom-up classification approach can be used, where smaller spatial units are aggregated into increasingly larger units forming the different levels of the hierarchy (Melles et al., 2014). A suitable fundamental unit (i.e. the lowest level of

☆ Program Title: Reach Affinity Tool (RAFT); Developer: Bastian Schmidt; Hardware required: IBM-compatible 64-bit computer system running Windows 7 or higher; Other software required: Microsoft Excel version 14 or later; First available: April 15, 2020; Program Size: <1 MB, Program Language: MATLAB R2016b; RAFT is available free of charge for download at https://github.com/RiverStreamEcologyLab.

* Corresponding author.
E-mail address: bastian.schmidt@ontario.ca (B.J. Schmidt).

https://doi.org/10.1016/j.envsoft.2020.104776
Received 21 August 2019; Received in revised form 17 April 2020; Accepted 16 June 2020
Available online 4 July 2020

the hierarchy) in river networks is the interconfluence reach, referred to as a reach from herein. Reaches have a clear spatial definition and there are existing methods to delineate them using Geographic Information Systems (GIS) and digital elevation models (DEM; Fig. 1). The reaches and their catchments (i.e. areas of land contributing water to them) can be attributed with a variety of landscape data thought to influence ecohydrological processes. This reach-based landscape data is simply an inventory of fundamental spatial units. A process is needed that will bring these reach units together to form larger, ecologically meaningful units using a set of objective and repeatable aggregation rules. These larger units are called stream segments. The segments represent the next hierarchical level of the stream classification and they are thought to be relatively homogenous units for abiotic and biotic characteristics and persistent over space and time (Seelbach et al., 1997; Seelbach et al., 2006). The conceptual foundation for the reach to segment aggregation rules has been laid by existing theoretical knowledge pertaining to how lotic systems change from headwaters to river mouth starting with the geomorphological concepts of Horton (1945) and Strahler (1957), to the ecological thinking introduced with the River Continuum Concept (RCC; Vannote et al., 1980), followed by the Serial Discontinuity Concept (SDC; Ward and Stanford, 1983) and more recently the Network Composition Hypothesis (NCH; Jones and Schmidt, 2017a).

We present a software tool called Reach Affinity Tool (RAFT). The tool is capable of objectively applying a series of upstream catchment area (i.e. size) and similarity rules based on the concepts outlined by the NCH to classify streams across large geographic extents (>1,000,000 km^2) at high spatial resolution (≤1:100,000 scale or ≤30 m cell size) using readily available spatial digital data (e.g. DEMs, geology, land cover). The software has already been applied during the development of the aquatic ecosystem classification (AEC) for the province of Ontario, Canada (Jones and Schmidt, 2017b). Although we are presenting RAFT from an ecological perspective, it is not limited to ecological applications. Other potential uses for RAFT include clustering reaches into segments of homogenous valley morphology (e.g. incision depth, flood plain width), potential for bank erosion and sediment transport (e.g. stream power, bed material, and sinuosity) and nutrient loading potential (e.g. landcover, geology, and soils) to name a few examples. The spatial structure and attribute content of RAFT input data is entirely in the hands of the user. For example, if the user would like to explicitly include dams and barriers in their results then they must prepare their input stream network data to include break-points at these features to create unique reaches prior to running RAFT. In the example we provide for Ontario, anthropogenic factors were intentionally not included because (1) we were interested in assessing fundamental natural drivers (e.g., geological setting) of stream type and (2) including anthropogenic factors (e.g., land use) would require that the classification be continuously updated which defeats its purpose. Anthropogenic influences are addressed by a separate stream stress analysis we conducted, which can be overlaid on the stream classification (unpublished data).

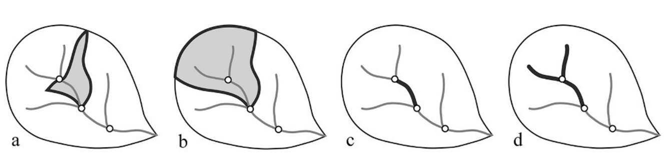

Fig. 1. The four scales of landscape summary that were derived using Arc Hydro layers. The heavy black lines illustrate each scale in the context of a hypothetical stream network: a) reach contributing area, b) upstream catchment area, c) reach channel (single raster cell width or buffered), and d) upstream channel for the catchment. The areas are the landscape that drains to the reach based on a digital elevation model and the D8 flow direction method of watershed delineation (Wilson and Gallant, 2000).

2. Program design and operation

RAFT is a spatially aware, network-based software tool that clusters stream reaches into stream segments at the next hierarchical level. The reaches that come together to form the segments should be of similar size and landscape character. The degree of similarity can be adjusted by the user via a variety of settings that can be manipulated using a graphical user interface (GUI). RAFT is based on the Valley Affinity Search Technique (VAST), which was developed by researchers at the University of Michigan (Seelbach et al., 2006; Brenden et al., 2008a). With their permission we used their MS-Excel Visual Basic for Applications (VBA) source code as a foundation. Our goal was to adapt and enhance the algorithms contained in the VAST code for classifying Ontario's rivers. Our software allows for the inclusion of lake influences within the network, it enhances the stream reach size and similarity aggregation rules, while greatly increasing processing stability and speed needed for use across large geographic areas.

2.1. Input data requirements

Input data for RAFT is composed of tabular data describing reaches within a river network where the characteristics of each reach line is described as a record (i.e. row) in a table. The lines of a network are called edges and the junction points between edges are called nodes. The edges of a network have an inherent direction describing the direction of movement along them (e.g. water flowing downstream). This directionality is crucial when working with stream networks. Two key attributes of the reach attribute table (i.e. fields or columns) are called the from-node and to-node identifiers. We used Arc Hydro (Maidment, 2002) to generate the stream network. Arc Hydro is spatial database structure that includes a set of network building and manipulation tools for use within ArcGIS (ESRI, 2012). Arc Hydro provides the benefit of generating a variety of spatial layers that can be used to collect and summarize information about the landscape draining to each reach (Fig. 1; also see Brenden et al., 2006).

Understanding streams in landscapes featuring many lakes requires the lakes to be taken into consideration when developing descriptive stream network algorithms (Brenden et al., 2006; Jones, 2010). RAFT includes the option to use reach data that includes waterbody inlet/outlet nodes in addition to the required basic confluence nodes required by Arc Hydro (Fig. 2).

Other means can be used to generate the required data table given that it includes the mandatory fields of Table 1 and the table is formatted as a comma separated text file. It is recommended that the choice of landscape variables and clustering thresholds is hypothesis driven, ecologically meaningful and match the scale of the process. For example, the amount of groundwater contributions to a stream (i.e. base flow) plays a large part in determining summer water temperatures, which in turn determines the type of fish species that can tolerate those

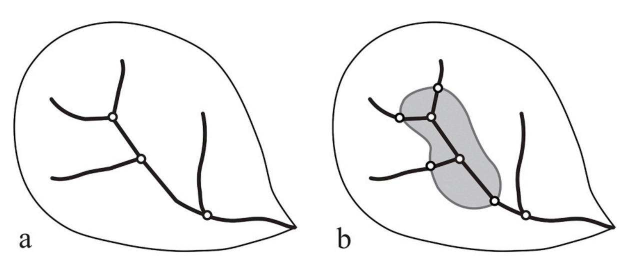

Fig. 2. An illustration of lake integration into the basic interconfluence reach network fabric. The points represent network nodes and the lines between the nodes are unique reaches (i.e. network edges). Illustration a) depicts the basic network nodes at confluences only, while b) shows the addition of network nodes at lake inlets and outlets, where new virtual network connector reaches are initiated.

Table 1
The mandatory input table structure RAFT requires to run. Field names marked with asterisk' require valid values for each reach (i.e no nulls). All other fields must be present but can be filled with inert placeholder values (e.g. 0, −999). The type and number of input landscape character variables used for creating the segments is entirely dependent on the user's classification objectives. It is recommended that the choice is hypothesis driven, representative of relevant physical process and matching the scale at which the physical process operates.

Field Name	Description	Type	Source
ObjectID	series of unique consecutive integers starting with 0	long integer	Arc Hydro
Arcid	series of unique consecutive integers starting with 1	long integer	Arc Hydro
from_node*	integer code of the node from which the reach originates	long integer	Arc Hydro
to_node*	integer code the node at which the reach terminates	long integer	Arc Hydro
Shape_Leng	length of the reach	double	Arc Hydro
HydroID	series of unique integers	long integer	Arc Hydro
GridID*	series of unique integers	long integer	Arc Hydro
NextDownID	integer code of the GridID of the next downstream reach	long integer	Arc Hydro
DrainID	series of unique integers	integer	Arc Hydro
Virtual	integer code indicating if the reach is real (0) or virtual (1)	integer	user added
UCA*	upstream catchment area at the reach terminus (units = m^2)	double	user added
Shreve*	Shreve stream order	integer	user added
Strahler*	Strahler stream order	integer	user added
SimpleLen	length of the reach based on a simplified reach geometry	double	user added
LI	integer code indicating upstream lake influence (1) or none (0)	integer	user added
SegmentID	integer code for the segment that the reach belongs to	long integer	user added
Variable 1*	Landscape characteristic variable	–	user added
Variable 2*	Landscape characteristic variable	–	user added
Variable n	Landscape characteristic variable (as many as needed)	–	user added

temperatures. The hydrologic process scale of this variable is the geology driven base flow contributions of the entire upstream catchment area, not just the reach contributing area (see Fig. 1). A second recommendation is keeping the list of variables as parsimonious as possible because as more variables are included in the similarity comparison it becomes less likely to find meaningful differences amongst reaches (Houle et al., 2010). Melles et al. (2014) provide a comprehensive summary of existing aquatic classifications and the landscape variables they have used. Some specific examples include catchment surficial geology, catchment slope, catchment landcover, reach July mean temperature, valley width and channel sinuosity (Snelder and Biggs, 2007; Brenden et al., 2008b). Not all variables included in the input table need to be used for the reach similarity comparisons. Setting a variable's weight to zero will cause RAFT to ignore it during the comparison but it still produces segment summary statistics for those variables, while a value of −999 will cause RAFT to ignore the variable entirely.

2.2. RAFT clustering algorithm

The probability of an ecological change downstream from a tributary is higher when the two merging streams are of similar size. Conversely, small tributaries flowing into to larger main stems have a lower probability of causing ecological changes (Jones and Schmidt, 2017a). The size relationship between the main stem and the tributary can be described by the confluence symmetry ratio (CSR), which is calculated by dividing the drainage area of the main stem by the drainage area of the tributary (Benda et al., 2004). In addition to size, the probability of change downstream from the confluence increases as the landscape character of the two drainages becomes more dissimilar. This means that even a small tributary with a low CSR can influence downstream characteristics on the larger main stem if it is draining a very different type of landscape. For example, a relatively small, warm and turbid tributary flowing into a large cold clear trout stream may change the main stem more than a similarly sized tributary that is also cold and clear. These relationships set up simple expectations about the relative importance of confluences in stream networks and provide a physical basis for the likelihood of change downstream of tributaries and network heterogeneity of habitats (Jones and Schmidt, 2017a).

RAFT groups reaches into segments based on upstream catchment area and similarity in landscape characteristics using two stages of processing. During both stages only adjacent reaches are compared (i.e. those connected directly downstream). The first stage divides the entire network into intermediate spatial units called neighbourhoods based on upstream catchment area (UCA) rules at confluence network nodes. The second stage joins reaches into segments within the neighbourhoods using landscape similarity comparisons. Neighbourhoods represent a classification by size while segments refine the size classification by considering landscape similarity.

2.3. Grouping reaches into neighbourhoods

Abrupt changes in the volume of flow, temperature, and sediment along the length of a stream are identified using the CSR. A CSR equal to one indicates that both reaches at the confluence have the same UCA. As tributary size decreases relative to the main stem CSR approaches zero. Virtual reaches within lakes (i.e. virtual network connectors within waterbody polygons) are automatically grouped into lake neighbourhoods, which subsequently become lake segments with one segment per lake.

Three rules are applied using the CSR values as illustrated in Fig. 3. The CSR rules are not applied at the confluences of 1st Strahler order headwater streams which are automatically assigned to the same

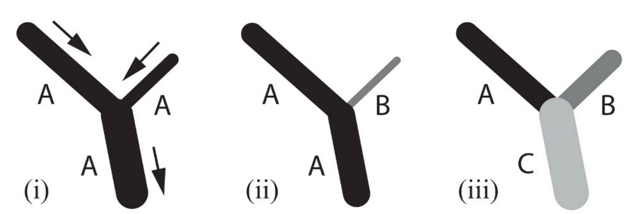

Fig. 3. Schematic representation of the rules applied while grouping stream reaches into neighbourhoods using the confluence symmetry ratio (CSR). The arrows indicate stream flow direction. In scenario i) the tributary is neither too small nor too large (e.g. CSR = 0.3) compared to the main stem, allowing all three reaches to be assigned to the same neighbourhood A. In scenario ii) the tributary is too small (e.g. CSR = 0.1) to cause a split in neighbourhood A and the tributary is assigned to a new neighbourhood B. The tributary in scenario iii) is large enough (e.g. CSR = 0.9) to cause a split in neighbourhood A, initiating a new neighbourhood C while the tributary itself is assigned to a new neighbourhood B.

neighbourhood to prevent unnecessary complexity in headwater areas at the size comparison stage, thus allowing the subsequent landscape similarity comparison to be the determining factor during clustering.

1. Stream reaches between the lower and upper CSR (e.g. 0.25–0.5) are allowed to join together into a neighbourhood because they have similar sizes, and thus, potentially similar ecological characteristics (Fig. 3i).
2. If the CSR is below a lower threshold (e.g. <0.25), a tributary is considered too small to significantly affect the main channel (Fig. 3ii). In this scenario, the main channel neighbourhood remains uninterrupted while the tributary becomes part of another neighbourhood. The rationale is that a small tributary should not become part of the main stem neighbourhood because they likely have different channel morphology (e.g., riparian shading, bankfull width).
3. Conversely, when the CSR at a confluence exceeds the upper threshold (e.g. >0.5) a new neighbourhood is initiated beginning with the reach directly downstream of the confluence (Fig. 3iii). The reasoning is that the combined volume of water in the downstream channel increases enough to change channel morphology (e.g., channel width, shading, temperature, riparian influence).

Subsequently, a fourth rule is applied that prevents large main stem river neighbourhoods from becoming too large. The mainstems of rivers of higher Strahler order have large drainage areas and few tributaries are large enough relative to these mainstem to split the neighbourhood based on the upper CSR threshold (Fig. 3iii). However, many small tributaries may join the mainstem causing its flow volume to gradually increase without causing any abrupt changes in stream character. The larger the mainstem catchment grows, the less likely it is that a tributary will be large enough to cause a split and as a result large portions of the river are likely to be grouped into a single neighbourhood (Fig. 4a).This is problematic because the stream reaches at the upstream end of such a neighbourhood might have a bankfull width of 25 m (i.e. UCA = 1000 km^2) whereas at the reaches at the downstream end might be 50 m wide (i.e. UCA = 2000 km^2) and therefore the reaches of this neighbourhood should not be considered ecologically homogeneous. Consequently, neighbourhoods with such an unacceptably wide range of reach UCAs need to be divided.

RAFT determines the best location to make the split within each of these neighbourhoods by calculating a ratio of a neighbourhood's minimum and maximum UCAs called the Neighbourhood Upstream Catchment Area Ratio (NUCAR). The NUCAR is a relative measure of the difference in UCA within a neighbourhood. Within each neighbourhood, RAFT identifies the reaches whose UCA fall between the lower and upper threshold of the user defined NUCAR window (e.g. 1.5 to 1.75).

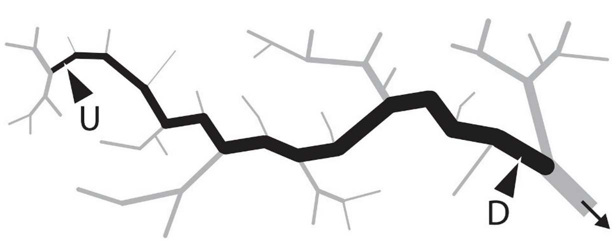

Fig. 4a. The black line illustrates a hypothetical neighbourhood defined by the CSR rules within a network. The neighbourhood is composed of multiple reaches starting upstream with *reach U* having a UCA of 150 km^2 and a furthest downstream *reach D* which has a UCA of 350 km^2 (NUCAR = 350 km^2/150 km^2 = 2.33). None of the small tributaries flowing into this neighbourhood were large enough to produce a split according to the CSR size rules yet the stream has more than doubled in UCA. This neighbourhood exceeds the upper NUCAR threshold of 2.0 it is passed to the NUCAR processing algorithm. The arrow indicates the direction of stream flow.

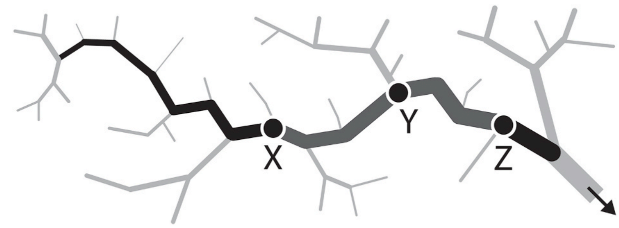

Fig. 4b. Using the upper and lower NUCAR thresholds RAFT creates a temporary window within the neighbourhood, which is searched for the largest tributary. In this example the upper NUCAR threshold of 2.0 is exceeded at a very small tributary at *confluence Z* and the lower NUCAR threshold of 1.5 is exceeded at *confluence X*, thus defining the extent of the search window (illustrated by the dark grey line between *confluence X* and *Z*). RAFT then finds the largest confluence within the search window at *confluence Y*.

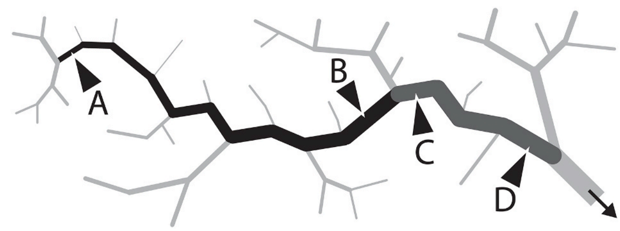

Fig. 4c. In this example RAFT splits the network neighbourhood into two new sub-neighbourhoods at the largest confluence it has found using the NUCAR window. The two new neighbourhoods are illustrated as the black line between *reach A* and *B*, and the dark grey line between *reach C* and *D*. The entire NUCAR process is repeated until all neighbourhoods within the processing extent fall below the upper NUCAR threshold (i.e. no neighbourhoods will have reaches that double in UCA).

The window allows RAFT some flexibility to find the largest tributary for making the split. This prevents neighbourhoods from being divided at small insignificant tributary confluences that can push the NUCAR above the upper threshold (Fig. 4b and 4c). The process repeats until all neighbourhoods have a NUCAR below the upper threshold value.

Stream width increases with UCA at a decreasing rate (Caissie, 2006). RAFT can account for this variable rate of width increase by allowing the user to apply an exclusion cut-off value below which the NUCAR process is not applied (e.g. UCA<50 km^2). For example the NUCAR for a mid-order neighbourhood (e.g. 350 km^2/150 km^2 = 2.33) is the same as the NUCAR for a low order neighbourhood (3.5 km^2/1.5 km^2 = 2.33). Ecologically speaking the difference in width may not be different enough in the low order neighbourhood (e.g. riparian vegetation shading unaffected) to warrant a split, hence a user may want to use the NUCAR exclusion rule.

An optional fifth rule can be applied during the neighbourhood creation process. It makes use of the zone of influence (ZOI) of lakes below their outlets. Lakes and rivers are intimately connected in an alternating series of lentic and lotic reaches in the Canadian Shield geologic region which is approximately 4.6 million km^2 or nearly half the land area of Canada and other similar geologic regions around the world (e.g. Scandinavia, Russia). Rapid and predictable changes in the ecological characteristics of streams occur at the interface with lakes (Jones, 2010). In many cases, water leaving the lake will be warmer and carry large amounts of seston (e.g., dissolved organic carbon, phytoplankton, zooplankton). Flows will be more constant (i.e. less flashy). The stream's water at the lake outlet might be clearer as particulate matter precipitates out in the lake. The combination of changes in the abiotic and biotic conditions fundamentally changes the ecology of the outlet stream, however changes are context dependent. For example, large rivers with catchment areas above 1000 km^2 maybe too large to be influenced by lakes. To make use of this optional rule the user must estimate a ZOI below every lake outlet based on some attenuation function where each reach in the network is assigned a lake influence (LI) value ranging from zero to one, where one is the maximum influence a lake a can exert, usually at the lake outlet. This maximum value attenuates to zero some distance downstream from the outlet indicating that a reach is not influenced by an upstream lake. It is the responsibility of the user to assign LI values to the reaches prior to running RAFT if they wish to include it during clustering.

2.4. Grouping reaches into segments within the neighbourhoods

RAFT's second stage groups reaches into segments within neighbourhoods based on their landscape similarities (e.g., geology, land cover, channel slope). Segments cannot expand outside their neighbourhood boundary but a neighbourhood can contain multiple segments. This process begins by sorting all reaches within a neighbourhood in descending Shreve order (Shreve, 1966) which means that processing will proceed from downstream to the upstream. RAFT starts a new segment using the first reach in the sorted list. The attributes of this first reach are compared with the attributes of its adjacent upstream reach along the main channel using a Euclidean distance (ED) measurement (Romesburg, 1989). When the Euclidean distance is below a user defined affinity threshold (AT) the adjacent reach is considered to be similar and part of the same segment. The algorithm continues the comparison process along the mainstem while smaller adjacent tributary reaches in the neighbourhood along the way are set aside for the moment. As the segment grows each new candidate reach is compared to each one of the reaches already part of the segment. Once the ED of the most upstream reach of the mainstem is evaluated the algorithm then moves on to the tributaries previously set aside, again working from largest to smallest Shreve order.

The ED to AT comparison can use one of several different methods: 1) the complete linkage (CLINK) method, which uses the largest ED between the candidate reach and the reaches of the segment, 2) the single linkage (SLINK) method, which uses the smallest ED between the candidate reach and the reaches of the segment, and 3) the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method, which uses the average of the EDs of the segment to the candidate reach. Details about differences amongst these methods can be found in Romesburg (1989). The ED calculations use equal attribute weighting by default, but the user can specify custom weights to emphasize importance of one or more variables.

Before any segment is considered finished it is treated to a *within* segment double-check, where the affinity of each reach to every other reach in the segment is checked once more. This ensures that all reaches in the segment remain under the AT because reaches added earlier in the process might exceed the AT when compared to reaches added later due to the processing order. If a reach is found with an affinity at or above the AT then it is assigned its own segment and the remainder of the reaches are returned for re-clustering. If no such reaches are found, then the segment is finished and a new segment is started for processing. The algorithm runs until all reaches in a neighbourhood have been assigned to a finished segment.

The last stage of the algorithm is one more double-check *amongst* segments. It may be the case where a reach has a higher affinity to an adjacent segment than to the segment to which it is currently assigned due to processing order. An affinity check is performed at the interface of segments but not across neighbourhood boundaries. Reaches with higher affinity (i.e. lower ED) to an adjacent segment are reassigned to that segment. This process is iterative until no reaches can be reassigned, indicating that optimal segmentation stability has been achieved.

2.5. Graphical user interface

The user interacts with RAFT using its graphical user interface (GUI). The GUI allows the user to load their input data, set all the necessary parameters required for RAFT and to specify output name and location. When RAFT is running the GUI also displays performance metrics and general progress status (Fig. 5).

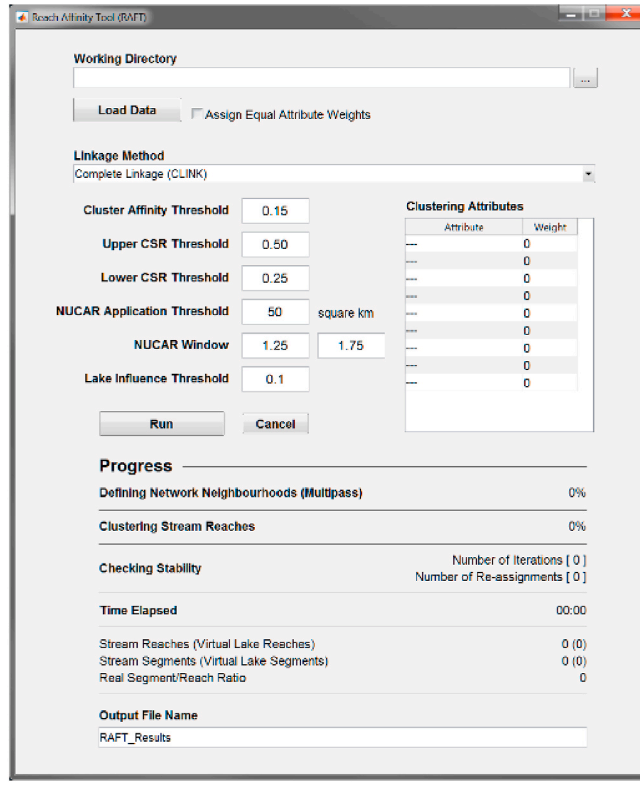

Fig. 5. A screen capture of RAFT's graphical user interface.

2.6. User defined threshold choices

The RAFT algorithm requires several user defined threshold values to produce optimal results. The threshold values will influence the size and structure of the resulting segments. There are no right or wrong choices, but the chosen values should take the user's tolerance for reach heterogeneity within segments into consideration. Higher similarity requirements will produce segments that are smaller (i.e. composed of fewer reaches) and more homogenous in character at the cost of increased segment count and classification complexity. Considering the rapid speed at which RAFT operates we encourage users to experiment with different threshold values to find the set that optimally suits their requirements.

The CSR thresholds are used to determine the size and structure of network neighbourhoods. The user must provide a lower and upper value or use default values. The lower threshold must be less than the upper threshold and both must be between zero and one. Decreasing the lower threshold value will result in more small tributaries being allowed to join a neighbourhood. Increasing the upper threshold will result in increasingly larger tributaries to join the neighbourhood. The larger the difference between the two CSR thresholds the larger neighbourhoods that are created (i.e. more reaches).

The LI threshold works in tandem with the CSR thresholds. Its value must be a between zero and one. LI has a strong effect on the way neighbourhoods are created because reaches that are considered to be influenced by lakes cannot be joined to reaches that are free of lake influence. The LI threshold value allows the user to specify a cut-off below which the neighbourhood creation part of the RAFT algorithm ignores non-zero LI values. This gives the user some flexibility to exclude inconsequentially small LI values. This is especially useful when using LI attenuating functions that produce asymptotic results (e.g. natural decay). RAFT ignores all input LI values during the neighbourhood creation process when the threshold is set to 1.

The NUCAR window is defined by a pair of threshold values that work together in tandem. They refine the size and structure of network neighbourhoods after the CSR thresholds have been applied. The upper NUCAR dictates the maximum reach UCA difference allowed within a neighbourhood. Increasing the upper NUCAR threshold will allow larger neighbourhoods to form at the cost of increased inter-neighbourhood reach UCA dissimilarity. It is recommended to keep the NUCAR window width below 0.5 to allow for flexibility while making results more interpretable. The NUCAR application threshold allows the user to prevent the NUCAR rules from being applied to reaches below the thresholds UCA value. Setting this value to zero will result in the NUCAR rules to be applied to all reaches in the processing extent, while a value equal to the maximum UCA of the processing extent will prevent the NUCAR rules from being applied entirely.

The choice of AT affects the size of the segments RAFT creates within neighbourhoods. Lower AT values create smaller, more homogenous segments (i.e. fewer reaches and higher within segment similar) while higher values create larger, more heterogeneous segments (i.e. more reaches and lower within segment similarity). The theoretical range of ED values is a function of the range of the landscape characteristics that are used to calculate them. For example, three landscape variables that range from zero to one, using equal weights that sum to one (i.e. 1/3), should result in a set of ED values also between zero and one, meaning that the AT must also fall inside this range.

When the ED is zero, the landscape characteristics of the reaches are identical, while ED of one indicates perfectly opposite reaches. In practice these extremes are never reached thus we advise users to become acquainted with distribution of actual ED value within the entire processing extent prior to using RAFT. This distribution can be calculated by finding the ED of every reach to all other reaches. If the maximum ED is 0.6 then it would not be wise to choose an AT of 0.5 because the threshold would rarely be exceeded during the reach comparisons, producing large segments with unacceptable amounts of landscape character heterogeneity. Conversely, choosing an AT that is too low (e.g. 0.1) would result in highly homogenous segments. However, such segments would be so small that they would not reduce complexity much beyond the input reaches. Both scenarios defeat the purpose of clustering. It is recommended that several potential AT values be tested to allow evaluation of the different results against the hypothesized expectations, field measurements and the objectives of the clustering exercise.

2.7. RAFT tabular output

RAFT outputs an MS-Excel workbook containing several worksheets. The segment reaches worksheet contains information for the individual reaches, including the SegmentID to which they have been assigned and the variable values used to calculate ED during clustering. The segment summaries sheet contains a variety of summary statistics of each segment's composite reach input variables. The summaries include the variables used for clustering as well as the inert variables (i.e. those assigned a weight of zero). The neighbourhood sheet contains the neighbourhood identifier associated with each reach, capturing the results of size rule clustering alone. The distances sheet provides the absolute (i.e. global) ED of the reach to all other reaches of the segment and relative (i.e. local) within segment distance values. The absolute values are the raw ED similarity measures of each reach compared to all the other reaches inside the segment using the chosen comparison method (e.g. UPGMA). The relative local values are a range scaled measure of each reach's similarity within its segment. A relative local distance of zero indicates that the reach has the smallest within segment ED while a value of one indicates the reach with largest within segment ED. The settings sheet contains relevant metadata for each analysis run.

3. A practical demonstration of RAFT

The Aquatic Ecosystem Classification (AEC) for Ontario, Canada, serves as a real world application of RAFT ability to cluster reaches into ecologically homogenous segments (Jones and Schmidt, 2017b).

RAFT was used to process all the streams and rivers of Ontario totalling ~710,000 reaches across ~1,000,000 km^2 area of land resulting in ~335,000 segments and ~180,000 virtual lake segments. The processing was achieved by dividing the provinces into 47 smaller work units containing up to 100,000 reaches each. RAFT runs very efficiently with the largest work units finishing in a matter of minutes on workstation class computer (in 2018). We do not see any reason why larger areas with more reaches (>100,000) could not be processed during a single RAFT run.

The Saugeen River watershed serves as an example at a suitable scale for illustrating the reach to segment clustering process that was applied across the whole province. The Saugeen River flows into Lake Huron at Southampton, Ontario (outlet at 44°30′3″N, 81°22′26″W), and drains an area of 3983 km^2. It has a diverse landscape of post-glacial geologic deposits (e.g. gravel, sand, clay) with underlying mesozoic bedrock (e.g. shale, limestone) and a temperate climate. The spatial data was generated using Arc Hydro tools with the Ontario Integrated Hydrology 30 m

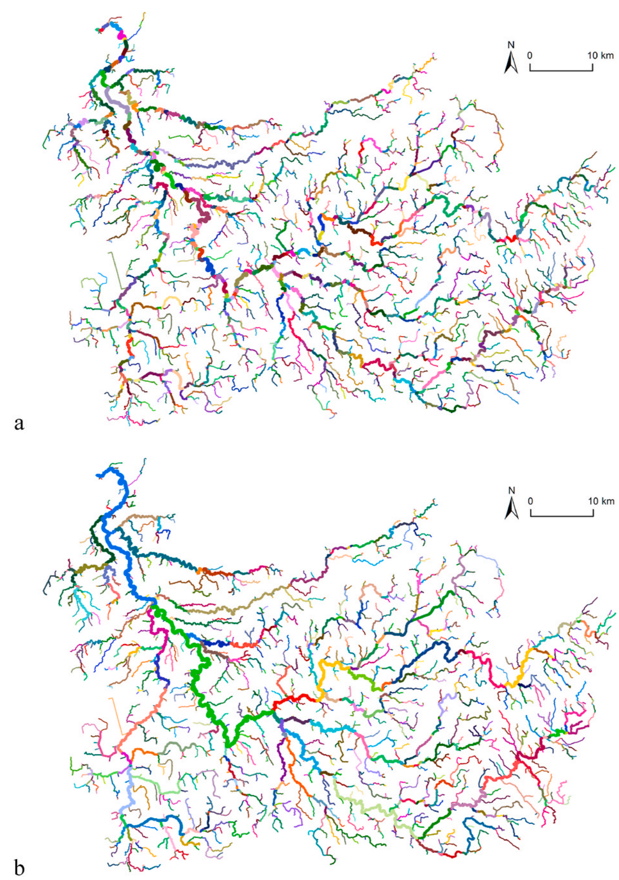

Fig. 6. a) The full complexity of the 2957 stream reaches of Saugeen River, Ontario, Canada (outlet at 44°30′3″N, 81°22′26″W), and b) the reduced complexity of the 2245 clustered stream segments. The colours of the lines represent individual reaches and segments. The higher order main stem segments are large (i.e. composed of many reaches) due to the homogenizing landscape averaging effect described by the Network Composition Hypothesis (NCH; Jones and Schmidt, 2017a). Conversely, the segments in the headwaters are generally smaller because of the increased potential for landscape heterogeneity amongst their component reaches (i.e. higher dissimilarity). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Settings, thresholds and landscape character variables used to create the demonstration data of the Saugeen River watershed.

Setting/Threshold	Setting
Linkage Method	CLINK
Affinity	0.15
Upper CSR	0.50
Lower CSR	0.25
NUCAR Application (km^2)	50
Lower NUCAR	1.25
Upper NUCAR	1.75
Lake Influence	0.10
Landscape Characteristic	**Weight**
Stream Temperature	0.333
% Upstream Channel Clay Geology	0.333
Reach Channel Slope	0.333

DEM (OMNRF, 2016) and the Ontario Hydro Network (OMNRF, 2015). The Saugeen River watershed contains 2957 reaches (Fig. 6a). RAFT clusters these reaches into 2245 segments (Fig. 6b) using the thresholds and landscape variables found in Table 2. We used three landscape variables for segment clustering. The first two are surrogates based on UCA geologic character. They are baseflow index (BFI), which is a measure of the proportion of groundwater contribution potential and low flow turbidity potential. The third variable is reach channel slope which is a surrogate for flow velocity. The input landscape variables used for the ED comparison were all range scaled to between 0 and 1 (Legendre and Legendre, 2012) and given equal weights. These settings correspond to the values used to produce the first version of the AEC for Ontario (Jones and Schmidt, 2017b). The higher order main stem segments are large (i.e. composed of many reaches) due to the homogenizing landscape averaging effect described by the Network Composition Hypothesis (NCH; Jones and Schmidt, 2017a). Conversely, the segments in the headwaters are generally smaller because of the increased potential for landscape heterogeneity amongst their component reaches (i.e. higher dissimilarity).

For the lake influence (LI) all reaches in the processing extent were sorted in ascending Shreve order to ensure that processing occurs from the outermost branches of the network to the pour point. The initial LI at the beginning of a lake outlet reach was set to one. The terminal LI of the reach was calculated by multiplying the reach length by an attenuation rate, which is the rate of change of LI per kilometre of reach length (e.g. Lake Influence Rate of Change, LIROC $= 1.66 \cdot \mathrm{UCA}_{\mathrm{reach}}^{-0.6}$), then subtracting the product of the multiplication from the initial LI. The initial

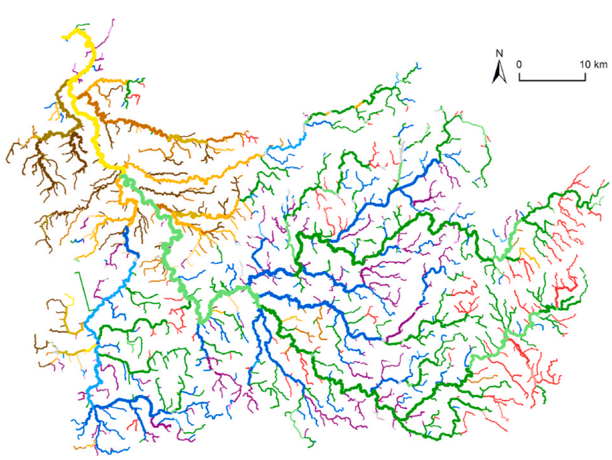

Fig. 7. The 16 classes of the first version of the Aquatic Ecosystem Classification for Ontario in the context of the Saugeen River, Ontario, Canada (outlet at 44°30′3″N, 81°22′26). The classes represent the highest hierarchical level of this classification system. The classes are composed of stream segments which in turn are composed of stream reaches.

LI at a confluence is the combination of the terminal LI of the confluent tributaries using an UCA weighted average. The final LI value assigned to a reach is the average of the initial and terminal LI. If the LI attenuates to zero along the reach, the proportion of reach length at zero is used to weight the averaging of the final LI value. First Shreve order streams without upstream headwater lakes are simply assigned a final LI of zero. The process is repeated until all reaches in the processing extent have been assigned an LI value.

The segments can be classified to form larger units, and thus, form another hierarchical classification level. For the AEC, the segments where grouped into classes based on their averaged reach attributes. This classification was achieved aspatially (i.e. no flow adjacency required) by binning the segments based on their groundwater water inflow potential (i.e. base flow), turbidity potential and DEM derived channel slopes. The binning process produced 16 stream classes, a drastic reduction in complexity (Fig. 6b vs Fig. 7).

4. Conclusions

Modern stressors like water taking, pollution, land use changes, flow alterations, climate change and over exploitation are threatening the ecological health of river systems around the globe. Resource managers tasked with mitigating or preventing further decline of degraded rivers, or preserving pristine systems, need new tools to help them gain a new perspective on their rivers at reach to regional scales. RAFT and resultant classifications allow these resource managers to rapidly gain such a perspective by reducing the inherent multi-scale stream network complexity using an objective stream reach aggregation method. The resulting segments are thought to represent the smallest useful units for resource management (Seelbach et al., 1997). Segments can be aggregated into additional spatial hierarchies (e.g. classes, zones) to further reduce complexity (Jones and Schmidt, 2017b). We recommended that users clearly define their management objectives or hypotheses because it will improve their choices of landscape variables, the process scales at which they operate and clustering thresholds. RAFT can be applied over large geographic extents using widely available global scale GIS landscape data (e.g. DEM, geology, land cover) making it suitable for national applications even in regions of the world that are deficient in locally collected GIS or field data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Brendan Quigley at the Ontario Provincial Geomatics Service Centre for assisting with the initial migration of the Michigan Valley Affinity Search Technique (VAST) code from VBA to MATLAB. We also appreciate the contributions of Paul Seelbach, Lizhu Wang, Kevin Wehrly, James McKenna and Catherine Riseng at the University of Michigan in Ann Arbor for sharing their experience with clustering the stream segments of Michigan. This research was funded by the Government of Ontario.

References

Baker, E.A., 2006. A Landscape-Based Ecological Classificationsystem for River Valley Segments in Michigan's UpperPeninsula. Michigan Department of Natural Resources,Fisheries Research Report 2085, Ann Arbor.
Benda, L., Andras, K., Miller, D., Bigelow, P., 2004. Confluence effects in rivers: interactions of basin scale, network geometry, and disturbance regimes. Water Resour. Res. 40 (5), 1–15.
Brenden, T.O., Clark Jr., R.D., Cooper, A.R., Seelbach, P.W., Wang, L., Aichele, S.S., Bissell, E.G., Stewart, J.S., 2006. A GIS framework for collecting, managing, and analyzing multiscale landscape variables across large regions for river conservation

and management. In: Hughes, R.M., Wang, L., Seelbach, P.W. (Eds.), Landscape Influences on Stream Habitats and Biological Assem-Blages, vol. 48. American Fisheries Society, Symposium, pp. 49–74 (Bethesda, Maryland).

Brenden, T.O., Wang, L., Seelbach, P.W., Clark, R.D., Wiley, J.M., Sparks-Jackson, B.L., 2008a. A spatially constrained clustering program for river valley segment delineation from GIS digital river networks. Environ. Model. Software 23 (5), 638–649.

Brenden, T.O., Wang, L., Seelbach, P., 2008b. a river valley segment classification of Michigan streams based on fish and physical attributes. Trans. Am. Fish. Soc. 137, 1621–1636.

Caissie, D., 2006. The thermal regime of rivers: a review. Freshw. Biol. 51 (8), 1389–1406.

Cullum, C., Rogers, K.H., Brierley, G., Witkowski, E.T.F., 2016. Ecological classification and mapping for landscape management and science: foundation for the description of patterns and processes. Prog. Phys. Geogr. 40 (1), 38–65.

ESRI, 2012. ArcGIS Desktop: Release 10.1. Environmental Systems Research Institute, Redlands, California.

Frissell, C.A., Liss, W.J., Warren, C.E., Hurley, M.D., 1986. A hierarchical framework for stream habitat classification - viewing streams in a watershed context. Environ. Manag. 10 (2), 199–214.

Horton, R.E., 1945. Erosional development of streams and their drainage basins - hydrophysical approach to quantitative morphology. Geol. Soc. Am. Bull. 56 (3), 275–370.

Houle, M.E., Kriegel, H.-P., Kroeger, P., Schubert, E., Zimek, A., 2010. Can shared-neighbour distances defeat the curse of dimensionality. In: Proceedings of the 22nd International Conference on Scientific and Statistical Database Management. https://doi.org/10.1007/978-3-642-13818-8. Heidelberg, Germany.

Hynes, H.B.N., 1975. The stream and its valley. Verhandlungen der Internationalen Vereinigung fuer Theoretische und Angewandte Limnologie 19, 1–15.

Illies, J., Botosaneanu, L., 1963. Problèmes et méthodes de la classification et de la zonation écologique des eaux courantes, considerées surtout du point de vue faunistique. SIL Communications 12 (1), 1–57. https://doi.org/10.1080/05384680.1963.11903811, 1953-1996.

Jones, N.E., 2010. Incorporating lakes within the river discontinuum: longitudinal changes in ecological characteristics in stream-lake networks. Can. J. Fish. Aquat. Sci. 67, 1350–1362.

Jones, N.E., Schmidt, B.J., 2017a. Tributary effects in rivers: interactions of spatial scale, network structure, and landscape characteristics. Can. J. Fish. Aquat. Sci. 74, 503–510.

Jones, N.E., Schmidt, B.J., 2017b. Aquatic Ecosystem Classification for Ontario's Rivers and Streams: Version 1, Ontario Ministry of Natural Resources and Forestry Science and Research Technical Note, p. 25. TN–04 2017.

Legendre, P., Legendre, L., 2012. Numerical Ecology, third ed. Elsevier Press, Amsterdam, NL, p. 990.

Maidment, D.R. (Ed.), 2002. Arc Hydro: GIS for Water Resources, ume 1. ESRI Inc, ISBN 1589480341, p. 203.

Melles, S.J., Jones, N.E., J Schmidt, B., 2014. Evaluation of current approaches to stream classification and a heuristic guide to developing classifications of integrated aquatic networks. Environ. Manag. 53, 549–566.

Omernik, J.M., 1987. Ecoregions of the conterminous united-states. Ann. Assoc. Am. Geogr. 77 (1), 118–125.

Omernik, J.M., Bailey, R.G., 1997. Distinguishing between watersheds and ecoregions. J. Am. Water Resour. Assoc. 33 (5), 935–949.

OMNRF (Ontario Ministry of Natural Resources and Forestry), 2015. User Guide for Ontario Hydro Network (OHN) Version 1.2. Provincial Mapping Unit, p. 68.

OMNRF (Ontario Ministry of Natural Resources and Forestry), 2016. User Guide for Integrated Hydrology: Elevation and Mapped Water Features for Provincial Scale Applications. Provincial Mapping Unit, p. 44.

Poff, N.L., 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. J. North Am. Benthol. Soc. 16 (2), 391–409.

Poole, G.C., 2002. Fluvial landscape ecology: addressing uniqueness within the river discontinuum. Freshw. Biol. 47 (4), 641–660.

Romesburg, H.C., 1989. Cluster Analysis for Researchers. Lulu Press, Morrisville, NC, p. 334.

Seelbach, P.W., Wiley, M.J., Kotanchik, J.C., Baker, M.E., 1997. A Landscape Based Ecological Classification System for River Valley Segments in Lower Michigan (MI-VSEC Version 1.0). Michigan Department of Natural Resources, Ann Arbor.

Seelbach, P.W., Wiley, M.J., Baker, M.E., Wehrly, K.E., 2006. Initial classification of river valley segments across Michigan's lower peninsula. In: Hughes, R.M., Wang, L., Seelbach, P.W. (Eds.), Landscape Influences on Stream Habitats and Biological Assemblages, vol. 48, pp. 25–48.

Shreve, R., 1966. Statistical law of stream numbers. J. Geol. 74, 17–37.

Snelder, T., Biggs, B., 2007. Multi-scale river environment classification for water resources management. JAWRA Journal of the American Water Resources Association 38, 1225–1239.

Southwood, T.R.E., 1977. Habitat, the templet for ecological strategies? J. Anim. Ecol. 46, 337–365.

Strahler, A.N., 1957. Quantitative analysis of watershed geomorphology. Trans. Am. Geophys. Union 38 (6), 913–920.

Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R., Cushing, C.E., 1980. River continuum concept. Can. J. Fish. Aquat. Sci. 37 (1), 130–137.

Ward, J.V., Stanford, J.A., 1983. Serial discontinuity concept of lotic ecosystems. In: Fontaine, T.D., Bartell, S.M. (Eds.), Dynamics of Lotic Systems. Ann Arbor Science, Ann Arbor, pp. 29–42.

Wilson, J.P., Gallant, J.C. (Eds.), 2000. Terrain Analysis: Principles and Applications. John Wiley and Sons, Inc, ISBN 0-471-32188-5, p. 479.