# High Performance Computing to Support Multiscale Representation of Hydrography for the Conterminous United States

Lawrence V. Stanislawski[1], Yan Liu[2], Barbara P. Buttenfield[3], Kornelijus Survila[2], Jeffrey Wendel[1], and Abdurraouf Okok[1]

[1]U.S. Geological Survey, Center of Excellence for Geospatial Information Science, 1400 Independence Road, Rolla MO 65401
Email: lstan@usgs.gov, jwendel@usgs.gov, aokok@usgs.gov

[2]University of Illinois at Urbana-Champaign,
Email: yanliu@illinois.edu, survila2@illinois.edu

[3]University of Colorado-Boulder, Boulder, CO 80309-0260
Email: babs@colorado.edu

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

## 1. Introduction

The National Hydrography Dataset (NHD) for the United States furnishes a comprehensive set of vector features representing the surface-waters in the country (U.S. Geological Survey 2000). The high-resolution (HR) layer of the NHD is largely comprised of hydrographic features originally derived from 1:24,000-scale (24K) U.S. Topographic maps. However, in recent years (2009 to present) densified hydrographic feature content, from sources as large as 1:2,400, have been incorporated into some watersheds of the HR NHD within the conterminous United States to better support the needs of various local and state organizations. As such, the HR NHD is a multi-resolution dataset with obvious data density variations because of scale changes. In addition, data density variations exist within the HR NHD that are particularly evident in the surface-water flow network (NHD flowlines) because of natural variations of local geographic conditions; and also because of unintentional compilation inconsistencies due to variations in data collection standards and climate conditions over the many years of 24K hydrographic data collection (US Geological Survey 1955).

The Center of Excellence for Geospatial Information Science (CEGIS) of the U.S. Geological Survey (USGS), working in collaboration with the University of Colorado-Boulder, has developed custom tools to automate the generalization of the HR NHD from multi-resolution content to 24K and smaller scales. The generalization workflow entails enrichment, pruning, and simplification steps. The NHD data are subdivided into a hierarchy of basin and subbasin watersheds. Processing for the analysis in this paper uses the 8$^{th}$ level of the hierarchy, known as Hydrologic Unit Code 8 (HUC8) subbasins. By processing HUC8 subbasins, enrichment assigns upstream drainage area and values reflecting local stream density to each linear flow network feature in the subbasin (Stanislawski and Buttenfield 2011b). Subsequently,

density values for linear features are aggregated into density classes or partitions, and each density partition is separately pruned to a target density in a manner that maintains topological connectivity of the features (Stanislawski 2009). Network pruning is based on upstream drainage area, NHD reachcode addresses, and feature names from the Geographic Names Information System (GNIS) database (Stanislawski 2009, Yost and Carswell 2009). Polygonal hydrographic features and other associated linear features are then pruned using size-based and other criteria determined from NHD Standards (U.S. EPA and U.S. DOI 1999). Retained features are simplified to reduce excess details using operations that are tailored for local geographic context (Buttenfield, Stanislawski, and Brewer 2011, Stanislawski and Buttenfield 2011a). An additional toolset is available to validate results through conflation and nonparametric statistical methods (Stanislawski, Buttenfield, Samaranayake 2010).

To facilitate production operations, a batch process fully automates the generalization workflow from enrichment through pruning and enables sequential processing of multiple subbasins. The batch generalization process normalizes, through pruning, the heterogeneous HR data to produce uniform scale data sets at 24K and several smaller scales for each subbasin. However, the process requires target density estimates for all density partitions in a subbasin at each desired scale. Over 2100 subbasins exist within the conterminous United States, rendering this step a substantial time-consuming bottleneck in the workflow.

To address this issue, an automated workflow to estimate the multi-scale target densities for all subbasins in the conterminous United States was developed and tested using commercially available and customized geoprocessing tools (Stanislawski, Falgout, and Buttenfield 2015). The process estimates natural drainage density patterns at 24K and smaller scales from elevation-derived drainage channels (Stanislawski and others 2012). The commercially-available geoprocessing tools function under a Windows operating system with single-threaded processing tools. The workflow requires 4 or more hours to complete for a single HUC8 subbasin. This workflow rate applied on a single-thread would require at least 8000 hours (about a year) of processing time to complete the more than 2100 subbasins in the country. Given that all contingencies have not been evaluated for the various conditions in the country, it is likely that refinements to the workflow will be needed, and a much faster processing alternative is critical to an efficient production implementation. Tests of the workflow using Windows emulation on a Linux cluster (Stanislawski, Falgout, Buttenfield 2015), which allows up to 10 simultaneous processing threads did not improve the throughput, because the Windows emulation server could not efficiently handle multiple simultaneous processing threads and its performance degrades due to the overhead of emulating Windows in Linux computing environment.

Consequently, the time-consuming tasks in the workflow have been converted to use open source geoprocessing methods, which can be efficiently implemented on a high-performance Linux computing cluster. The new open-source version of the workflow has demonstrated the ability to leverage the high throughput computing capability of the cluster for processing 30 subbasins in about one hour. The cluster used for this work has the capacity to process up to 100 subbasins simultaneously. The new open source workflow represents a feasible solution to achieve production goals for cartographic generalization. This paper reports progress towards estimating target drainage densities for the conterminous United States using the open source alternative on a Linux cluster. In addition, some generalization results based on the estimated target densities are presented.

# 2. Methods

## 2.1 Natural Drainage Density Estimation Workflow

The objective of target drainage density estimation is to estimate drainage density variations that only reflect the natural local variations of the geomorphological characteristics within a subbasin at 24K and smaller scales. Working at the HUC8 level is a logistically viable solution given the need to balance the processing volume for the nation with target density estimations that are sufficiently localized for realistic characterization of stream density patterns. A synopsis of the process follows, along with estimates of per subbasin processing times from the initial workflow using commercially available and customized geoprocessing tools on a Windows machine. For additional details see Stanislawski and others (2012) and Stanislawski, Falgout, Buttenfield (2015).

1) Flow network parameters for the total length, number of features, and minimum length of first order tributaries that exist at 24K are derived through a process that summarizes the 24K version HR NHD flowline features in the subbasin. The 24K version contains no data that were densified to finer resolutions. In addition, polygons are generated for areas devoid of 24K flowline features in the subbasin. Processing time for this step is about 2 to 6 minutes per subbasin using customized commercially available geoprocessing tools on a Windows machine.
2) The 24K flow network parameters and devoid area polygons (if identified in the subbasin) are used to extract drainage channels from the 1/3$^{rd}$ arc-second (nominally 10-meter cell resolution) digital elevation model (DEM) for the subbasin, which is available from the USGS 3D Elevation Program (3DEP). The algorithm extracts drainage channels using a weighted flow accumulation model, whereby weights are determined from factors that influence stream geomorphology; specifically, surface runoff, terrain slope, soil permeability, soil depth, vegetation cover, and ground water. The number and location of extracted channels is governed by the 24K network parameters, devoid areas, and weights. Processing time using customized Windows tools average about 4 hours per subbasin, but may take up to 30 hours depending on the size and complexity the subbasin.
3) Subsequently, line-density partition polygons that summarize line-density variations of HR NHD flowlines within a subbasin are generated. As mentioned previously, line-density variations may be caused by natural conditions or data compilation inconsistencies in the HR NHD. This process takes 1 to 4 minutes on a Windows machine.
4) The HR NHD density partitions are overlain with a raster line-density dataset generated for the 24K extracted drainage channels. Zonal averages of 24K density are computed for each HR NHD density partition. Then, target densities for 1:50,000-scale (50K) and smaller scales are estimated for each partition using the 24K partition densities and an adjusted Radical Law relation. The Radical Law was adjusted to fit stream densities derived from benchmark hydrographic datasets at 1:100,000, 1:500,000, 1:1,000,000, and 1:2,000,000. This step takes 5 to 10 seconds through a Windows process.

### 2.2 Open Source Tools on Linux Cluster

The drainage channel extraction process (step 2) is the computing-intensive step. Therefore, this step has been implemented through freely available open source tools. The Geospatial Data Abstraction Library (GDAL, http://gdal.org) is used for vector operations, and the Terrain Analysis Using Digital Elevation Models tools (TauDEM, https://github.com/dtarb/TauDEM) are used for raster operations that extract the elevation-derived channels. GDAL is deployed on a Linux cluster as C/C++/Python programming libraries. TauDEM employs a parallel programming model, i.e., the message passing interface (MPI), to enable spatial data decomposition, runtime communication among processors for data and message exchange, and parallel input/output (IO) for processing large DEMs beyond what desktop software can handle. Specific TauDEM functions used in the drainage channel extraction process include pit removal, determination of flow-direction, weighted flow accumulation, and channel extraction.

The process is developed as portable Python software and deployed on a five-node Linux compute cluster at USGS. Each node is comprised of 20 processing cores and 64 Gigabytes of shared Random Access Memory (RAM). Rapid access to file storage is achieved through a parallel Lustre file system on the high-speed Infiniband interconnect. All steps in the workflow are programmed in Python. To accelerate the numerical performance of a scripting language like Python, we also deployed Numba and LLVM libraries to provide native machine code for common numerical functions used in Python, instead of running the corresponding Python functions with runtime interpretation. The choice of Python is also better for portability to other GIS software environments because Python is more broadly supported than C/C++.

Job scheduling on the cluster is handled through the Simple Linux Utility for Resource Manager (SLURM). A shell script uses SLURM to request appropriate resources from the cluster and execute jobs. A single job performs the channel extraction process for one subbasin. SLURM automatically schedules the execution jobs to ensure full use of available processor and memory resources.

## 3. Preliminary Results and Discussion

### 3.1 Sample of Extraction with Open Source Tools

A sample subbasin of elevation-derived drainage channels extracted with the open source/TauDEM workflow are shown in Figure 1 in comparison to the existing HR NHD flowline stream features. The sample subbasin is the Lake Mead subbasin that includes the Grand Canyon. The large lake and relatively high variation in elevation and slope in this subbasin create complex conditions that are more difficult to extract a fully connected network of drainage channels than in subbasins with less variable conditions. Density variations that follow 7.5-minute topographic map boundaries (inset, Figure 1 a) are evident in the NHD stream features because of inconsistent data compilation. However, only natural density variations are evident in the elevation-derived channels.

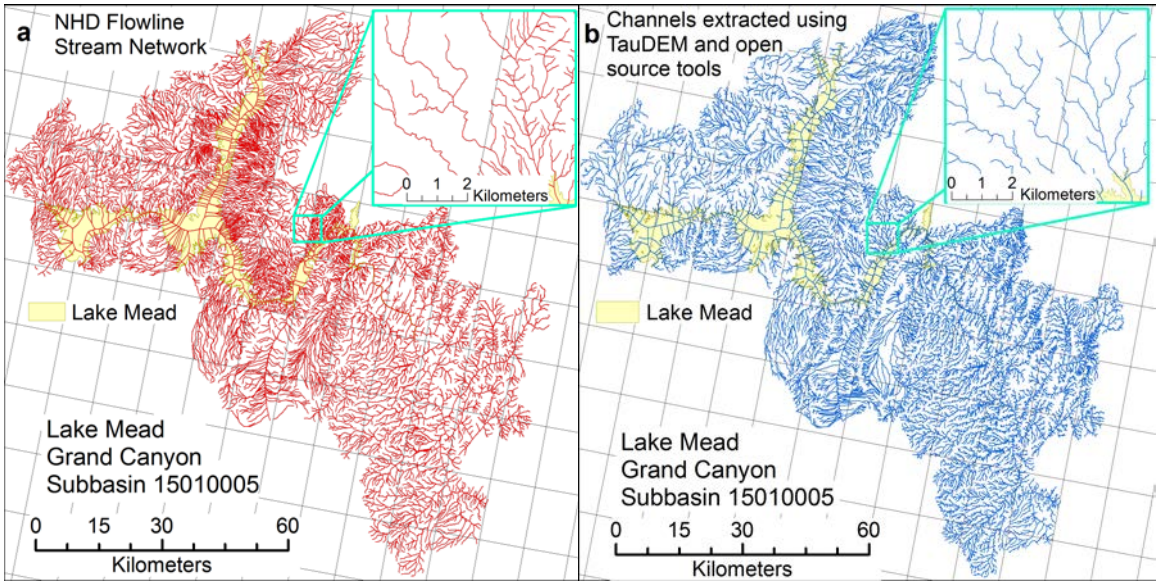


Figure 1. a) NHD Flowline features (red lines), and b) channels (blue lines) extracted from 1/3rd arc-second elevation model for the Lake Mead, Grand Canyon, National Hydrography Dataset (NHD) subbasin #15050005. Dark gray grid lines are 7.5-minute topographic map boundaries. Drainage channels were extracted using a weighted flow accumulation model implemented through open source and TauDEM tools.

The similarity of these two sets of linear features is estimated with the Coefficient of Line Correspondence (CLC), which is computed as the sum of the length of the matching features in both datasets divided by the sum of the length of all features in both datasets (Stanislawski, Buttenfield, and Doumbouya 2015). The CLC for the Lake Mead subbasin indicates 69 percent of the lines are matching between the two datasets, with 54 percent matching first order tributaries and 87 percent matching higher order tributaries. CLC values comparing the NHD flowlines to the extracted lines from the earlier commercial software methods show 65 percent matching, with 50 percent matching first order features and 84 percent matching higher order features. Thus, for this subbasin, channels extracted with the open source tools match slightly better with the NHD flowlines than channels extracted with the commercial tools.

A portion of mismatching first order (headwater) features can be attributed to cartographic constraints (minimum length thresholds) that limited the number of headwater features originally collected for the NHD (Colson and others 2008, Fritz and others 2013, Caruso 2014, Stanislawski, Buttenfield, Doumbouya 2015). However, many mismatching features are NHD features that are omitted from the derived channels because of inconsistent NHD compilation. That is, NHD stream networks on some maps are over collected (too dense) compared to adjacent maps (Figure 1a, see inset).

### 3.2 Scaling the Solution to Multiple Subbasins

The 24K natural drainage density pattern depicted through the extracted channels for the 16 subbasins in NHD subregion 0601 is shown in Figure 2. The number and density of extracted 24K channels in a subbasin are regulated by the subbasin parameters for the extraction process, which are derived from the best available estimate of 24K NHD flowlines. Within each subbasin the distribution of channel density is controlled through the weights that are modelled to reflect stream formation conditions, and this within subbasin relation is supported by a visual comparison of the

density pattern of subregion 0601 and the spatial pattern of associated model weights (to be presented at workshop). However, it is also evident in Figure 2 that some density variations follow subbasin boundaries, which may be an artefact of extracting channels by HUC8 subbasin. Further testing is needed to investigate whether some type of adjustment of the extraction parameters or the level of the extraction watershed (HUC8, HUC6, HUC4, etc.) could improve the natural drainage channel extraction process.

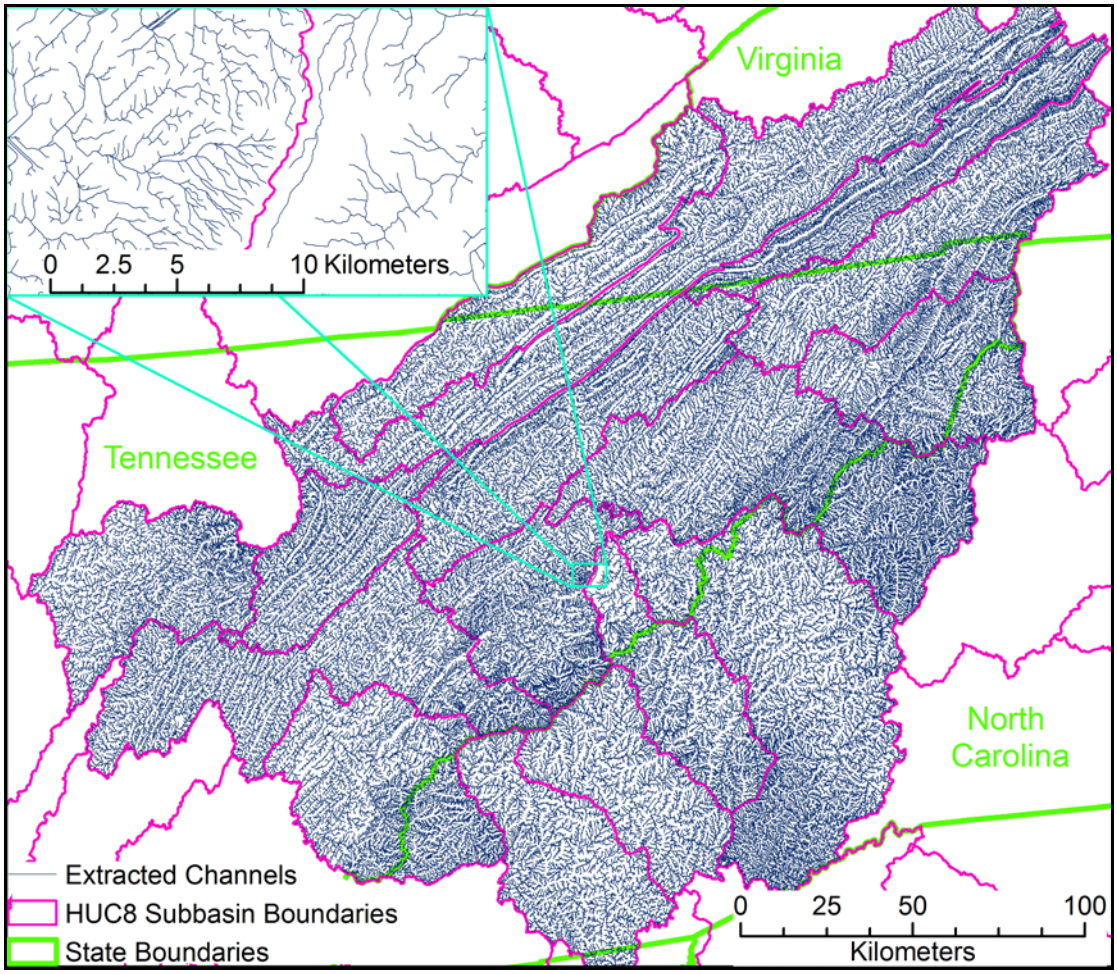


Figure 2. 1:24,000-scale extracted channels estimating natural drainage density pattern for the 16 subbasins in National Hydrography Dataset subregion 0601. Drainage channels are derived from 1/3rd arc-second elevation data using a weighted flow accumulation model.

Furthermore, validation and refinement of the model and parameters are needed. Passalacqua, Belmont, and Foufoula-Georgiou (2010) propose analysis of terrain curvature to model headwater starting points for geomorphic channel extraction from elevation data. Tarboton, Bras, Rodriguez-Iturbe (1991) suggest the use of flow accumulation thresholds that extract channels with a constant average elevation drop between stream orders. These alternatives can help validate or refine the weighted flow accumulation model. Therefore, the rapid data processing afforded through the open source high-performance computing is crucial to these research objectives.

### 3.3 Thinning to Reduce Compilation Inconsistency

The HR NHD features for HUC4 region 0601 are shown in Figure 3, which demonstrates the data inconsistency issues that exist in some parts of the HR NHD. Compilation inconsistencies in the NHD have been an issue of concern for some time. Obvious compilation inconsistencies are evident where flowline network densities vary along 7.5-minute map boundaries. Inconsistency is also evident where local resolution content has been included in the southeast subbasin and other smaller subwatersheds. The generalization workflow presented in this paper automatically identifies and resolves these inconsistencies.

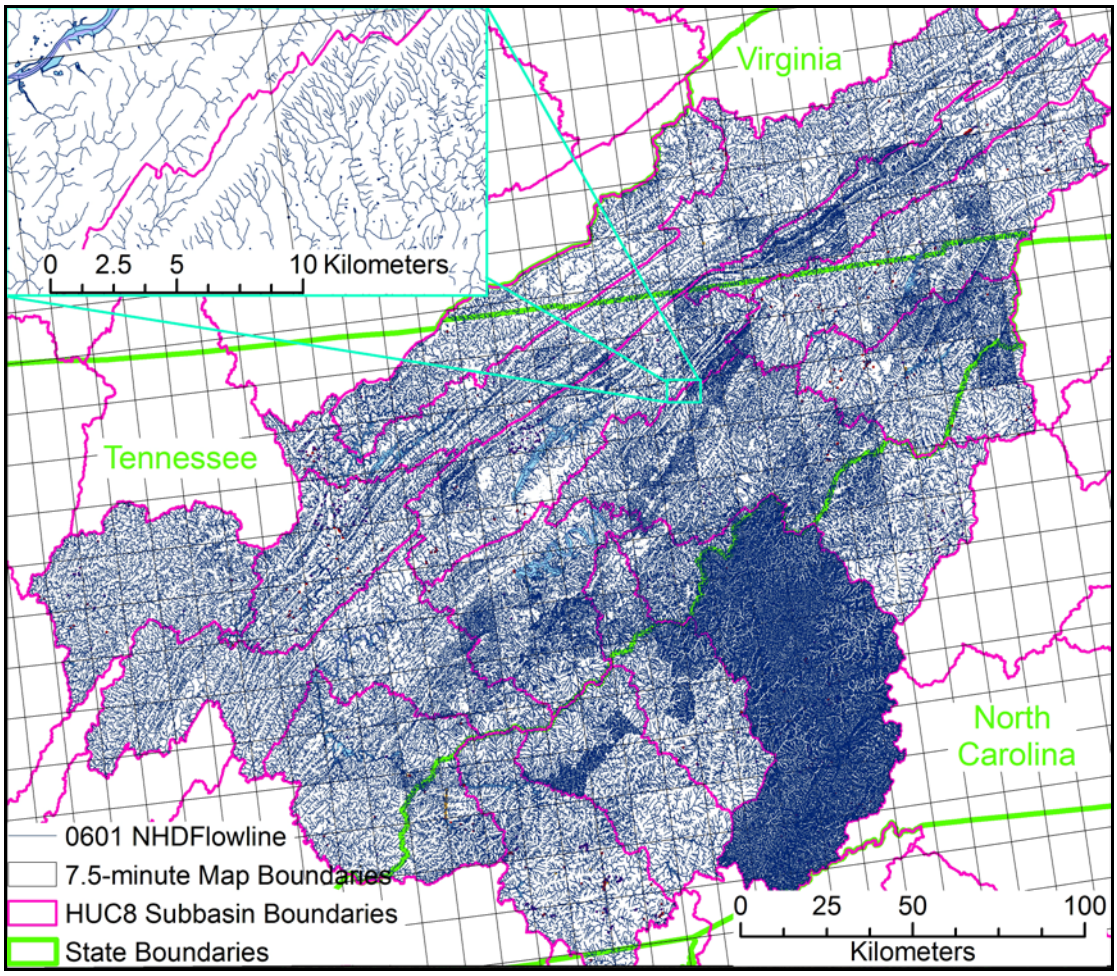


Figure 3. High-resolution National Hydrography Dataset features for HUC4 subregion 0601. 1:24,000-scale 7.5-minute map boundaries are shown with the black grid lines. Compilation inconsistencies that follow 7.5-minute map boundaries are evident, along with the inclusion of local resolution content in several subbasin and smaller watersheds.

Results of thinning the HR NHD features for HUC4 region 0601 to 1:100,000-scale (100K) is shown in Figure 4. Target 100K densities (along with several smaller scale densities) for the HR line density partitions are predicted from the 24K density pattern extracted through the weighted flow accumulation model (Figure 2). Stratified pruning included in the batch NHD Generalization process was used to thin these data

to 100K. Gross compilation inconsistencies have been corrected in these results. Additional scales of the data, which include 1:24,000, 1:50,000, 1:100,000, 1:250,000, 1:500,000, 1:1,000,000, 1:2,000,000, and 1:5,000,000, are furnished through the batch thinning process.

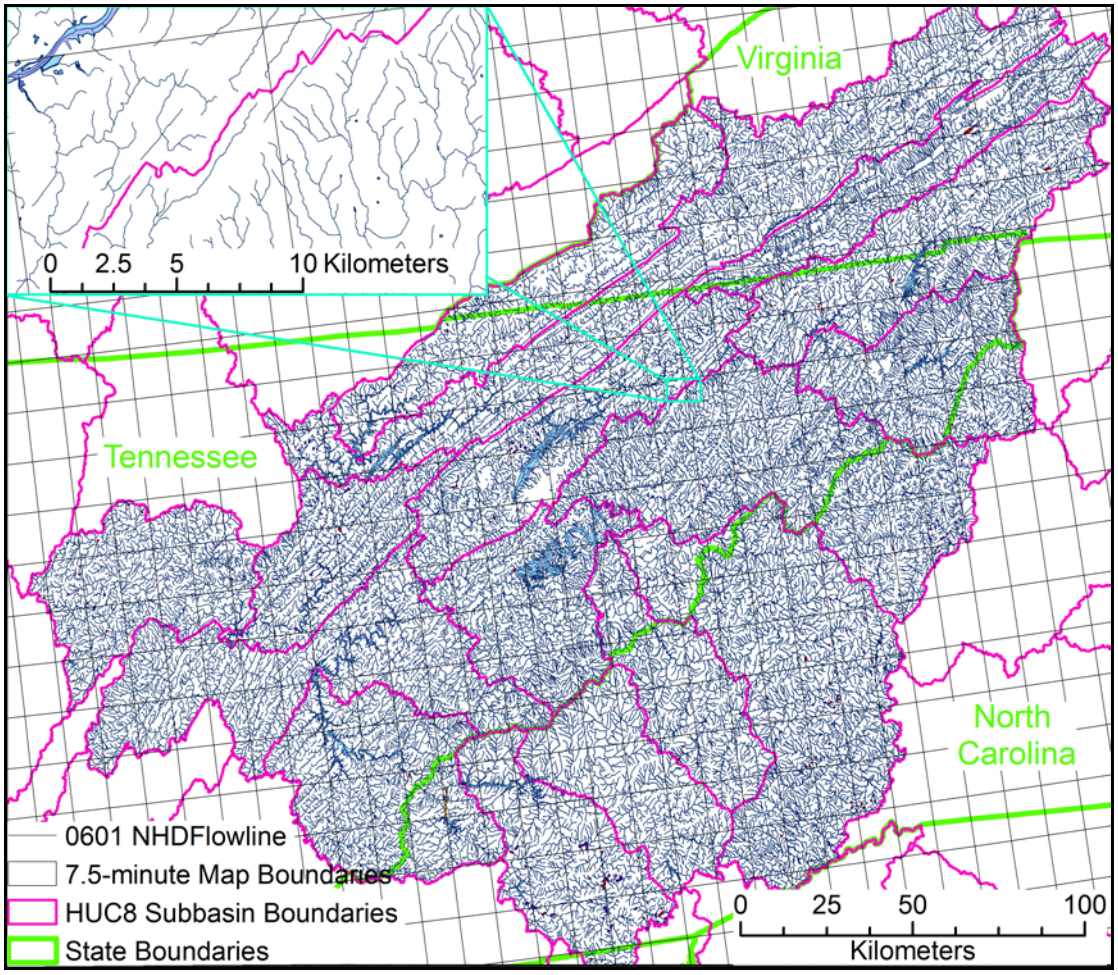


Figure 4. High-resolution National Hydrography Dataset features for HUC4 subregion 0601 thinned to 1:100,000-scale based on target density estimates derived through the weighted flow accumulation model.

## 4. Conclusions

This paper demonstrates results from an automated hydrography generalization capability that is supported through an open source/TauDEM workflow for estimating 24K natural stream density patterns within the conterminous United States. The generalization processing eliminates density variations caused by data compilation inconsistencies and maintains variations that reflect natural terrain conditions. Thus far, results indicate the 24K natural stream density estimation process is sufficient to generalize the HR NHD to 24K and smaller scales for watersheds with rough to partially rough terrain, but it does not perform as well in flat coastal plains. Alternative or augmented density estimation procedures are likely needed for relatively flat, coastal or swampy watersheds.

Further research could improve the 24K natural drainage density estimation process by extending its capability for hydrographic feature collection and hydrologic

analyses. Research options could test spatial modelling of the extraction parameters or processing larger watershed units (for instance, HUC6 or HUC4 basins). Further work should also validate the extraction process by comparison with other geomorphological approaches for extracting drainage channels.

The target density estimation process has been completed for 447 HUC8 subbasins in NHD regions 6, 7, and 10, which span various terrain and climate conditions within the conterminous United States. Processing the channel extraction step for these three regions takes less than 16 hours to complete using the open source/TauDEM methods on a Linux cluster. This is a substantial improvement over the commercially available software running on a Windows operating system, which can require longer run times for single subbasins. Moreover, implementing the commercial solution for simultaneous processing requires added complex programming. The high throughput approach achieved by the open source tools can process more NHD regions in the same amount of time by using more computing power available through government or academic supercomputing environments (e.g., http://xsede.org). The open source/TauDEM procedures provide better support for research and high volume cartographic mapping requirements for the United States than other commercial options that have been reviewed.

## 5. References


Buttenfield, B P, Stanislawski, L V, and Brewer, C A, 2011, Adapting generalization tools to physiographic diversity for the United States National Hydrography Dataset, *Cartography and Geographic Information Science* 38 (3):289–301. doi:10.1559/15230406382289.

Caruso, B. S. 2014. "GIS-Based Stream Classification in a Mountain Watershed for Jurisdictional Evaluation." *JAWRA Journal of the American Water Resources Association* 50 (5):1304–1324. doi:10.1111/jawr.12189.

Colson, T., J. Gregory, J. Dorney, and P. Russell. 2008. "Topographic and Soil Maps Do Not Accurately Depict Headwater Stream Networks." *National Wetlands Newsletter* 30 (3): 25–28.

Fritz, K. M., E. Hagenbuch, E. D'Amico, M. Reif, P. J. Wigington Jr., S. G. Leibowitz, R. L. Comeleo, J. L. Ebersole, and T.-L. Nadeau. 2013. "Comparing the Extent and Permanence of Headwater Streams from Two Field Surveys to Values from Hydrographic Databases and Maps." *JAWRA Journal of the American Water Resources Association* 49 (4): 867–882. doi:10.1111/jawr.2013.49.issue-4.

Passalacqua, P, Belmont, P, Foufoula-Georgiou, E, 2010, Automatic geomorphic feature extraction from lidar in flat engineered landscapes, *Water Resources Research* 48, W03528, doi:10.1029/2011WR010958.

Stanislawski, L V, 2009, Feature pruning by upstream drainage area to support automated generalization of the United States National Hydrography Dataset, *Computers, Environment and Urban Systems* 33 (5): 325–333. doi:10.1016/j.compenvurbsys.2009.07.004.

Stanislawski, L V, and Buttenfield, B P, 2011a, Hydrographic generalization tailored to dry mountainous regions, *Cartography and Geographic Information Science* 38 (2): 117–125. doi:10.1559/15230406382117.

Stanislawski, L V, and Buttenfield, B P, 2011b, A raster alternative for partitioning line densities to support automated cartographic generalization, *25th International Cartography Conference*, Paris, July 3-8.

Stanislawski, L V, Buttenfield, B P, and Doumbouya, A, 2015, A rapid approach for automated comparison of independently derived stream networks, *Cartography and Geographic Information Science* 42 (5): 435–448, http://dx.doi.org/10.1080/15230406.2015.1060869

Stanislawski, L V, Buttenfield B P, Samaranayake V A, 2010, Generalization of hydrographic features and automated metric assessment through bootstrapping, *13th Workshop of the International Cartographic Association Commission on Generalisation and Multiple Representation*, Zurich, September 12-13.

Stanislawski, L V, Doumbouya, A T, Miller-Corbett, C D, Buttenfield, B P, and Arundel-Murin, S T, 2012, Scaling stream densities for hydrologic generalization, in *Seventh International Conference on Geographic Information Science*, September 18-21, Columbus, Ohio.

Stanislawski, LV, Falgout, J, and Buttenfield B P, 2015, Automated extraction of natural drainage density patterns for the conterminous United States through high-performance computing, *The Cartographic Journal*, 52(2):185-192.

Tarboton, D G, Bras, R L, and Rodriguez-Iturbe, I, 1991, On the extraction of channel networks from digital elevation data, *Hydrologic Processes* 5(1): 81-100.

U.S. EPA, and U.S. DOI, 1999, Standards for National Hydrography Dataset – High Resolution, U.S. Environmental Protection Agency and U.S. Department of the Interior, National Mapping Program Technical Instructions, July. U.S. Geological Survey.

U.S. Geological Survey, 1955, Map publication scales, United States Geological Survey, in *Geological Survey Topographic Instructions*, Book 1.PartB, U.S. Geological Survey, Reston, Virginia.

U.S. Geological Survey, 2000, The National Hydrography Dataset: Concepts and Contents (February 2000), United States Geological Survey. http://nhd.usgs.gov/chapter1/chp1_data_users_guide.pdf.

Yost, A Y, and Carswell, W J Jr., 2009, Geographic Names: U.S. Geological Survey Fact Sheet 2009-3016, 2 p