r.accumulate: Efficient computation of hydrologic parameters in GRASS

Huidae Cho, Ph.D., GISP, PE (MD), M.ASCE, CFM
Institute for Environmental and Spatial Analysis
University of North Georgia
GRASS GIS Core Developer

February 2, 2020
FOSDEM 2021
Me

- Huidae Cho /hɪdɛ tɕo/
- Teaching geospatial science and computing at the University of North Georgia
- GRASS GIS core developer (20 years)
- ArcGIS developer (12 years)
- Water resources engineer (10 years)
- https://github.com/HuidaeCho
- grass4u@gmail.com
Source code

- https://github.com/OSGeo/grass-addons/tree/master/grass7/raster/r.accumulate
Web-based hydrologic modeling system (WHydroMod)

- `r.topmodel`: A GRASS module for the Topography Model (TOPMODEL)
- USGS online data
- NCDC web API
WHydroMod for Texas

- http://txmod.isnew.info
- Proof-of-concept implementation for Texas
- GNU Affero General Public License Version 3
- Transparent system to the user
- Open source stack
  - GRASS GIS
  - PostgreSQL
  - PyWPS
  - MapServer
  - OpenLayers
  - Apache
- Bing Maps API for base maps
Typical TOPMODELing
The WHydroMod way
Data flow
Challenges

Web users are impatient

But it takes minutes to download online data & process DEM

- Blue & red USGS gages
- Automate heavy data processing over night
- Allow the user to initiate new processing and share results
- **Improve the performance of geospatial computation**
Important hydrologic parameters

- Flow direction
- Flow accumulation
- Longest flow path
Flow direction

Figure 1: GRASS GIS drainage encoding
Flow accumulation

A raster (think of a matrix with cell values) that downtraces rain drops following flow directions.
Longest flow path

FP is the watercourse from one point to another.

\[ \overrightarrow{LFP} \in \left\{ \overrightarrow{FP}_i \mid \| \overrightarrow{FP}_i \| \geq \| \overrightarrow{FP}_j \| \ \forall j \neq i \right\} \]

Yes! LFP is the “longest” flow path.

There can be more than one in some case.
Motivation

The current LFP algorithm

- Peter Smith (1995) → the resolution of DEM was limited
- grid-based → can take a long time
- grid output → potentially invalid vector output
- calculation of two flow length grids for each outlet
- not for a large number of watersheds

Peter Smith’s method
Downstream flow length

Flow length starting from the outlet

$FL_{i,j}$ is the flow length from cell $i$ to cell $j$.

$$DFL_i = \begin{cases} 0 & i = 0 \text{ at the outlet} \\ FL_{i-1,i} + DFL_{i-1} & i \geq 1 \end{cases}$$

Note

- 0 at the outlet
- high at a headwater
Upstream flow length

Flow length starting from a headwater

\[ UFL_i = \begin{cases} 
0 & \text{i = 0 at a headwater} \\
FL_{i-1,i} + \max(UFL_{i-1}) & \text{i \geq 1}
\end{cases} \]

Note

- 0 at a headwater
- high at the outlet
- taking the maximum at a confluence
DFL + UFL

Note

- maximum DFL at the headwater on the LFP
- maximum UFL at the outlet
- both maximums are the same → LFL (longest flow length)
- any DFL+UFL cells on the LFP have this maximum value
LFP

LFP defined by all the cells with a value of LFL

\[ \text{LFL} = \max (DFL + UFL) \]
Typical procedure for multiple outlets

1. Set `i = 1` and `n = Number of outlets`.
2. If `i ≤ n`, go to step 3; otherwise, stop.
3. Delineate the watershed for outlet `i`.
4. Clip FDR to the watershed.
5. Calculate DFL.
6. Calculate UFL.
7. Store LFL.
8. Convert the extracted cells to line vector LFL.
9. Extract DUFL cells with a value of LFL.
10. Calculate DUFL.
11. Calculate UFL.
12. Calculate LFL = max(DUFL).

Stop if `i = n`. Repeat steps 2-11 until `i = n`.

Note: FOSDEM 2021
Critical problem

Green arrows: flow directions, blue line: LFP

Hydrologically invalid!
Other problems

Slow!

*Esri’s Flow Length tool is limited.*
Divide-and-conquer approach

Purely vector-based approach

r.accumulate GRASS GIS addon
Divide

Divide the problem into smaller pieces in a recursive way.

\[ \overrightarrow{\text{LFP}}_i \in \left\{ \overrightarrow{\text{LFP}}_j + \overrightarrow{\text{FP}}_{ji} \; \forall j \in \text{UP} \right\} \]

where

- \( \text{UP} = \{\text{Upstream neighbors of cell } i\} \)
- \( 0 \leq \|\text{UP}\| \leq 8 \)

Define a function of \( j \) that returns the length of the longest flow path at cell \( i \).

\[
 f(j) = \begin{cases} 
 \|\overrightarrow{\text{LFP}}_j\| + \|\overrightarrow{\text{FP}}_{ji}\| = \|\overrightarrow{\text{LFP}}_i\| & \text{if } \text{UP} \neq \emptyset \\
 0 & \text{otherwise}
\end{cases}
\]

Now, the problem becomes finding all \( \arg\max_{j \in \text{UP}} f(j) \) by traversing upstream cells starting from the outlet. The search stops when \( \text{UP} = \emptyset \).
Conquor

Intuitively, the longest longest flow length

$$LFL_{\text{max}} = s \cdot FAC\sqrt{2}$$

where $s$ and FAC are the cell size and flow accumulation, respectively

Based on Hack’s law, the shorted longest flow length

$$LFL_{\text{min}} = s\sqrt{FAC}$$

Between upstream neighbor cells $i$ and $j$, cell $i$ is on the LFP if $LFL_{\text{min},i} > LFL_{\text{max},j}$. 
Stack overflow and solution

- Depending on the watershed size, recursion can consume all stack memory and cause a stack overflow
- Convert recursion to iteration using a heap-based stack
- A single loop until the stack with upstream cells depletes
Benchmark experiments: r.accumulate vs. Arc Hydro

Is paid software always better than open and free software?
Data

- National Elevation Dataset (NED) 1 arc-second (approximately 30 meters)
- 27 NED maps
- National Atlas of the United State state boundaries for masking
- Generated 100 outlets randomly
System specifications

- CPU: Intel Xeon E5620 2.40 GHz
- Memory: 48 GB
- OS: Linux kernel version 4.4.14
- GIS: GRASS development version 7.7.svn revision r74124
Results and discussions
LFP geometries

Both methods produced almost identical LFPs except for some areas

Blue: r.accumulate, Red: Arc Hydro
Performance comparison with Arc Hydro
Efficient delineation of a massive number of subwatersheds
Conclusions

- The fewer raster operations, the faster!
- The performance of r.accumulate is linearly growing with the subwatershed size.
- While that of Arc Hydro exponentially growing.
- The new approach is cost-efficient and can be used for interactive hydrologic modeling (e.g., web).
References