2014

# HydroTerre Strahler Network Service For Any Level 12 HUC Catchment In The USA.

Lorne Leonard

Chris Duffy

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

# HYDROTERRE: SELECTING UP-STREAM LEVEL-12 HUCS USING DEPTH-FIRST GRAPHS ANYWHERE IN THE CONTINENTAL USA

LORNE LEONARD (1), CHRISTOPHER J DUFFY (1)
*(1): Department of Civil & Environmental Engineering, The Pennsylvania State University United States of America*

We demonstrate our strategy to enhance the National Hydrography Datasets so that a user can select all the up-stream Hydrological Unit Code-12s for large-scale hydrological modeling using depth-first graphs within HydroTerre. The fundamental issue in the selection process is determining flow direction between HUC-12s across the continental USA. This article discusses the strategies to correct the assumptions, data constraints, and how data is filtered to create the depth-first graphs. The depth-first graph is tested against CONUS major river basins by detecting if inner holes exist and by inspecting graph geodesic values. A simple two-level geodesic graph hierarchy is parsed for hydrological modeling. The Penn State Integrated Hydrological Model serves as an example, although the simple xml file generated for web services to represent the graph is adaptable for other models. The value added to NHD includes how each HUC-12 is connected with each other by specifying multiple directions if they exist, where the intersections occur, and all adjacent HUC-12 neighbors.

## INTRODUCTION

The National Hydrography Dataset (NHD) provided by the United States Geological Survey (USGS) is digital data containing national hydrological features such as streams, rivers, and lakes across the United States. NHD contains Geographic Information Science layers of watershed boundary datasets that contain Hydrologic Units that divide the US into the smallest to the largest hydrologic units [1]. In this article, we focus on the smallest unit, the level-12 Hydrological Unit Code (HUC-12), a sub watershed typically ranging from 10 to 40,000 acres in size [2]. The HUC-12 serves as the watershed boundary selection process for users retrieving data from HydroTerre http://www.hydroterre.psu.edu/ [3]. Users can download the essential terrestrial variables such as soils, land cover, and elevation from HydroTerre that serve as the minimal data for watershed modeling anywhere in the continental USA (CONUS). In this article, we address how a web user can gain access to all the HUC-12 identification keys that are upstream of the selected HUC-12 anywhere in the CONUS. The selection hierarchy is an important step towards database driven large-scale hydrological modeling on a distributed computing environment using high performance computing.

### Why create a derived NHD product?

First, we acknowledge the important and valuable service that USGS provides to the community by creating and supporting NHD data. We acknowledge that our use is not their intended audience, yet we hope our results will improve their data products. For this article, we are using NHDPlus version 2.1, dated January 15th 2013. NHD provides a field called HU_12_DS that represents the downstream designation from the selected HUC-12 [4]. By simply collecting all the unique HUC-12 keys, all 83,016 of them, and checking that the HU_12_DS keys exist, 38 unique HU_12_DS were found to not exist, and were assigned to 221 HUC-12s. Manually fixing 221 designation keys is resolvable; however, when the designation is not a neighbor to the selected HUC-12, computer graph algorithms will fail to make the connections. By simply creating a neighbor list of HUC-12s that border each unique HUC-12, the assigned NHD HU_12_DS field was compared to the neighbor list to identify which HUC-12s required inspection. Unfortunately, 1940 HUC-12s were found where the designation existed, but did not touch the source HUC-12. In some instances, the designation is technically correct, however, when there is a HUC-12 in between the source and target a graph algorithm will fail to create a connection. Again, it is possible to correct these manually and create geometry for those HUC-12s that do not exist, but the fundamental problem with the HU_12_DS field is the assumption of one designation. Logic errors such as these are not possible to resolve without creating new derived datasets that complement the existing NHD datasets, so that users can select all the upstream HUC-12 catchments. In this article, we demonstrate our strategy to enhance the NHD datasets so that a user can select all the upstream HUC-12s for large-scale hydrological modeling.

## METHOD TO CREATE GRAPH HIERARCHY

To have a HUC-12 upstream selection hierarchy, a graph data structure is needed, with a graph root representing the user's selected HUC-12. The depth-first search algorithm [5] [6] was chosen so the order of HUC-12 simulations could be calculated. By having this sequence, and knowing the stream connections between HUC-12s, it will be feasible to model every HUC-12 on a separate compute node, in a high performance-computing environment such as a compute cluster and simply share data (flow exchanges between HUCs) using Message Passing Interfaces [7]. As discussed in the introduction, the hurdle to create the graph network is determining the flow direction between HUC-12s. This section focuses on the attempted strategies and the chosen solution to determine the flow directions using NHD datasets. Also addressed are the strategies to correct the assumptions, data constraints, and how data was filtered to create the depth-first graph.

### Flow direction attempts

To have a selection hierarchy, a graph is needed that shows the connections between HUC-12s and the direction of the stream flow-line so we can identify upstream versus downstream. Each HUC-12 becomes a graph node. The graph edge was created by buffering each intersection (between the HUC-12 boundary and flow-lines) and identifying the HUC-12 at the intersections of the buffer circle and flow-line. By ignoring flow-direction, or by saying the graph is un-directed [8], this simple strategy selects all the downstream and upstream catchments using a depth-first graph (does not select closed basins as expected). However, the goal is to obtain only the upstream catchments, thus the flow-direction is critical. Rather than leaping in at the HUC-12 scale, a strategy was devised at the HUC-4 (1,271,879 to 30,542,646 acres) as only 202 catchments exist in the CONUS [9]. The devised strategy was to buffer the HUC and flow-line intersections with a 2-metre distance. Each intersection was assigned a unique identification and

then was buffered at 15, 45, 60, 75, and 90-meter intervals that were within the stream segment reach. HydroTerre uses 30-meter resolution elevation data and these buffer distances were chosen to obtain multiple elevation values. This elevation data was used to create flow direction grids using TauDEM software for the CONUS [10]. Then with each buffer, a set of new intersection points were created where the buffer intersected the same stream flow-line that went through the HUC boundary. Each stream flow-line segment has a unique identification to verify the same stream was used rather than adjacent streams. With this set of points, each HUC key was then assigned elevation and flow direction values. These techniques were successful at the HUC-4 scales due to using a coarser quality stream flow and smoother catchment boundaries. However, we had no success at the HUC-12 scale that would not require significant time spent checking every catchment by hand/eye. Why these steps failed at the HUC-12 catchment scale are predominately due to regions being flat within the stream reach. Thus, slope was negligible and did not inform us of flow direction. Calculated flow direction conflicted with the vector direction of the HUC-12 intersection and stream flow path. Short stream reaches and the lack of "smooth" stream reaches and/or catchment boundaries produced erroneous results making the technique ineffective.

**Technique to determine and correct flow direction**

As described earlier, each HUC-12 is a graph node and the main issue that requires resolving is the direction of the graph edge. The chosen solution is based on the assumption that the digitized direction is from down-stream to up-stream. Unfortunately, it is unclear and unlikely that all professionals involved with the stream flow-line generation took this approach, but the results are adequate for our goals. Each flow-line was converted to polylines, and the vertices that constitute the polyline are scaled from zero (start) to 100 percent (end). With all the HUC-12 and stream flow-line intersections, each was buffered with a distance of 2 millimeters. Having a very small buffer eliminates the issues described earlier with short stream reaches, vertices that "jump" to new locations or automatically dissolved to other edges that exist within the stream flow-lines. Each circle buffer was assigned a unique identification for any manual or automated corrections. Then, the intersection points created between the flow-lines and the buffer circle are merged with the polyline flow-lines. This merger creates new vertices within the polyline that are assigned a percentage that represent the vertex position along the polyline. By having these percentages, it is now possible to identify whether the vertex is downstream (lower percent) or upstream (higher percent) in relation to the intersection of HUC-12 boundary and stream flow-line. In addition, these points are assigned a unique key, the underlying HUC-12 key and the unique identification of the circle buffer. Now we have all the graph edges with directions.

Upstream depth-first graphs for all the major river basins in the CONUS were generated to create a selection list of HUC-12s. The selection list is then converted to a polygon shape file and dissolved to eliminate individual HUC-12 boundaries. If there were holes within the polygon (programmatically detected as inner polylines), these are either closed basins or HUC-12s where the assumption is incorrect. HUC-12s marked as closed basins were automatically added. HUC-12s where the assumption failed were visually inspected and added to a correction file. This entire process was repeated until no more holes were detected within the shape file. The main reason for regions not being included were due to streams going back and forth across the HUC-12 boundary. Within the developed graph application, all these connections are collected and compared to the NHD version. If they all match, then the connection is assumed correct. However, when the HUC-12 NHD dataset designation did not match and the stream flow

geometry conflicted, the developed software flagged them and they were added to a correction list edited by hand that overrides the graph network. It is likely there are cases when this is incorrect, but we have to assume that the geometry is correct with regard to catchment boundaries and stream delineation. Thus, as USGS updates NHD data products we can simply update the graph by using the newest product and create a correction list. Rather than "correcting" stream and catchment geometry ourselves that may conflict with the national product.

**Constraints: closed basins and filtered HUC-12s**

There are numerous scattered areas that are simply closed basins. This result was expected as no streams within these HUC-12s cross the catchment boundary. However, what should the protocol be when the selected hierarchy has a closed basin on the outer edge that meets the hierarchy? Should the graph only select the individual closed basin and ignore all the HUC-12s that feed into the closed basin? Is it acceptable to assume that a closed basin surrounded by non-closed basins be added to the selection hierarchy? These questions depend on the use of the graph network data and are model dependent. The goal of the derived data product is to be model agnostic and each of these cases are tagged in the results so the model software can determine the course of action that suits the user and software needs. However, to evaluate that the graph hierarchy creates no inner holes, the closed basin and any hierarchies that feed into the closed basin are automatically added to the graph as discussed earlier.

HUC-12s along the ocean with their designation as the ocean were not included in the graph hierarchy due to the way they were delineated in the NHD dataset with buffers along the coastline. HUC-12s cover long and thin coastal regions that would produce a selection hierarchy of the entire east coast. The Great Lakes have hundreds of HUC-12s that have streams that reach them and while technically correct, are not practical for our goals. Thus, an ignore selection criteria was created for oceans and lakes identified as HUC-12s. The other constraint is HUC-12s along the Canadian and Mexican borders that connect to watersheds within these two countries that are not included by NHD. The most significant problems are HUC-12s that are part of the Columbia, Pen Oreille and Flathead rivers in the Pacific North and Canada. The CONUS HUC-12s are not digitized into Canada, but future NHD versions will include this information [11].

## RESULTS

The graph hierarchy has been tested in twi ways; the first is by checking for missing HUC-12s within the major river basins. The second is by creating graphs for each HUC-12 and then determining the highest graph geodesic, or graph distance [12] [13]. This section describes these results as well as the interface for web users to gain access to the data and the resultant xml file for hydrological modeling use. The section concludes with a simple demonstration explaining how the xml file is used in Penn State Integrated Hydrological Modeling (PIHM) tools [14].

**Checking against major river basins**

The major river basins were used to determine that the catchment generated from the selection hierarchy contains no holes. This was programmatically done by determining that no inner holes existed within the shape file and the results are shown in Figure 1. The hierarchy graph was generated by designating the graph root at, or near, the river mouth. These are not definitive representations of the river basins, but clearly indicate that the data generation for the graph selection is working for large areas within the defined constraints.

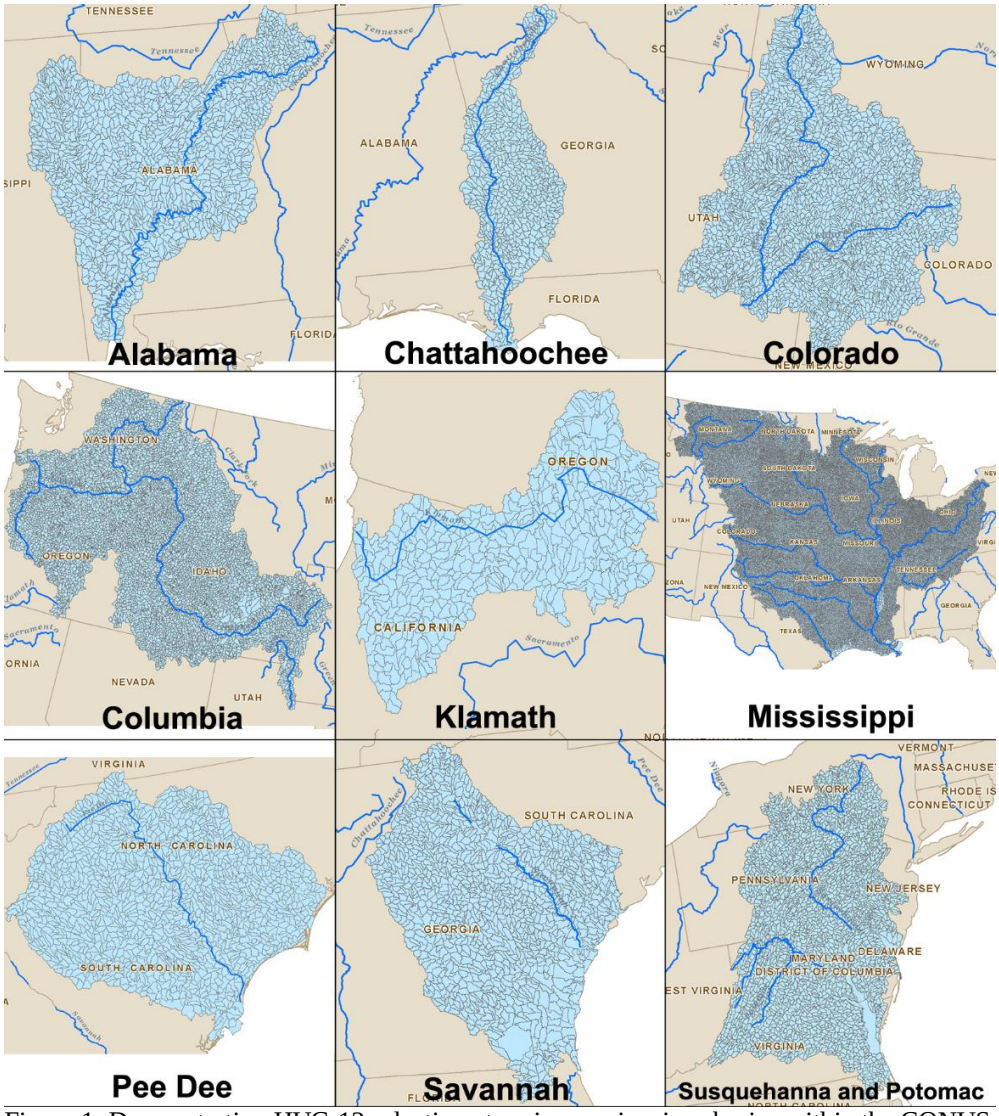

Figure 1: Demonstrating HUC-12 selection at various major river-basins within the CONUS. Black lines are the CONUS HUC-12s. Shaded blue areas are the selected HUC-12s. Dark blue lines are the coarse river flow lines.

**Highest graph geodesic**

The second method for evaluating the graph selection hierarchy was to test every HUC-12 in the CONUS and identify the highest graph geodesic at each HUC-12. Within the CONUS, the geodesic range is from one to 560. For example, you would expect regions along the mouth of the Mississippi river basin, which is the largest basin in the CONUS, to reflect the highest graph geodesics in the United States. Graph geodesic does not represent the number of HUC-12s in the graph but a way to group or bin HUC-12s. Binning HUC-12s by their graph geodesic is a naïve way to simulate similar graph geodesic HUC-12s at one time. This is not necessarily the best method, but is sufficient for smaller graph geodesics that cover large portions of the CONUS. The goal is to establish a hierarchy that will enable modeling on small cluster environments, but is also independent from the model. It is expected that users of these data hierarchies will adapt

the structures for their modelling needs and performance goals. Graph geodesics below 64 represent the majority of HUC-12s that could be modeled in this manner. By knowing the average number of HUC-12s with range at each geodesic, it is now possible to estimate how much input and output data, memory, and network speed is required, when a user selects by graph geodesic. This will help us create thresholds for distributing model services via web services.

**Web interface to gain access to selection hierarchy xml file**

Access to the HUC-12 selection hierarchy is available via a web browser at http://www.hydroterre.psu.edu/Development/HUC12_hierarchy/HUC12_Selection_Hierarchy.aspx . There are four steps to get access to the data. First, the user specifies their email address in order to receive an email about where to download the data from the service via a web link. The second step is for the user to select a HUC-12 watershed using the point tool and zooming into their area of interest. The third step is for the user to identify the reason for the download so the service can collect information about how the data is used. The last step is for the user to click on the generate button and within a few minutes the xml file is generated, and an email sent to the user.

**Summary of selection hierarchy xml file**

As an example, the HUC-12 key 051402060704 was selected as the root HUC-12. In approximately 15 seconds, 6445 HUC-12s were selected as part of the upstream graph and an xml file was created. The selection hierarchy generated is an xml file due to its flexibility and ease of sharing data independent of software and operating systems. Briefly, the structure is as follows; a summary description that includes version, root source, number of unique HUC-12s, and number of closed basins. There is a separate list of HUC-12s versus closed basins, allowing modelers to determine on their own how to use the closed basin data.

The HUC-12 hierarchy list includes the highest graph geodesic value followed by a list of HUC12 objects. Each object includes the number of stream intersections, the graph geodesic, and HUC-12 identifications. The graph geodesic starts at zero, which represent the root node that the user selected. Part of the HUC-12 object is a NHD object, which indicates if this HUC is a closed basin or ocean HUC. The original NHD designation (NHD_DS) object is included for users to compare results and have alternative solutions. Each HUC_12 object includes a list of neighboring HUC-12s that touch its boundaries. Having neighbors also aids in users creating a different type of graph than the one presented. Each HUC-12 object has a connection list, indicating target (downstream) and source (upstream) of all streams that exit the HUC-12. The streams that enter/exit the HUC-12 are listed as a set of intersection objects, which indicate on what point the HUC-12 lies on, the X and Y position in Albers_Conical_Equal_Area coordinates, and the unique identification stored by the database.

A closed basin hierarchy list follows and is similar to the HUC-12 hierarchy list. When software reading the HUC-12 list encounters a closed basin, the software will need to access the closed basin hierarchy to retrieve its properties. That way, if the user wants to ignore closed basins, or treat them separately, they can. If a closed basin has multiple hierarchies upstream, these will also be included within the closed-basin object. The final object included in the xml file is a “where string” list of all unique HUC-12s included in both the HUC-12 and closed basin hierarchy lists. That way, any GIS software can quickly create a selection list based on object identification,

maximum width representing number of HUC-12s per where string row. Further details about the xml file structure are available at the http://www.hydroterre.psu.edu/ website.

**Simple demonstration**

In this section, one way the xml could be parsed for hydrological modeling is explained using PIHM to distribute HUC-12 modeling in a cluster or cloud configuration. The first step is to use the "WhereClause" object to select all the HUC-12s for a visual inspection of the catchment boundary representing the area of interest. The next step is to determine the amount of disk storage required. On average, 10 gigabytes of data is required as PIHM inputs per HUC-12 for 30-year duration. Thus, one can quickly determine from "UniqueCombinedHUC12s" object the amount of storage needed. The same process is necessary for PIHM model outputs, 30 gigabytes of output per HUC-12 are needed, and we can calculate output storage requirements. This is a trivial example, but an important feature for larger catchment studies.

Assuming disk storage is available, the next step is to create three lists of "HUC12" objects. The first list is created by parsing all the "HUC12_Hierarchy" objects. Within each HUC12 object the "graph geodesic" and "HUC12_ID" is stored. Followed by the "HUC12_Connection" and the "Intersection" locations. The connections inform us which direction the streams follow. Thus, once PIHM has finished a simulation time step, the results are parsed to the next HUC-12 simulation at the locations identified in the "Intersection" list. The second list is the result of parsing all the "Closed_Basins_Hierarchy" objects. The process is identical to parsing the "HUC12_Hierarchy" objects list except, if the closed basin is only one HUC-12, the "Highest_Graph_Geodesic" equals zero, this HUC-12 is added to the second list. Thus, all these simulations can be run independently. Otherwise, if a closed basin contains a hierarchy list, these are added to the third list. The third list object is independent of the first list thus is simulated separately. In hydrological modeling, one would start from the highest graph geodesic and work backwards to the root. Since the selection hierarchy could be used for other purposes, the graph geodesic is left intentionally this way, to keep the root node as one object.

## CONCLUSION

In this article, we describe the steps to create level-12 Hydrological Unit Code (HUC-12) depth first search hierarchy graph networks (Strahler) for any HUC-12 in the CONUS. The graph process is derived by using United States Geological Survey National Hydrology Datasets (NHD). NHD specifies a HUC-12 designation (HU_12_DS) and we describe the issues we encountered. The fundamental problem with HU_12_DS is the assumption of only one designation. Our attempts to resolve flow direction between HUC-12s were described and the assumption that digitized flow lines are from down-stream to up-stream. Included are the strategies to correct the assumptions, data constraints, and how data was filtered to create the depth-first graphs with HUC-12s. Hierarchy graph networks for a subset of major river basins were demonstrated with a discussion of how we validated our strategy. How users can access the graph as an xml file format via a web application was described, with a simple demonstration of how the depth first search hierarchy graph can be used in Penn State Hydrological Model for simulating HUC-12s in the CONUS. However, the xml file structure is simple to parse and we anticipate other uses. The value added to NHD includes how each HUC-12 is connected with each other by specifying multiple directions, if they exist, where the intersections occur, and all adjacent HUC-12 neighbors.

## FUTURE DIRECTIONS

Although every HUC-12 hierarchy has been generated for the CONUS and checked for holes within the entire catchment polygon, we acknowledge this does not guarantee the accuracy of the catchment. In addition, the graph is not accurate along the Canadian and Mexican borders as data is missing. While validating major river basins against HUC-4 boundaries, HUC-12s were identified that require intervention, as a stream flows in two directions across the boundary. We propose that difficult HUC-12 direction flow assignments be open to the community for discussion and resolution. For example, perhaps closed basins are modified to include groundwater flow between HUC-12 boundaries, creating a new selection hierarchy. In the meantime, we foresee the next version of this tool suite to enable web users to identify selection hierarchies that are inaccurate, flag for discussion by the community and update the graph database via online tools, enabling versioned datasets that adapt to community needs. This tool suite will be incorporated into the existing HydroTerre infrastructure so users can gain rapid access to Essential Terrestrial Variables for major river basins anywhere in the continental USA.

## REFERENCES


[1] *Seaber, P. R., Kapinos, F. P., & Knapp, G. L., "Hydrologic unit maps", U.S. Geological Survey, (1987).*

[2] *United States Geological Survey.,"USGS HUC", Retrieved from http://water.usgs.gov/GIS/huc.html , (2013).*

[3] *Leonard, L., & Duffy, C. J, "Essential Terrestrial Variable data workflows for distributed water resources modeling.", Environmental Modelling & Software, Vol.50, (2013), pp 85-96*

[4] *McKay, L., Bondelid, T., Dewald, T., Rea, A., Johnston, C., & Moore, R., "NHD Plus Version 2 : User Guide", (2012).*

[5] *Skiena, S. S., "The algorithm design manual", Santa Clara, CA: TELOS--the Electronic Library of Science, (1998).*

[6] *Cormen, T. H., Leiserson, C. E., & Rivest, R. L., "Introduction to algorithms", Cambridge, Mass.; New York: MIT Press ; McGraw-Hill, (1990).*

[7] *Gropp, W., Ewing, L., & Rajeev, T., "Using MPI-2: advanced features of the message-passing interface", Cambridge, Mass.: MIT Press, (1999).*

[8] *Trudeau, R. J., "Introduction to Graph Theory", New York: Dover Pub, (1993).*

[9] *United States Department of Agriculture., "Watershed Boundary Dataset", Retrieved from http://www.nrcs.usda.gov/ ,(2013).*

[10] *Tarboton, D. G., "TauDEM Hydrology Research Group", Retrieved from http://hydrology.usu.edu/taudem/taudem5.0/index.html ,(2011).*

[11] *Simley, J. "USGS National Hydrography Dataset Newsletter", Retrieved from http://nhd.usgs.gov/newsletters/News_Nov_13.pdf, ,(2013).*

[12] *Harary, F., "Graph theory.". Mass: Addison-Wesley, (1969).*

[13] *Buckley, F., "Distance in graphs.", Redwood: Addison-Wesley, (1990).*

[14] *Qu, Y., & Duffy, C. J., "A semidiscrete finite volume formulation for multiprocess watershed simulation.", Water Resources Research, Vol.43, No. 8(2007)*