SoftwareX 11 (2020) 100401

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

OpenCLC: An open-source software tool for similarity assessment of linear hydrographic features

Ting Li [a,b], Lawrence V. Stanislawski [c], Tyler Brockmeyer [c], Shaowen Wang [a,b,*], Ethan Shavers [c]

[a] Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[b] CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[c] U.S. Geological Survey, Center of Excellence for Geospatial Information Science, Rolla, MO, USA

ARTICLE INFO

Article history:
Received 26 December 2018
Received in revised form 16 November 2019
Accepted 10 January 2020

Keywords:
CyberGIS
Line similarity assessment
Hydrography
National Hydrography Dataset

ABSTRACT

The National Hydrography Dataset (NHD) is a foundational geospatial data source in the United States that enables extensive and diverse environmental research and supports decision-making in numerous contexts. However, the NHD requires regular validation and update given possible inconsistent initial collection and hydrographic changes. Furthermore, systems or tools that use NHD data must manage regular updates that occur within the high-resolution version of the NHD (NHD HR). This research contributes to filling this gap by establishing an open-source software tool named OpenCLC, which automatically identifies matching and mismatching line features between two sets of hydrographic flowlines. Aside from identifying differences among two version of NHD lines, results can be applied to improve the quality of NHD HR content. OpenCLC significantly outperforms the best available commercial off-the-shelf software in computational scalability, and it is made widely available as part of the CyberGIS Toolkit to benefit broad environmental and geospatial science communities.

Code metadata

Current Code version	*V0.8*
Permanent link to code / repository used of this code version	https://github.com/ElsevierSoftwareX/SOFTX_2018_242
Legal Code License	*GNU General Public License, version 3*
Code Versioning system used	*git*
Software Code Language used	*C, python*
Compilation requirements, Operating environments & dependencies	*Compilers: GNU/Intel; OS: Linux (RedHat, Debian, Ubuntu, CentOS, SUSE); Dependencies: GDAL*
If available Link to developer documentation / manual	https://github.com/cybergis/cybergis-toolkit/
Support email for questions	CyberGIS Helpdesk (help@cybergis.org)

1. Introduction and background

The National Hydrography Dataset (NHD) is a comprehensive vector dataset of surface-water features within the United States ([1] U.S. Geological Survey, 2000) that is used for geomorphometric, hydrologic, and watershed research ([2] Sheng et al. 2007; [3] Maceyka and Hansen, 2016; [4] Schneider et al. 2017; [5] Vanderhoof et al. 2017; [6] Wu and Lane, 2017; [7] Liu et al. 2018). Two NHD datasets are currently available. The medium-resolution NHD, compiled from 1:100,000-scale source data, is a legacy dataset that is no longer maintained by the U.S. Geological Survey. The high-resolution NHD (NHD HR) is compiled from 1:24,000-scale (24K) or finer-scale source data and is regularly updated. NHD HR is the most up-to-date and detailed hydrography dataset for the United States ([8] U.S. Geological Survey, 2018). Multiple efforts are in progress to improve the accuracy of the NHD HR. For instance, drainage lines or other hydrographic features are often derived from recent digital elevation models (DEMs), lidar, or other data ([9] Poppenga et al. 2013; [10] Stanislawski, Buttenfield, and Doumbouya, 2015; [5] Vanderhoof et al. 2017). This paper describes an open-source software tool named OpenCLC

* Correspondence to: Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Room 2046, Natural History Building, MC-150, 1301 W. Green St., Urbana, IL 61801, USA.
E-mail address: shaowen@illinois.edu (S. Wang).

https://doi.org/10.1016/j.softx.2020.100401

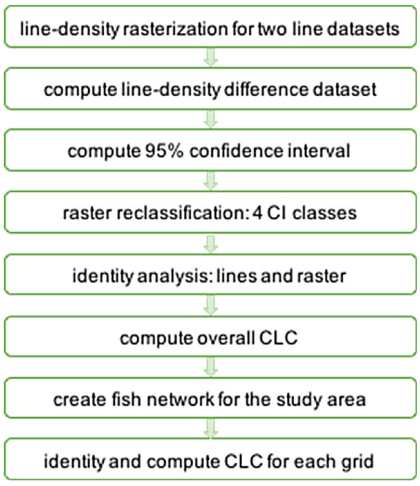

Fig. 1. The workflow of OpenCLC.

that automates a process of identifying matching and mismatching lines between two sets of lines representing similar features. The matching and mismatching lines determined through this process provide locations where verification can be focused to improve NHD content.

The process described here was first used by [10] Stanislawski, Buttenfield, and Doumbouya (2015) to generate the Coefficient of Line Correspondence (CLC), a metric used to estimate similarity for sets of elevation-derived drainage lines and NHD flowlines. Initial tests were conducted on thirty Hydrologic Unit Code 8 (HUC-8) subbasins, each of which includes between 700 to 21,000 surface-water flowlines. Early versions of the process were implemented with ArcGIS® tools customized through Python and required between 10 min to an hour of processing per subbasin. Because there are more than 2000 HUC-8 subbasins in the United States, widespread implementation was not practical. The new OpenCLC is implemented with Python and C programming languages to enable parallel processing within high-performance computing (HPC) environments. OpenCLC is part of the CyberGIS Toolkit that benefits broad environmental and geospatial science communities ([11] Wang, 2010; [12] Wang et al. 2016). CyberGIS Toolkit is a suite of programs and applications for GIS processing in HPC. The efficiency of OpenCLC is demonstrated by comparing elevation-derived drainage lines to NHD flowlines for all HUC-8 subbasins in the conterminous United States.

2. Workflow

Determining the CLC is an ordered process: rasterize the line density of each input dataset; calculate the difference between the two line density rasters; compute a confidence interval for matching cells (cells with difference values not significantly different from zero) based on the difference raster; reclassify the study area based on the difference raster and confidence intervals; identify the parts of line features within the matching or mismatching cells in the reclassified raster; compute the CLC of any given area as the sum of the length of all matching lines in both input datasets and dividing by the sum of the length of all lines in both datasets ([10] Stanislawski, Buttenfield, and Doumbouya, 2015) (Fig. 1). Each subbasin can be further partitioned using a spatial grid, and the CLC value is calculated in each smaller partition.

OpenCLC is designed to exploit the power of HPC for the CLC computation process and allow nationwide comparison of surface-water flow networks. It is implemented in Python and C and uses the Geospatial Data Abstraction Library (GDAL, http://gdal.org). As input, OpenCLC takes two datasets of line features in shapefile format (.shp), both with the same spatial reference and similar spatial extents (e.g., stream networks should be from the same subbasin or watershed). The user specifies two additional parameters, search radius and cell size, which are used in line-density rasterization. OpenCLC outputs two line-density rasters and their difference raster, a reclassified raster, a shapefile of matching and mismatching features between each input line dataset, and a shapefile of a spatial grid with CLC values.

3. Implementation

3.1. Line density rasterization

A straightforward way to calculate a line-density raster for an input line dataset is to create a circular buffer at each cell center, and then intersect this circle with the input lines to compute the line density for each cell. However, a buffer and intersection analysis at each cell can be computationally intensive. Each polyline in a shapefile is composed of a series of line segments. In OpenCLC, the calculation of line density at each cell is decomposed into calculating the intersected length of a single line segment and a circle (buffer boundary) centered on a pixel of interest, which is recomputed for all candidate line segments. Fig. 2 shows an efficient way to calculate the length of intersection between a single line segment and a circle using straightforward math calculation without buffer or intersection analysis. The lengths of all line segments from a dataset that intersect a cell's circle are summed to determine the final line density for that cell.

Instead of looping through all the line segments to calculate the length of intersection and density at each cell, the line-density rasterization process calculates line density for cells near each line segment within an input line dataset and accumulates partial results to get the final line density raster. Empty cells without nearby lines can be skipped. A query of cells near each line segment is easier to run than a query of lines near each cell. As shown in Fig. 3, the line-density rasterization procedure loops through each line segment, identifies cells within distance r of the bounding box for the line segment, and adds the segment-intersection length to the sum for each cell.

3.2. Identifying matching and mismatching features

Once the line density of each input line dataset has been created, a difference raster between the two line-density rasters is calculated. A confidence interval for matching cells (difference values not significantly different from zero) is calculated from the difference raster. Applying the vertical accuracy test procedures from the National Standard for Spatial Data Accuracy ([13] U.S. Federal Geographic Data Committee, 1998), an upper bound for a 95% confidence interval for matching line features is estimated by multiplying the root mean square error (RMSE) for the entire difference raster by 1.96. The difference raster is computed as the line-density raster for line dataset 1 (LD1) minus the line-density raster for line dataset 2 (LD2). Then the difference raster is reclassified into four classes based on the confidence interval: values less than the lower bound (clc_code = 1, LD1 match, LD2 do not match); values within the confidence interval (clc_code = 2, LD1 and LD2 match); values greater than the upper bound (clc_code = 3, LD1 do not match, LD2 match); and empty area, where line density of LD1 and LD2 is close to 0 (clc_code = 4, LD1 and LD2 do not match). Optionally, if waterbody polygons that overlap lines in both datasets are available, such as NHD double-line streams, lakes, ponds, and reservoirs, then they can be used

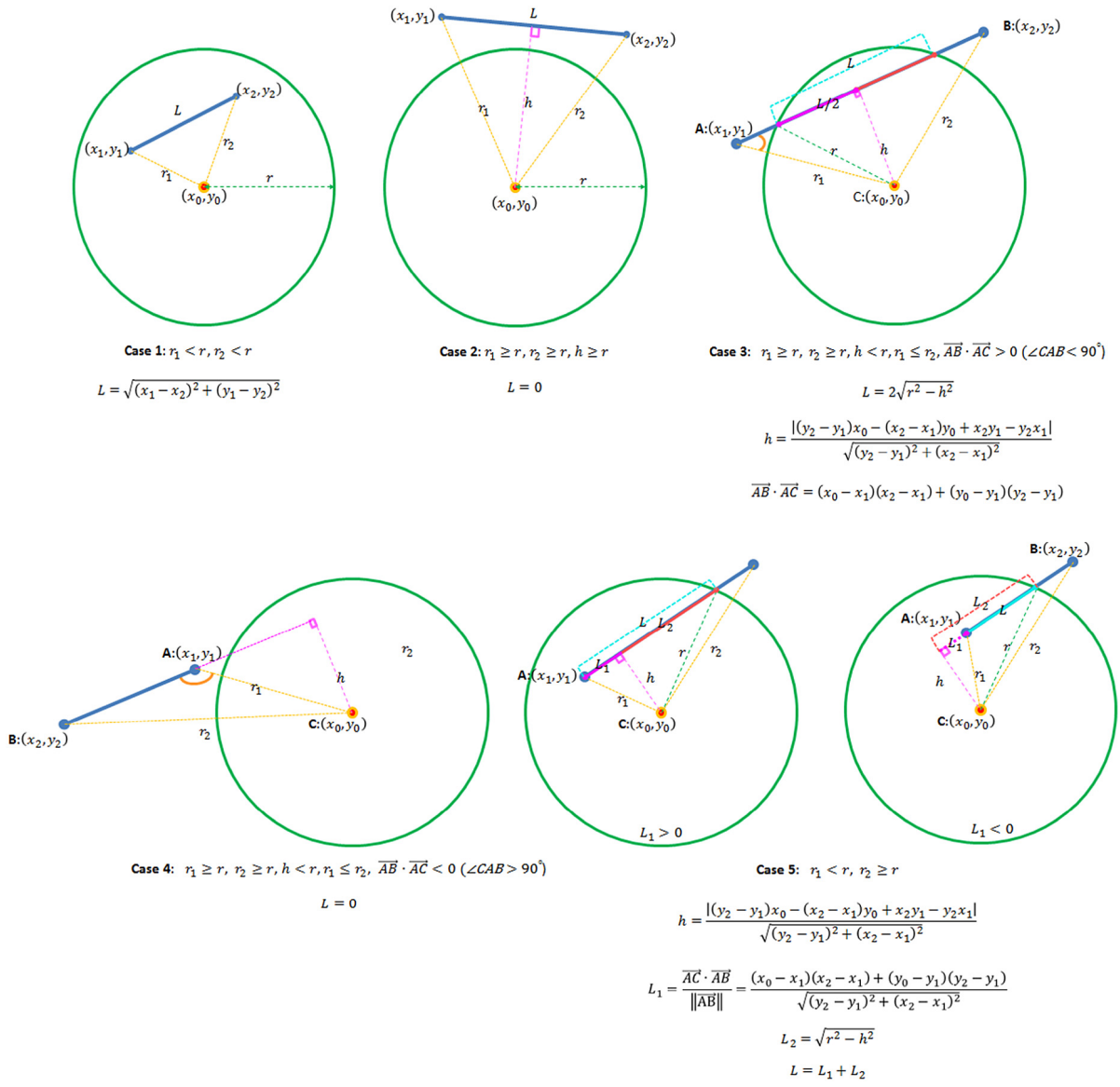

Fig. 2. Length of intersection (L) between line and circle.

by the program. The program will rasterize the waterbody polygons and categorize them as matching pixels (clc_code = 2) in the reclassified raster. This step allows better matching of flow lines within waterbodies, because they are subject to positional variation within the banks of these polygons. The reclassified difference raster is used to identify matching and mismatching features in the two input line datasets.

A line-tracking algorithm uses the clc_code values of intersecting raster cells of the reclassified difference raster to determine the portions of lines passing through matching or mismatching cells. As shown in Fig. 4, a tracking point moves from the start point to the end point of each line, and the line is split each time the value of the raster cell changes. Each polyline feature in the two input line datasets will be split into multiple parts depending on intersecting cell values in the reclassified difference raster. Each polyline part becomes a new polyline feature in the output shapefile with the additional clc_code attribute field. Given the line-density difference defined as LD1 density minus LD2 density, full or partial line features from LD1 with clc_code values of 3 or 4 are mismatching features, and features with clc_code values of 1 or 2 are matching features; full or partial line features from LD2 with clc_code values of 1 or 4 are mismatching features, and features with clc_code values of 2 or 3 are matching features.

3.3. CLC calculation

Upon identifying matching and mismatching lines in the two datasets, a CLC metric is computed as follows (Stanislawski et al. 2015):

$$CLC = \frac{M_1 + M_2}{M_1 + M_2 + O_1 + O_2}$$

M_1 is the sum of the length of matching features in LD1, M_2 is the sum of the length of matching features in LD2, O_1 is the sum of the length of features in LD1 that are omitted from the LD2, and O_2 is the sum of the length of features in the LD2 that are omitted from the LD1. CLC values estimate the proportion of lines that are matching and range between 0 and 1. A CLC value of 1 indicates all features match in both datasets, and 0 indicates no matching features.

The CLC metric is calculated for the entire study area where the two input line datasets are compared. It can also be calculated for all partitions in a grid (fishnet) that overlays the study area,

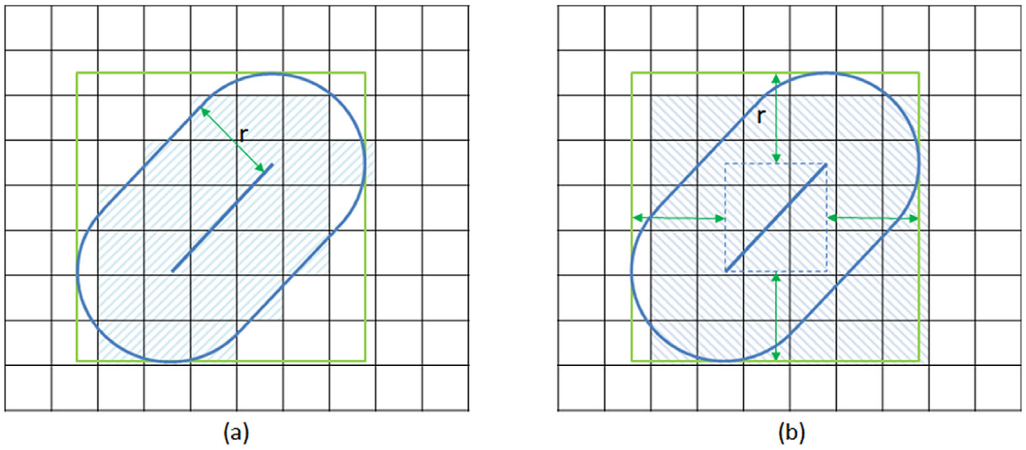

Fig. 3. (a) For each line segment, only nearby cell centers that fall within the buffer of radius r will be influenced by it when calculating line density; (b) the bounding box of line buffer with radius r is the same as the result of extending the bounding box of the line with r.

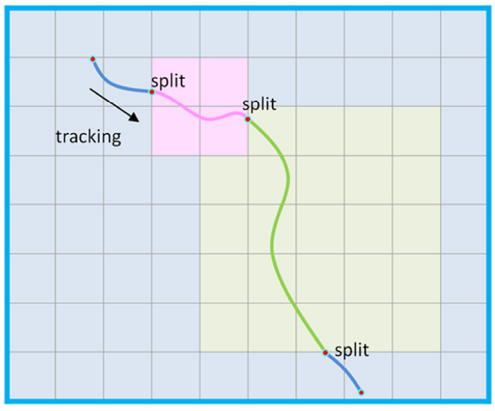

Fig. 4. Illustration of the line tracking algorithm. Different cell colors represent different class values in the reclassified raster. A tracking point moves along the line and splits the line when the cell value changes.

which generates a spatial distribution of CLC values. To generate the CLC distribution, an identity analysis finds the portion of all lines with the clc_code attribute that fall within each grid cell. Subsequently the CLC metric is determined for each subset of lines that fall within each grid cell. Relatively lower values in the CLC distribution indicate locations where further verification or improvement efforts may be focused.

4. Evaluation

This section demonstrates the use of the OpenCLC tool in an HPC environment to estimate the similarity between elevation-derived drainage lines and associated NHD flowlines in the conterminous United States.

4.1. Data and computing

Because NHD HR is compiled at multiple resolutions (or scales), the data generally must be thinned to a common scale in order to be used for analysis or display purposes. To accomplish this task, a reference drainage network that depicts natural drainage patterns at 24k was extracted for the conterminous United States from 1/3rd arc-second DEM data (nominal 10-m cell resolution) using a weighted flow accumulation model ([14] Stanislawski, Falgout and Buttenfield, 2015). Given a 24K-based drainage pattern, the NHD HR can be thinned to 24K and several smaller scales using a stratified pruning process ([15] Stanislawski, 2009; [16] Stauffer, Finelli and Stanislawski, 2016). The elevation-derived drainage lines were extracted from each of the 2119 HUC-8 subbasins in the conterminous United States, which furnished between 17 and 2800 km of drainage lines each. OpenCLC was tested on a 12-node Linux cluster, with each node having 20 processing cores and 128 GB of RAM. Job execution and computational resources—such as nodes, processors, memory and processing time—were managed through the Slurm Workload Manager. The workflow allowed simultaneous comparison of up to 240 HUC-8 subbasin datasets by processing one subbasin on each available core. Rapid access to file storage was provided through a parallel shared Lustre file system on a high-speed Infiniband network.

4.2. Results

OpenCLC processing of an individual HUC-8 subbasin with a 24-m cell resolution and radius for line-density computations required less than two minutes using a single processing core. This is a substantial improvement over the ArcGIS process that required between 10 min to an hour. Simultaneous processing of up to 240 subbasins completed all 2119 subbasins in the conterminous United States in about two hours. This task compares over 10 million km of flowlines, comprised from over 30 million vector features with more than 160 million vertices, and these values can be doubled to account for the elevation-derived drainage lines. The spatial distribution of HUC-8 CLC values determined through OpenCLC is shown in Fig. 5. Processing failed in 35 of the 2119 subbasins, which is a 1.65 percent failure rate. Overall, poorer matching is evident in the southwest, where many features are ephemeral features. This is an expected result, because ephemeral features were not included in model parameters for extracting elevation-derived drainage lines. These results identify specific subbasins where obvious problems exist with the extraction model, and where model improvements should be focused.

Fig. 6 shows the spatial distribution of matching and mismatching features identified by the OpenCLC, along with the gridded CLC values. The gridded results clearly identify sections with relatively high proportions of mismatching features that may require more detailed assessment (Fig. 6c), and the vector results furnish a more detailed assessment (Fig. 6b). A detailed analysis reveals that a majority of mismatching features occur among first-order (headwater) tributaries for subbasin 0303003 (Fig. 6a, b), and this is generally the case revealed for all tested subbasins.

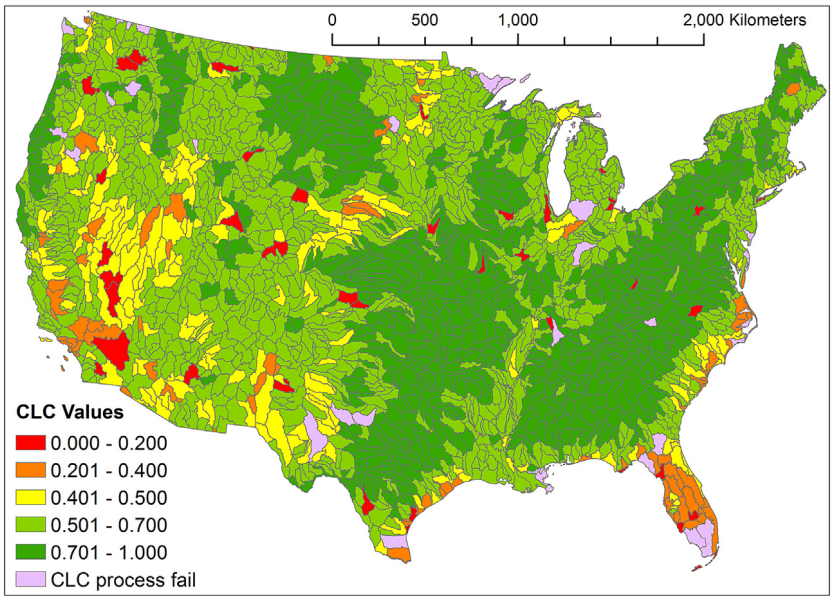

Fig. 5. Distribution of CLC values comparing 1:24,000-scale (24K) elevation-derived drainage lines to 24K NHD flowlines for the HUC-8 subbasins in the conterminous United States. CLC values estimate the proportion of lines that are matching; 1 indicates all features match in both datasets, and 0 indicates no matching features. CLC processing failed for 35 (pink shade) of 2119 subbasins.

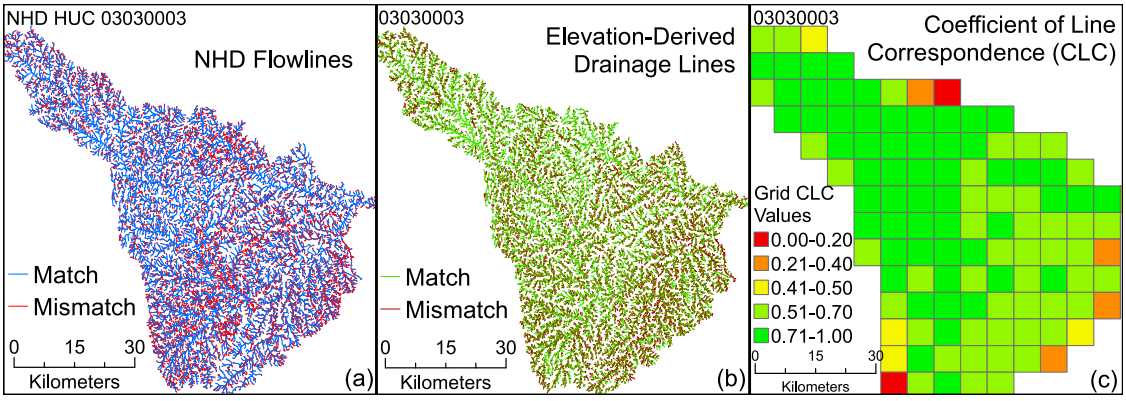

Fig. 6. Distribution of matching and mismatching (a) flowlines and (b) elevation-derived drainage lines determined by OpenCLC for the Deep River watershed in North Carolina, which is assigned 8-digit Hydrologic Unit Code (HUC8) 03030003 within the National Hydrography Dataset (NHD). Grid CLC values (c) show the distribution of CLC values for each about 6.6-km-by-6.6-km grid in the watershed.

5. Conclusion

As demonstrated by this analysis, OpenCLC is an effective tool for identifying differences between elevation-derived drainage lines and NHD flowlines that can greatly assist management and update of the NHD. Through an HPC-based implementation, OpenCLC provides a rapid assessment of the similarity between two different sets of hydrographic lines for very large regions or countries. The entire analysis for the conterminous United States required about two hours with OpenCLC, whereas other available tools would have taken several weeks for this process. This performance improvement unlocks other possible uses for the tool, such as refining or comparing drainage-line extraction models for the entire country.

OpenCLC is intended for automated comparison of two sets of linear features representing similar phenomenon of any type—such as comparing two sets of linear road features, or two sets of contour lines. This initial version of OpenCLC, as described in this paper, is configured to compare two sets of surface water drainage lines having similar positional accuracy. In this version, it is expected that the line-density resolution parameter is deduced from the positional accuracy of the input datasets. As in the above analysis, given the source scale of 24K for the input flowlines, it was deduced that a 24-m cell resolution was adequate for the line-density rasterization process and subsequent CLC computation. However, in order to compare two linear feature datasets with substantively different accuracies, the program would need to be modified to allow inputs of two resolution parameters and to process two line-density datasets with different resolutions. Such development is expected for future versions of the software.

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

This paper and associated materials are based in part upon work supported by the U.S. Geological Survey under grant number G14AC00244 and the National Science Foundation (NSF) under grant numbers 1443080 and 1664119. The work used the

ROGER supercomputer, which is supported by NSF under grant number 1429699. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. Assistance received from Anand Padmanabhan and Zewei Xu at the CyberGIS Center for Advanced Digital and Spatial Studies at the University of Illinois at Urbana-Champaign on data processing and software testing is greatly appreciated.

Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

[1] US Geological Survey. The National Hydrography Dataset: Concepts and Contents (2000). United States Geological Survey; 2000, https://nhd.usgs.gov/chapter1/chp1_data_users_guide.pdf, last accessed 1 March 2018.

[2] Sheng J, Wilson JP, Chen N, Devinny JS, Sayre JM. Evaluating the quality of the national hydrography dataset for watershed assessments in metropolitan regions. GISci Remote Sens 2007;44(3):283–304. http://dx.doi.org/10.2747/1548-1603.44.3.283.

[3] Maceyka A, Hansen WF. Enhancing hydrologic mapping using lidar and high resolution aerial photos on the Francis Marion National Forest in coastal South Carolina. In: Stringer Christina E, Krauss Ken W, Latimer James S, editors. Headwaters to Estuaries: Advances in Watershed Science and Management—Proceedings of the Fifth Interagency Conference on Research in the Watersheds. March (2015) 2-5, North Charleston, South Carolina. E-Gen. Tech. Rep. SRS-211, Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station; 2016, p. 302.

[4] Schneider A, Jost A, Coulon C, Silvestre M, Théry S, Ducharne A. Global-scale river network extraction based on high-resolution topography and constrained by lithology, climate, slope, and observed drainage density. Geophys Res Lett 2017;44:2773–81. http://dx.doi.org/10.1002/2016GL071844.

[5] Vanderhoof MK, Distler HE, Lang M. Integrating Radarsat-2, Lidar, and Worldview-3 imagery to maximize detection of forested inundation extent in Delmarva Peninsula, USA. Remote Sens 2017;9(105). http://dx.doi.org/10.3390/rs9020105, 2017.

[6] Wu Q, Lane CR. Delineating wetland catchments and modelling hydrologic connectivity using lidar and aerial imagery. Hydrol Earth Syst Sci 2017;21(2017):3579–95. http://dx.doi.org/10.5194/hess-21-3579-2017.

[7] Liu YY, Maidment DR, Tarboton DG, Zheng X, Wang S. A cyberGIS integration and computation framework for high-resolution continental-scale flood inundation mapping. J Am Water Resour Assoc 2018;54(4):770–84.

[8] US Geological Survey. Hydrography: NHDPlus High Resolution, National Hydrography Dataset, Watershed Boundary Dataset. United States Geological Survey; 2018, https://nhd.usgs.gov/index.html, last accessed 1 March 2018.

[9] Poppenga SK, Gesch DB, Worstell BB. Hydrography change detection: The usefulness of surface channels derived from LiDAR DEMS for updating mapped hydrography. J Am Water Resour Assoc 2013;49(2):371–89. http://dx.doi.org/10.1111/jawr.12027.

[10] Stanislawski LV, Buttenfield BP, Doumbouya A. A rapid approach for automated comparison of independently derived stream networks. Cartogr Geogr Inf Sci 2015;42(5):435–48. http://dx.doi.org/10.1080/15230406.2015.1060869.

[11] Wang S. A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. Ann Assoc Amer Geogr 2010;100(3):535–57.

[12] Wang S, Liu Y, Padmanabhan A. Open cyberGIS software for geospatial research and education in the big data era. SoftwareX 2016;5:1–5. http://dx.doi.org/10.1016/j.softx.2015.10.003.

[13] US Federal Geographic Data Committee. Geospatial Positioning Accuracy Standard, Part 3: National Standard for Spatial Data Accuracy. FGDC-STD-007.3-1998, Reston, VA: Federal Geographic Data Committee; 1998, p. 25.

[14] Stanislawski LV, Falgout J, Buttenfield BP. Automated extraction of natural drainage density patterns for the conterminous United States through high-performance computing. Cartogr J 2015;52(2):185–92. http://dx.doi.org/10.1080/00087041.2015.1119466.

[15] Stanislawski LV. Feature pruning by upstream drainage area to support automated generalization of the United States national hydrography dataset. Comput Environ Urban Syst 2009;33(5):325–33.

[16] Stauffer A, Finelli E, Stanislawski LV. Moving from Generalization to the VisibilityFilter Attribute: Leveraging Database Attribution to Support Efficient Generalization Decisions. Sacramento, California: American Water Resource Association; 2016.