Managing Hydrologic Data, Observations and Terrain Analysis

David Tarboton

Utah State University

http://www.engineering.usu.edu/dtarb/
Outline

• The CUAHSI Hydrologic Information System

http://his.cuahsi.org/

• Terrain Analysis Using Digital Elevation Models

http://hydrology.usu.edu/taudem
Hydrologic Data Challenges

- From dispersed federal agencies
- From investigators collected for different purposes
- Different formats
  - Points
  - Lines
  - Polygons
  - Fields

Data Heterogeneity
The way that data is organized can enhance or inhibit the analysis that can be done.

I have your information right here ...
It is as important to represent **hydrologic environments** precisely with data as it is to represent **hydrologic processes** with equations.
Data models capture the complexity of natural systems

NetCDF (Unidata) - A model for Continuous Space-Time data

ArcHydro – A model for Discrete Space-Time Data

Terrain Flow Data Model used to enrich the information content of a digital elevation model

CUAHSI Observations Data Model: What are the basic attributes to be associated with each single data value and how can these best be organized?
Data intensive science synthesizes large quantities of information (Hey et al., 2009).

- exploiting advanced computational capability for the analysis and integration of large new datasets to elucidate complex and emergent behavior.
- In hydrology, the image at left (Ralph et al., 2006) illustrates connection between extreme floods recorded in USGS stream gages and atmospheric water vapor from space based sensors.
- Satellite remote sensing and massive datasets enhance understanding of multi-scale complexity in processes such as rainfall and river networks.
CUAHSI HIS

The CUAHSI Hydrologic Information System (HIS) is an internet based system to support the sharing of hydrologic data. It is comprised of hydrologic databases and servers connected through web services as well as software for data publication, discovery and access.
The CUAHSI Community Hydrologic Information System

- University of Texas at Austin – David Maidment, Tim Whiteaker, James Seppi, Fernando Salas, Jingqi Dong, Harish Sangireddy
- San Diego Supercomputer Center – Ilya Zaslavsky, David Valentine, Tom Whitenack, Matt Rodriguez
- Utah State University – Jeff Horsburgh, Kim Schreuders, Stephanie Reeder, Edward Wai Tsui, Ravichand Vegiraju, Ketan Patil
- University of South Carolina – Jon Goodall, Anthony Castronova
- Idaho State University – Dan Ames, Ted Dunsford, Jiří Kadlec, Yang Cao, Dinesh Grover
- Drexel University/CUNY – Michael Piasecki
- WATERS Network – Testbed Data Managers
- CUAHSI Program Office – Rick Hooper, Yoori Choi, Conrad Matiuk
- ESRI – Dean Djokic, Zichuan Ye

CUAHSI
HIS
Sharing hydrologic data

http://his.cuahsi.org/

Support EAR 0622374
What is CUAHSI?

Consortium of Universities for the Advancement of Hydrologic Science, Inc.

- 110 US University members
- 6 affiliate members
- 12 International affiliate members (as of March 2009)

Infrastructure and services for the advancement of hydrologic science and education in the U.S.

http://www.cuahsi.org/
Data Searching – What we used to have to do

Searching each data source separately

Michael Piasecki
Drexel University
What HIS enables

- Searching all data sources collectively

Generic request

- NWIS
- NAWQA
- NARR
- MODIS
- NCEP
- ODM

Michael Piasecki
Drexel University
Finding Water Data with CUAHSI-HIS

Tim Whiteaker, Ph.D.
Research Associate
The University of Texas at Austin

http://his.cuahsi.org/movies/JacobsWellSpring/JacobsWellSpring.html
Web Paradigm

Catalog (Google)

Catalog harvest

Web Server (CNN.com)

Search

Access

Browser (Firefox)
CUAHSI Hydrologic Information System

Services-Oriented Architecture

**HydroCatalog**
Data Discovery and Integration

**HydroServer**
Data Publication
- ODM
- Geo Data

**HydroDesktop**
Data Analysis and Synthesis

Metadata Services

Search Services

WaterML, Other OGC Standards

Information Model and Community Support Infrastructure

OpenMI

R
What are the basic attributes to be associated with each single data value and how can these best be organized?

<table>
<thead>
<tr>
<th>Value</th>
<th>Offset</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>DateTime</td>
<td>OffsetType/</td>
<td>Quality Control</td>
</tr>
<tr>
<td>Variable</td>
<td>Reference Point</td>
<td>Level</td>
</tr>
<tr>
<td>Location</td>
<td>Source/Organization</td>
<td>Sample Medium</td>
</tr>
<tr>
<td>Units</td>
<td>Censoring</td>
<td>Value Type</td>
</tr>
<tr>
<td>Interval (support)</td>
<td>Data Qualifying</td>
<td>Data Type</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Comments</td>
<td></td>
</tr>
</tbody>
</table>
CUAHSI Observations Data Model

- A relational database at the single observation level (atomic model)
- Stores observation data made at points
- Metadata for unambiguous interpretation
- Traceable heritage from raw measurements to usable information
- Standard format for data sharing
- Cross dimension retrieval and analysis

A data value $V_i(s,t)$

Space, S

Time, T

Variables, V

“Where”

“When”

“What”
Stage and Streamflow Example

<table>
<thead>
<tr>
<th>ValueID</th>
<th>DataValue</th>
<th>ValueAccuracy</th>
<th>LocaDateTime</th>
<th>UTCOffset</th>
<th>SiteID</th>
<th>VariableID</th>
<th>MethodID</th>
<th>DerivedFromID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.18</td>
<td></td>
<td>05/01/2006 00:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>97</td>
<td>7.48</td>
<td></td>
<td>05/01/2006 00:00:00.000</td>
<td>-7</td>
<td>-7</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>193</td>
<td>7.22</td>
<td>22.88831542</td>
<td>05/01/2006 00:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>4.18</td>
<td></td>
<td>05/01/2006 01:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>98</td>
<td>7.48</td>
<td></td>
<td>05/01/2006 01:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4.17</td>
<td></td>
<td>05/01/2006 02:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>99</td>
<td>7.42</td>
<td></td>
<td>05/01/2006 03:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>7.42</td>
<td></td>
<td>05/01/2006 04:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>7.42</td>
<td></td>
<td>05/01/2006 05:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>4.17</td>
<td></td>
<td>05/01/2006 06:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>100</td>
<td>7.42</td>
<td></td>
<td>05/01/2006 06:00:00.000</td>
<td>-7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VariableID</th>
<th>VariableCode</th>
<th>VariableName</th>
<th>VariableUnitID</th>
<th>SampleMedium</th>
<th>ValueType</th>
<th>Irregular</th>
<th>TimeSupport</th>
<th>TimeUnitsID</th>
<th>DataTyp</th>
<th>GeneralCategory</th>
<th>NoDataValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>100006</td>
<td>Gage height</td>
<td>Gage height</td>
<td>1 Water</td>
<td>Field Observation</td>
<td>15</td>
<td>5</td>
<td>Continuous</td>
<td>Hydrologic</td>
<td>9999</td>
<td>Hydrologic</td>
<td>-9999</td>
</tr>
<tr>
<td>20000</td>
<td>Discharge</td>
<td>Discharge</td>
<td>2 Water</td>
<td>Derived Value</td>
<td>15</td>
<td>5</td>
<td>Continuous</td>
<td>Hydrologic</td>
<td>9999</td>
<td>Hydrologic</td>
<td>-9999</td>
</tr>
<tr>
<td>30000</td>
<td>Discharge, daily average</td>
<td>2 Water</td>
<td>Derived Value</td>
<td>24</td>
<td>Average</td>
<td>6</td>
<td>Instantaneous</td>
<td>Water Quality</td>
<td>9999</td>
<td>Hydrologic</td>
<td>-9999</td>
</tr>
<tr>
<td>40000</td>
<td>Dissolved oxygen concentration</td>
<td>3 Water</td>
<td>Field Observation</td>
<td>0</td>
<td>Instantaneous</td>
<td>6</td>
<td>Instantaneous</td>
<td>Water Quality</td>
<td>9999</td>
<td>Hydrologic</td>
<td>-9999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UnitsID</th>
<th>UnitsName</th>
<th>UnitsType</th>
<th>UnitsAbbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feet</td>
<td>Length</td>
<td>ft</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Cubic feet per second</td>
<td>Flow</td>
<td>ft³/s</td>
</tr>
<tr>
<td>3</td>
<td>Miligrams per Liter</td>
<td>Concentration</td>
<td>mg/L</td>
</tr>
<tr>
<td>4</td>
<td>Meters</td>
<td>Length</td>
<td>m</td>
</tr>
<tr>
<td>5</td>
<td>Minutes</td>
<td>Time</td>
<td>min</td>
</tr>
<tr>
<td>6</td>
<td>Hours</td>
<td>Time</td>
<td>hr</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MethodID</th>
<th>MethodDescription</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gage height measured with continuous data logger</td>
</tr>
<tr>
<td>2</td>
<td>Discharge derived from water stage using site specific rating curve</td>
</tr>
<tr>
<td>3</td>
<td>Daily average discharge derived from 15 minute continuous discharge values</td>
</tr>
<tr>
<td>4</td>
<td>Dissolved oxygen measured with a Hydrolab multiprobe field instrument</td>
</tr>
</tbody>
</table>
Independent of, but can be coupled to Geographic Representation

ODM

Observations Data Model

Sites
- SiteID
- SiteCode
- SiteName
- Latitude
- Longitude
...

CouplingTable
- SiteID
- HydroID

OR

e.g. Arc Hydro

Feature

HydroPoint
- HydroID
- HydroCode
- FType
- Name
- JunctionID

Waterbody
- HydroID
- HydroCode
- FType
- Name
- AreaSqKm
- JunctionID

Watershed
- HydroID
- HydroCode
- DrainID
- AreaSqKm
- JunctionID
- NextDownID

ComplexEdgeFeature

HydroEdge
- HydroID
- HydroCode
- ReachCode
- Name
- LengthKm
- LengthDown
- FlowDir
- FType
- EdgeType
- Enabled

SimpleJunctionFeature

HydroJunction
- HydroID
- HydroCode
- NextDownID
- LengthDown
- DrainArea
- FType
- Enabled
- AncillaryRole

HydroNetwork

Flowline

Shoreline
Importance of the Observations Data Model

- Provides a common persistence model for observations data

- Syntactic heterogeneity (File types and formats)
- Semantic heterogeneity
  - Language for observation attributes (structural)
  - Language to encode observation attribute values (contextual)

- Publishing and sharing research data
- Metadata to facilitate unambiguous interpretation
- Enhance analysis capability
WaterML and WaterOneFlow

WaterML is an XML language for communicating water data.
WaterOneFlow is a set of web services based on WaterML.

- Set of **query** functions
  - GetSites
  - GetSiteInfo
  - GetVariableInfo
  - GetValues

- Returns data in **WaterML**

```xml
<timeSeries>
  <sourceInfo xsi:type="SiteInfoType">
    <siteName>Colorado River at Austin, TX</siteName>
    <siteCode network="NWIS" siteID="4619631">08158001</siteCode>
    <geoLocation>
      <geogLocation xsi:type="LatLonPointType" srs="EPSG:4326">
        <latitude>30.24465429</latitude>
        <longitude>-97.694448</longitude>
      </geogLocation>
    </geoLocation>
  </sourceInfo>
  <variable>
    <variableCode vocabulary="NWIS" default="true" variable="Discharge">
      <variableName>Discharge, cubic feet per second</variableName>
      <units unitsAbbreviation="cfs" unitsCode="35">cubic feet per second</units>
    </variableCode>
    <values count="2545">
      <value dateTime="2006-12-31T00:00:00">129</value>
      <value dateTime="2006-12-31T00:15:00">129</value>
      <value dateTime="2006-12-31T00:30:00">129</value>
      <value dateTime="2006-12-31T00:45:00">129</value>
      <value dateTime="2006-12-31T01:00:00">124</value>
      <value dateTime="2006-12-31T01:15:00">129</value>
      <value dateTime="2006-12-31T01:30:00">124</value>
      <value dateTime="2006-12-31T01:45:00">124</value>
      <value dateTime="2006-12-31T02:00:00">124</value>
    </values>
  </variable>
</timeSeries>
```
WaterML as a Web Language

USGS Streamflow data in WaterML language

Discharge of the San Marcos River at Luling, TX
June 28 - July 18, 2002

This is the WaterML GetValues response from NWIS Daily Values
Open Geospatial Consortium
Web Service Standards

• Map Services
  - Web Map Service (WMS)
  - Web Feature Service (WFS)
  - Web Coverage Service (WCS)
  - Catalog Services for the Web (CS/W)

• Observation Services
  - Observations and Measurements Model
  - Sensor Web Enablement (SWE)
  - Sensor Observation Service (SOS)

These standards have been developed over the past 10 years .... .... by 400 companies and agencies working within the OGC

OGC Hydrology Domain Working Group evolving WaterML into an International Standard

http://www.opengeospatial.org/projects/groups/waterml2.0swg
HydroServer – Data Publication

Ongoing Data Collection

Point Observations Data

Historical Data Files

GIS Data

Internet Applications

Data presentation, visualization, and analysis through Internet enabled applications
HydroServer

Manager Interface

Browser Interface

Manager Interface

Browse Interface

Publish point data
Publish metadata
Publish spatial data
Service Interface

HydroServer

Capabilities Web Service
WaterOneFlow Web Service
Capabilities Configuration Tool
ODM Tools and Data Loaders

Capabilities Web Service
Spatial Data Services
Spatial Database
Geo Spatial Database
Spatial Data File

HydroServer Website
Map Website
Time Series Analyst Website

View server portal
View spatial data
View point data

Publish point data
Publish metadata
Publish spatial data

Load and Edit data
Configure server
HydroCatalog

- Search over data services from multiple sources
- Supports concept based data discovery

Service Registry

Hydrotagger

Harvester

Water Metadata Catalog

Search Services

Discovery and Access

CUAHSI Data Server

3rd Party Server e.g. USGS

WaterML

GetSites
GetSiteInfo
GetVariableInfo
GetValues

WaterOneFlow Web Service

http://hiscentral.cuahsi.org
Overcoming Semantic Heterogeneity

- ODM Controlled Vocabulary System
  - ODM CV central database
  - Online submission and editing of CV terms
  - Web services for broadcasting CVs

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigator 1:</td>
<td>“Temperature, water”</td>
</tr>
<tr>
<td>Investigator 2:</td>
<td>“Water Temperature”</td>
</tr>
<tr>
<td>Investigator 3:</td>
<td>“Temperature”</td>
</tr>
<tr>
<td>Investigator 4:</td>
<td>“Temp.”</td>
</tr>
<tr>
<td>Sunshine duration</td>
<td>Turbidity</td>
</tr>
</tbody>
</table>

From Jeff Horsburgh
Dynamic controlled vocabulary moderation system

ODM Data Manager

ODM Tools

ODM Controlled Vocabulary Website

XML

 ODPM Controlled Vocabulary Moderator

Master ODM Controlled Vocabulary

Local Server

Local ODM Database

ODM Data Manager

ODM Website

ODM Controlled Vocabulary Web Services

From Jeff Horsburgh

http://his.cuahsi.org/mastercvreg.html
HydroDesktop – Data Access and Analysis

- Integration from multiple sources
- Thematic keyword search
- Search on space and time domain
HydroModeler

An integrated modeling environment based on the Open Modeling Interface (OpenMI) standard and embedded within HydroDesktop

In development at the University of South Carolina by Jon Goodall, Tony Castronova, Mehmet Ercan, Mostafa Elag, and Shirani Fuller
Integration with “R” Statistics Package
37 Water Data Services on HIS Central from 12 Universities

- University of Maryland, Baltimore County
- Montana State University
- University of Texas at Austin
- University of Iowa
- Utah State University
- University of Florida
- University of New Mexico
- University of Idaho
- Boise State University
- University of Texas at Arlington
- University of California, San Diego
- Idaho State University

Dry Creek Experimental Watershed (DCEW)
(28 km² semi-arid steep topography, Boise Front)

68 Sites
24 Variables
4,700,000+ values

Published by Jim McNamara, Boise State University
Water Agencies and Industry

- USGS, NCDC, Corps of Engineers publishing data using HIS WaterML
- OGC Hydrology Domain Working Group evaluating WaterML as OGC standard
- ESRI using CUAHSI model in ArcGIS.com GIS data collaboration portal
- Kisters WISKI support for WaterML data publication
- Australian Water Resources Information System Water Accounting System has adopted aspects of HIS
- NWS West Gulf River Forecast Center Multi-sensor Precipitation Estimate published from ODM using WaterML
CUAHSI Water Data Services Catalog

69 public services
18,000 variables
1.9 million sites
23 million series
5.1 billion data values
(as of June 2011)

The largest water data catalog in the world

maintained at the San Diego Supercomputer Center
Open Development Model

- http://hydrodesktop.codeplex.com
- http://hydroserver.codeplex.com
- http://hydrocatalog.codeplex.com
Summary

• **Data Storage** in an *Observations Data Model* (ODM) and publication through HydroServer

• **Data Access** through internet-based *Water Data Services* using a consistent data language, called WaterML from HydroDesktop

• **Data Discovery** through a *National Water Metadata Catalog* and thematic keyword search system at Central HydroCatalog (SDSC)

• **Integrated Modeling and Analysis** within HydroDesktop

The combination of these capabilities creates a common window on water observations data for the United States unlike any that has existed before.
Terrain Analysis Using Digital Elevation Models (TauDEM)

David Tarboton\textsuperscript{1}, Dan Watson\textsuperscript{2}, Rob Wallace,\textsuperscript{3} Kim Schreuders\textsuperscript{1}, Jeremy Neff\textsuperscript{1}

\textsuperscript{1}Utah Water Research Laboratory, Utah State University, Logan, Utah
\textsuperscript{2}Computer Science, Utah State University, Logan, Utah
\textsuperscript{3}US Army Engineer Research and Development Center, Information Technology Lab, Vicksburg, Mississippi

This research was funded by the US Army Research and Development Center under contract number W9124Z-08-P-0420
Deriving hydrologically useful information from Digital Elevation Models

- Raw DEM
- Pit Removal (Filling)
- Flow Field
- Flow Related Terrain Information
A parallel version of the TauDEM Software Tools

- Improved runtime efficiency
- Capability to run larger problems
- Platform independence of core functionality

Deployed as an ArcGIS Toolbox with tools that drive accompanying command line executables, available from http://hydrology.usu.edu/taudem/
The challenge of increasing Digital Elevation Model (DEM) resolution

- **1980’s DMA** 90 m
  - $10^2$ cells/km$^2$

- **1990’s USGS DEM** 30 m
  - $10^3$ cells/km$^2$

- **2000’s NED** 10-30 m
  - $10^4$ cells/km$^2$

- **2010’s LIDAR ~1 m**
  - $10^6$ cells/km$^2$
Website and Demo

• [http://hydrology.usu.edu/taudem](http://hydrology.usu.edu/taudem)
Grid Data Format Assumptions

- Input and output grids are uncompressed GeoTIFF
- Maximum size 4 GB
- GDAL Nodata tag preferred (if not present, a missing value is assumed)
- Grids are square ($\Delta x = \Delta y$)
- Grids have identical extents, cell size and spatial reference
- Spatial reference information is not used (no projection on the fly)
Representation of Flow Field

Steepest single direction

\[
\begin{bmatrix}
48 & 52 \\
56 & 67 \\
\end{bmatrix}
\]

\[
\frac{67 - 52}{30} = 0.50
\]

Parallel Approach

- MPI, distributed memory paradigm
- Row oriented slices
- Each process includes one buffer row on either side
- Each process does not change buffer row
Illustrative Use Case: Delineation of channels and watersheds using a constant support area threshold

Steps

- Pit Remove
- D8 Flow Directions
- D8 Contributing Area
- Stream Definition by Threshold
- Stream Reach and Watershed
Pit Remove

Identifies all pits in the DEM and raises their elevation to the level of the lowest pour point around their edge. Pits are low elevation areas in digital elevation models (DEMs) that are completely surrounded by higher terrain. They are generally taken to be artifacts that interfere with the routing of flow across DEMs, so are

```
Executing: PitRemove logon & E:\Users\dtarb\Scratch\Logan\loganfel.tif
Start Time: Mon Sep 20 20:02:31 2010
Executing: CalculateStatistics E:\Users\dtarb\Scratch\Logan\loganfel.tif i l # E:\Users\dtarb\Scratch\Logan\loganfel.tif
Start Time: Mon Sep 20 20:02:33 2010
Executed (CalculateStatistics) successfully.
End Time: Mon Sep 20 20:02:33 2010 (Elapsed Time: 0.00 seconds)

Shell Command: mpiexec -np 6 "E:\Program Files (x86)\Taudem\TaudEN55\bin\PitRemove" -z "E:\Users\dtarb\Scratch\Logan\logan.tif" -fel "E:\Users\dtarb\Scratch\Logan\loganfel.tif" > "E:\Users\dtarb\Scratch\Logan\msgtmp.txt"
```

PitRemove version 5.0.4
Size: 8
Header read time: 0.059355
Data read time: 0.005522
Compute time: 0.234125
Write time: 0.077409
Total time: 0.376414

Executed PitRemove successfully.
End Time: Mon Sep 20 20:02:33 2010 (Elapsed Time: 2.00 seconds)
D8 Flow Direction (and Slope)

D8 Flow Direction

Calculates 2 grids. The first contains the D8 flow directions which are defined, for each cell, as the direction of the one of its eight adjacent or diagonal neighbors with the steepest downward slope. Flow Direction Coding: 1 - East, 2 - North East, 3 - North, 4 - North West, 5 - West, 6 - South West, 7 - South, 8 - South East.

D8FlowDir version 5.0.4
Size: 8
Header read time: 0.027037
Data read time: 0.009796
Compute Slope time: 0.102044
Write Slope time: 0.004855
Resolve Flat time: 0.353239
Write Flat time: 0.047082
Total time: 0.624034

Executed (D8FlowDirections) successfully.
End Time: Mon Sep 20 20:11:18 2010 (Elapsed Time: 3.00 seconds)
D8 Contributing Area

Calculates a grid of contributing areas using the single direction D8 flow model. The contribution of each grid cell is taken as one (or when the optional weight grid is used, the value from the weight grid). The contributing area for each grid cell is taken as its own contribution plus the contribution from upslope neighbors that...
Stream Definition by Threshold

Stream Definition by Threshold

- Input Accumulated Stream Source Grid: loganad8
- Input Mask Grid (optional)
- Threshold: 2000
- Input Number of Processes: 8
- Output Stream Raster Grid: E:\Users\dtarby\Scratch\Logan\logansrc.tif

Operates on any grid and outputs an indicator (1,0) grid identifying cells with input values $\geq$ the threshold value. The standard use is to use an accumulated source area grid as the input grid to generate a stream raster grid as the output. If you use the optional input mask grid, it limits the domain being evaluated to cells with mask values $\geq$ 0. When you use a D-infinity contribution area...
Stream Reach and Watershed

Stream Reach and Watershed

This function produces a vector network and shapefile from the stream raster grid. The flow direction grid is used to connect flow paths along the stream raster. The Strahler order of each stream segment is computed. The subwatershed draining to each stream segment (reach) is also delineated and labeled with the value identifier that corresponds to the WSNO (watershed number) attribute in the stream reach shapefile.
Some Algorithm Details
Pit Removal: Planchon Fill Algorithm

Planchon, O., and F. Darboux (2001), A fast, simple and versatile algorithm to fill the depressions of digital elevation models, *Catena*(46), 159-176.
Initialize( Z,F)

Do
for all grid cells $i$
    if $Z(i) > n$
        $F(i) \leftarrow Z(i)$
    Else
        $F(i) \leftarrow n$
    $i$ on stack for next pass
endfor
Send( topRow, rank-1 )
Send( bottomRow, rank+1 )
Recv( rowBelow, rank+1 )
Recv( rowAbove, rank-1 )

Until F is not modified

Z denotes the original elevation.
F denotes the pit filled elevation.
n denotes lowest neighboring elevation
i denotes the cell being evaluated

Iterate only over stack of changeable cells
Parallelization of Contributing Area/Flow Algebra

1. Dependency grid

Executed by every process with grid flow field \( P \), grid dependencies \( D \) initialized to 0 and an empty queue \( Q \).

\[ \text{FindDependencies}(P,Q,D) \]
for all \( i \)
for all \( k \) neighbors of \( i \)
if \( P_{ki} > 0 \) \( D(i) = D(i) + 1 \)
if \( D(i) = 0 \) add \( i \) to \( Q \)
next

2. Flow algebra function

Executed by every process with \( D \) and \( Q \) initialized from \( \text{FindDependencies} \).

\[ \text{FlowAlgebra}(P,Q,D,\theta,\gamma) \]
while \( Q \) isn’t empty
get \( i \) from \( Q \)
\( \theta_i = FA(\gamma_i, P_{ki}, \theta_k, \gamma_k) \)
for each downslope neighbor \( n \) of \( i \)
if \( P_{in} > 0 \)
\( D(n) = D(n) - 1 \)
if \( D(n) = 0 \) add \( n \) to \( Q \)
next \( n \)
end while

swap process buffers and repeat
Illustration of some other functions

Proportion flowing to neighboring grid cell 4 is $\alpha_1/(\alpha_1 + \alpha_2)$.
Proportion flowing to neighboring grid cell 3 is $\alpha_2/(\alpha_1 + \alpha_2)$.

D-Infinity Slope and Contributing Area

Value
High: 6.28319
Low: 0
Useful for modeling erosion and sediment delivery, the spatial dependence of sediment delivery ratio and contaminant that adheres to sediment
**Decaying Accumulation**

A decayed accumulation operator DA[.] takes as input a mass loading field $m(x)$ expressed at each grid location as $m(i, j)$ that is assumed to move with the flow field but is subject to first order decay in moving from cell to cell. The output is the accumulated mass at each location $DA(x)$. The accumulation of $m$ at each grid cell can be numerically evaluated

$$DA[m(x)] = DA(i, j) = m(i, j)\Delta^2 + \sum_{k \text{ contributing neighbors}} p_k d(i_k, j_k) DA(i_k, j_k)$$

Here $d(x) = d(i, j)$ is a decay multiplier giving the fractional (first order) reduction in mass in moving from grid cell $x$ to the next downslope cell. If travel (or residence) times $t(x)$ associated with flow between cells are available $d(x)$ may be evaluated as $\exp(-\lambda t(x))$ where $\lambda$ is a first order decay parameter.

Useful for a tracking contaminant or compound subject to decay or attenuation
### Capabilities Summary

#### Capability to run larger problems

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Processors used</th>
<th>Grid size</th>
<th>Theoretical limit</th>
<th>Largest run</th>
<th>RAM limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>TauDEM 4</td>
<td>1</td>
<td>0.22 GB</td>
<td>0.22 GB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sept 2009</td>
<td>Partial implementation</td>
<td>8</td>
<td>4 GB</td>
<td>1.6 GB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>June 2010</td>
<td>TauDEM 5</td>
<td>8</td>
<td>4 GB</td>
<td>4 GB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sept 2010</td>
<td>Multifile on 48 GB RAM PC</td>
<td>4</td>
<td>Hardware limits</td>
<td>6 GB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sept 2010</td>
<td>Multifile on cluster with 128 GB RAM</td>
<td>128</td>
<td>Hardware limits</td>
<td>11 GB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Single file size limit 4GB

At 10 m grid cell size
Improved runtime efficiency

Parallel Pit Remove timing for NEDB test dataset (14849 x 27174 cells ≈ 1.6 GB).

For 8 processor PC:
- Dual quad-core Xeon E5405 2.0GHz PC with 16GB RAM

For 128 processor cluster:
- 16 diskless Dell SC1435 compute nodes, each with 2.0GHz dual quad-core AMD Opteron 2350 processors with 8GB RAM

- $T \sim n^{-0.03}$
- $C \sim n^{-0.56}$
Improved runtime efficiency

Parallel D-Infinity Contributing Area Timing for Boise River dataset (24856 x 24000 cells ~ 2.4 GB)

8 processor PC
Dual quad-core Xeon E5405 2.0GHz PC with 16GB RAM

128 processor cluster
16 diskless Dell SC1435 compute nodes, each with 2.0GHz dual quad-core AMD Opteron 2350 processors with 8GB RAM

8 processor PC
Dual quad-core Xeon E5405 2.0GHz PC with 16GB RAM

128 processor cluster
16 diskless Dell SC1435 compute nodes, each with 2.0GHz dual quad-core AMD Opteron 2350 processors with 8GB RAM
Scaling of run times to large grids

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size (GB)</th>
<th>Hardware</th>
<th>Number of Processors</th>
<th>PitRemove (run time seconds)</th>
<th>D8FlowDir (run time seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Compute</td>
<td>Total</td>
</tr>
<tr>
<td>GSL100</td>
<td>0.12</td>
<td>Owl (PC)</td>
<td>8</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>GSL100</td>
<td>0.12</td>
<td>Rex (Cluster)</td>
<td>8</td>
<td>28</td>
<td>360</td>
</tr>
<tr>
<td>GSL100</td>
<td>0.12</td>
<td>Rex (Cluster)</td>
<td>64</td>
<td>10</td>
<td>256</td>
</tr>
<tr>
<td>GSL100</td>
<td>0.12</td>
<td>Mac</td>
<td>8</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>YellowStone</td>
<td>2.14</td>
<td>Owl (PC)</td>
<td>8</td>
<td>529</td>
<td>681</td>
</tr>
<tr>
<td>YellowStone</td>
<td>2.14</td>
<td>Rex (Cluster)</td>
<td>64</td>
<td>140</td>
<td>3759</td>
</tr>
<tr>
<td>Boise River</td>
<td>4</td>
<td>Owl (PC)</td>
<td>8</td>
<td>4818</td>
<td>6225</td>
</tr>
<tr>
<td>Boise River</td>
<td>4</td>
<td>Virtual (PC)</td>
<td>4</td>
<td>1502</td>
<td>2120</td>
</tr>
<tr>
<td>Bear/Jordan/Weber</td>
<td>6</td>
<td>Virtual (PC)</td>
<td>4</td>
<td>4780</td>
<td>5695</td>
</tr>
<tr>
<td>Chesapeake</td>
<td>11.3</td>
<td>Rex (Cluster)</td>
<td>64</td>
<td>702</td>
<td>24045</td>
</tr>
</tbody>
</table>

1. Owl is an 8 core PC (Dual quad-core Xeon E5405 2.0GHz) with 16GB RAM
2. Rex is a 128 core cluster of 16 diskless Dell SC1435 compute nodes, each with 2.0GHz dual quad-core AMD Opteron 2350 processors with 8GB RAM
3. Virtual is a virtual PC resourced with 48 GB RAM and 4 Intel Xeon E5450 3 GHz processors
4. Mac is an 8 core (Dual quad-core Intel Xeon E5620 2.26 GHz) with 16GB RAM
Scaling of run times to large grids

1. Owl is an 8 core PC (Dual quad-core Xeon E5405 2.0GHz) with 16GB RAM
2. Rex is a 128 core cluster of 16 diskless Dell SC1435 compute nodes, each with 2.0GHz dual quad-core AMD Opteron 2350 processors with 8GB RAM
3. Virtual is a virtual PC resourced with 48 GB RAM and 4 Intel Xeon E5450 3 GHz processors
Summary and Conclusions

- Parallelization speeds up processing and partitioned processing reduces size limitations
- Parallel logic developed for general recursive flow accumulation methodology (flow algebra)
- Documented ArcGIS Toolbox Graphical User Interface
- 32 and 64 bit versions (but 32 bit version limited by inherent 32 bit operating system memory limitations)
- PC, Mac and Linux/Unix capability
- Capability to process large grids efficiently increased from 0.22 GB upper limit pre-project to where < 4GB grids can be processed in the ArcGIS Toolbox version on a PC within a day and up to 11 GB has been processed on a distributed cluster (a 50 fold size increase)
Limitations and Dependencies

- Uses MPICH2 library from Argonne National Laboratory
  http://www.mcs.anl.gov/research/projects/mpich2/

- TIFF (GeoTIFF) 4 GB file size (for single file version)

- Capability to use multiple files to cover domain not yet on web site

- Processor memory

- Toolbox user interface only for ArcGIS 9.3.1
Are there any questions?