Workshop: Development of a HydroChemical Database, StreamChemDB , 13-14 Oct 2011, Corvallis OR

Summary

Purpose of this workshop:
- discuss stream chemistry data base structure, metadata, controlled vocabulary
- compare this with what is being used by other groups and data networks
- brainstorm about incentives to encourage site involvement (visualizations?)
- articulate questions and tools (flux calculations?) that this database could help address
- identify sources and draft proposals for funding for next steps

and engage in general and specific scheming about science syntheses.

Thanks everyone who was able to attend the Workshop last month. It was great discussion and interesting to see multiple ways of making data and metadata available across networks of sites and linking between networks. Detailed notes are below.

We talked about follow up workshops and recently LTER has announced its call for workshops. New this year is option of short term funding (~6-8 months) for postdoc salary. Because we have a postdoc familiar with the project, data compiled already, can discuss using webinars and meet at the ASM in Sept, we felt that going for additional funding for a postdoc to focus on analyses would be best use of time and effort rather than all of us travelling and meeting for a few days. LTER plans to only award 1-2 of these, so competition will be stiff but seems worth a try. We are also submitting a small supplemental proposal for cross site Experimental Forest funding.

Link to Presentations: We posted the presentations as pdfs at: http://www.fsl.orst.edu/efr/news.htm or google 'Stream chemistry synthesis project'. The document site is set up with logins so that only workshop participants can access the files. If the presenters want to edit their file, feel free. Note that several posters from this project that are being presented at AGU next week are here also.

username: EFRsynthesis
password: 4StreamChem

Followup and next steps:

- **LTER Short term PostDoc proposal -** Alba is submitting a proposal to examine fluxes for a conservative (Ca) and non-conservative (NO3-N) solute across sites. Questions include: How do methods of calculating loads compare across sites with very different hydrological regimes and different relationships between discharge and concentration? How do influences compare across a conservative and a non conservative analyte? Is there a single method that is appropriate for a broad range of sites? What are minimum frequencies of sampling that provide reasonable approximation of annual load across hydrologic regimes? She would be using StreamChem data base and HydroDB. If she gets funded and you want to participate and your data isn't in this database yet, let us know. We are working with a programmer to create a straightforward interface so that additional sites could easily upload their data and metadata to the database.

-Plan for **Workshop(s) for LTER All Scientists Meeting** in Sept 2012. One option would be to discuss and refine questions involving cross site syntheses – such as stream chemistry fluxes and looking at analyses that from comparison of fluxes. Other workshop could involve followup on databases, harvesters and linking data.

-**Posters at AGU** next week so come by if you are attending. **'Temporal trends in stream N and responses to disturbance'** poster is Monday afternoon 5th **(B13G-0648).** I'll be there. **'Variability in stream N and relevance to nutrient criteria'** poster is Thursday afternoon 8th **(B43D-0313)**. Stephen will be there. You can see these posters on the link above.

-**Manuscript** examining long term trends in nitrogen dynamics at 7 Experimental Forest sites is in final round of reviews by co-authors. This is the first manuscript from this cross site stream chemistry collaboration, and that led to this database creation. Alba is lead author and plans to submit to *Frontiers in Ecology* in next few weeks.

-Don submitted a proposal for **Production Workshop to LNO** to link SiteDB with hydrological data, climatic data, and chemistry data and to help connect the existing data harvesters with other networks, including CUAHSI HIS.

-Theresa is working with **LTERmaps group** to integrate crucial info (lat long, land use, disturbance) to Site DB that will help provide standard metadata across projects and info needed by HIS and other networks.

-**StreamChemDB** prototype download site and web pages are moving and being standardized. A programmer will be working with us for a few months to create interfaces and visualization and he will help automate the flux calculations using the database.

-**Continued search for funding** to 1) finalize controlled (preferred) vocabulary and standardize output (see notes) to facilitate communication across projects, 2) work with LNO on database interface with NIS, 3) add additional data, sites and metadata. Ideas?

Participants in workshop:

Emilio Mayorga, Univ of Washington, CZO and NANOOS;
Mark Williams, Univ of Colorado-Boulder-Niwot LTER and CZO;
Nate Booth-USGS, Madison - WQX and BioData;
Jeff Horsburgh, Utah State Univ, CUAHSI and Hydrologic information System;
Jay Jones, Univ. of Fairbanks, AK, Bonanza LTER;
Jody Potter, UNH, LUQ EF and LTER;
Yang Xia, LNO, ClimDB/HydroDB;
Stephen Sebestyen, USFS NRS, Marcel EF;
Alba Argerich, OSU, Postdoc on StreamChem project;
Effie Greathouse, OSU, Data coordinator, StreamChemDB;
Sherri Johnson, USFS PNW, Andrews EF and LTER;
Don Henshaw, USFS PNW, ClimDB/HydroDB and Andrews LTER;
Keli Goodman-NEON;
Doug Ryan, USFS PNW;
Suzanne Remillard, Andrews LTER;
Theresa Valentine, Andrews LTER;
Remotely through VTC- John Campbell, USFS NRS, Hubbard Brook EF and LTER;

Additional attendees for presentations: Remotely through VTC- Chuck Rhoades, USFS RMS, Fraser EF; Laurie Porth, USFS RMS, data coordinator; Ann Mebane, USFS WO; James Brunt, LNO; Kerstin Lehnert, EarthChemDB, Columbia Univ; Deb Hayes, USFS WO;
In Corvallis - Adam Kennedy, Linda Ashkenas, and Julia Jones, Andrews LTER

Agenda

13-Oct Forestry Sciences Lab, 3200 SW Jefferson, Corvallis, OR

8:30	Introductions Objectives for meeting	Sherri Johnson
9:00	Presentations	
	StreamChemDB	Effie Greathouse
	EarthChemDB	Kristin Lehnert
	CUAHSI HIS	Jeff Horsburgh
	WQX, Earth3, BioData	Nate Booth
	CZO	Mark Williams
	CLIM/HYDRODB_Site DB	Yang Xia and Don Henshaw
11:00-12:00	Discussion:	
	Integrating across databases, agencies, objectives	
	data, spatial attributes and metadata	
	applications of data- audiences	
1:00-5:00	Discussion:	
	Ways of highlighting the value of data -publications, web products	
	metadata - theory and levels	
	Units and time steps	
	Controlled vocabulary	
	Precision, accuracy	
	Questions and tools that a streamChem DB could help address?	

14-Oct

8:30- 12:00 Overview:

Brainstorm about incentives to encourage participation
visualizations
spatial data, temporal data, multidimensional stream chemistry
Discussion:
identify sources and draft proposals for funding for next steps
data management design
integrating timesteps, precision, conc, extrapolation into fluxes
database access and permissions

1:00-4:00 Discussion:
Database design
controlled vocabulary
Next steps - Proposal drafting

Presentations are available through: http://www.fsl.orst.edu/efr/news.htm or http://web.fsl.orst.edu/streamchem/ username: EFRsynthesis; password: 4StreamChem and click on 'available documents' on left bar.

Notes, Questions and Discussion

StreamChemDB- Effie Greathouse, Sherri Johnson

Notes: Effie discussed progress on creating cross site stream chemistry database and challenges with how sites describe units, precision, detection limits, methodology. Prototype at: http://oregonstate.edu/~greathoe/StreamchemDB.html; User name: chemdb; Password: lter.
QUESTIONS: Suzanne-Controlled vocabulary? Have you been collaborating with the LTER office with their effort of constructing a common vocabulary; Don-Reported unit vs. what is measured This is what they are doing at HJA, other LTERs; Nate-USGS-we keep units and speciation together; Jeff-HIS separate units, speciation, molecular compounds; Doug-which are the products related to the StreamChemDB?; Sherri-Trends, Disturbances, Fluxes, Nutrient Criteria.

CUAHSI Hydrologic System – Jeff Horsburgh

Notes: Point AND spatial data, Dynamic Controlled Vocabulary Moderation System
Interoperability between databases: WaterML-"language to communicate" Created by CUAHSI >100 universities. Loading data: They are able to load data that is coming in streaming automatically - e.g. every day upload data to the CUAHSI Hydrologic System
HydroDesktop: Visualization-search in space or time, thematic keyword search, databases/agency --> download the data in your hard disk or directly to R.
QUESTIONS: Sherri- How much is USFS involved? Jeff-Not much, they hope to include the data from StreamChemDB; Don- we expect to have this connection through the Hydro/ClimDB?; Jeff- You don't need to change your database, we need a change in the way to communicate with the data; Stephen- Is there space to host the data -servers?; Jeff- You can use SanDiego supercomputer, or a friend. In the future CUAHSI may evolve to be a data storage center.
To participate, Clim/hydroDB data needs to be in HIS format (can be link to a web service) with lat longs
James Brunt: Late 2012 CUOHSI web service for LTER (current development plan)

ClimDB/HydroDB- Don Henshaw, Yang Xia, Suzanne Remillard, James Brunt

Notes: Visualization 2 sites or more at the same time, SiteDB to feed Clim, Hydro, and StreamChem DB + LTER maps, PASTA-LTER network information system, EML
QUESTIONS: Keli-QA/QC is it done at the site? Don- they plan to do another workshop at the end of the month at HB

Digital Data Systems for Geochemistry-Kerstin Lehnert
Notes: Integrated Earth Data Applications IEDA, EarthChem
http://www.petdb.org/search
http://www.earthchem.org/
http://www.geoinfogeochem.org/
Sampling sites for the data-located through Google Maps
QUESTIONS: Stephen- Repository or harvesting system? Both;
Sherri-: Funding? Kerstin-Base funding from NSF ks...

WQX-USGS –NAWQA- Nate Booth
Notes: NAWQA -data warehouse, ACWI-Network of Reference Watersheds, US National Groundwater Monitoring Network, Next: Water Quality Portal -EPA, USGS together,Different networks:
EPA= clearinghouse more than creating their own data
Obj: evolving WaterML as a standard language for all hydrologic data
First trial: US Groundwater Data Portal http://cida.usgs.gov/gw_data_portal/
Next steps: Surface water interoperatibility Experiment, Water Quality Portal to be launched during the NWQMC, Also BioData, GeoData Portal http://cida.usgs.gov/climate/gdp/
QUESTIONS: Don: Storet less and less new data stored. Why? Nate- Less resources;
Mark Willliams: How to come to a common vocabulary among databases? Doug: Biossessment protocols which ones do you use? National level protocols; Sherri: How do you deal with metadata? Storet does not have much metadata; Jeff: CUAHSI does not QA/QC data, it´s the provider. who QA the data? New data network funded by NSF the use another method: New datasets even if they are an expansion of the old ones get a new identifier so the people that has used the old dataset they can cite it and it is still there; Nate: Diferrent versions related so when new data is added they can send an alert to the users. Stephen: Do you request corrections, do you reject the data if they do not pass your QI/QC?

Boulder Creek Critical Zone Observatory - Mark Wlliams
Notes: Metada: data+metadata together, Date difficult to have correct (day savings time?), Preferred vocabulary instead of Controlled vocabulary, Common database for all CZOs
QUESTIONS: When do you will know if WaterML standard work for water chemistry data? Emilio: ODM 2.0, Future versions of WaterML will work

Incentives to encourage participation?
Gives the sites back a “database” product that they can use, that makes the data more usable to them.
Tools: data management, analysis, visualization
Make it easier for sites to: explore their data, simple/easy to use/QAQC/simple graphs
Why not?: Protection of data before publication, several layers of release: Guarantee: sites have control, what is it that everyone can agree with. Download forms can be disadvantage to others.
Suggested acceptance form: more info/NADP: register once, cookie until you clear your cookies. Need to track downloads: usage can be tracked from IP addresses, tools available. Earth Cube: wider access/machine readability suffers, can set up IP address from server:

Citation with DOI, can have tiered layers of access. Use tracking of downloads to send targeted e-mails of errors, can automate this. USGS: option for higher level graphics, persistent identifiers: citations can then be tracked when people use DOI's.
Centralized data and sites; CUAHSI: people put money on the "curation" of the data
Answer scientific questions.
NSF asks for data management as a part of all their grants.
Begin with having some commonalities-common metadata. Repositories.
Jay: Why do not include lakes? Sherri: We will overlap with Gleon project
Doug: Unfunded mandate? Helps to find money (end of year) to give seed money to sites , as in HydroDB

Visualizations?
Calculation of fluxes (John-Update QUEST: In the future they plan to use HydroDB and StreamChemDB)
Calculation of different time steps -annual data, monthly data, daily…
Emilio: notify about updates or new tools: automatized download from USGS
PROVIDING TOOLS To make things more useful to the sites themselves, error checking (alternatively, track the users, you just sign in once and after that the system remembers you), label each database with a DOI.
Emilio: when you download the data there is a link to a webpage where there are notifications about your dataset.
Functions could include different levels of permission for data releases. Censored data, how to represent them in the downloads or plots?

NANOOS – Emilio shared example: http://www.nanoos.org/nvs/nvs.php
Good examples for filters in a geographic framework
Example: Tillamook, OR, Provider: USGS, Data source: CUAHSI-HIS (where we are pulling the data from)
See Christina River Basin
Cost ~ $100k

Who would be users of StreamChemDB?
Scientists -fluxes
Regulators-statistical tools, e.g. probability of exceedance, frequency calculations, return periods
Synthesis of results, reviews
Educators K-12: NSF apply for NSF educational programs K-12 (they are working in modifying HydroDesktop to implement it in the classrooms)
General public,
Tools for specific audiences

Controlled vocabulary:
Shared vocabulary instead of controlled vocabulary;
LTER has controlled vocabulary that should be reviewed
CUASHI example: extensive Controlled vocabulary:
Variable : code and name (code can have meeting) HIS.central.cuashi.org: search hydrodesktop
CUAHSI: http://his.cuahsi.org/mastercvreg/cv11.aspx
Hydrologic concept ontology : http://hiscentral.cuahsi.org/startree.aspx -star tree-VISUALIZATION!
San Diego SC Center wrote/provides web services to do this.
Next steps for vocab - Agree on exchange file and make it machine and human readable. Metadata and data included: Give options/guidance to sites so they can enhance

EcoTrends is not a good model.
DataOne https://www.dataone.org/
QUEST-composite method; It needs better than daily measurements of Q
15 minute instantaneous measurements; LOADEST-daily Q
Hydrologic regimes will determine sampling freq => fluxes calculations

Which is the best way to integrate across agencies?
Jeff: 1st share an information model what are the elements of metadata, data that everybody understands. Once agreed, you can encode it in different ways.
Stephen: Which data do we want to share? Long-term data from LTERs or also short-term data/intensive data.
Synthesis studies
Theresa-There is a lot of money invested in collecting data but maybe they do not have enough resources to maintain them or make them available-Data repositories
LTER+EFR high valuable data, not a lot of long-term data, and less in reference sites
Doug: Why? We need to demonstrate what will be different after doing this effort? Is it worth it? -economically.
Mark: Example of difficulty: Fluxes calculations, each site have different sampling frequency….
How to integrate databases? Hands-on.
Emilio: First integrate LTER+EFR using common good practices, and then to the rest of networks. If LNO is overwhelmed, CUAHSI can help to make the "connection" to the rest of networks s they do with CZO.
Jeff: Shared Vocabulary CZO is the same than the one used by CUAHSI. Difference: ways to add new vocabulary -CZO open forum, CUAHSI only 3 persons deciding.
Information models

Detection limits
CUAHSI: they do not have MDL but they include censored data
USGS: Null value if below detection, another field for MDL
StreamChemDB: Sent value from the lab, another field for MDL/IDL/others...
CZO: MDL explained in methods but not stored with the data, if below detection=0
EPA: 1/2 MDL
Emilio: How to store the data is different than how the user recovers the data
Jeff: You want to include also qualifiers ("result questionable", they do not categorize the qualifiers) and annotations (notes) for each data value.
Two levels to display to the user: QA/QI and all the data -INTERNALLY : machine read values + another column for MDL; Derived data vs Row data
Also include "Who collected the data" and "Data quality" -it implies a second QA/QI -Important for data harvesters such as USGS

Missing Data?
Jay: How does StreamChemDB differentiate between missing values and below detection? Effie: The database doesn't have missing values-if there is no sample there is no date.
Possible problem: if you present more than one analyte
Jeff: 99999 for a missing value
Jay: -99999

Units
How to maintain the precision when doing unit conversions?

USGS do not make conversions to avoid the problem. Unit and speciation separate (one column with mg/L and another one with NO3 or NO3-N).
Database diff from Information system that delivers data.
Issues: calculations of fluxes, €of different sampling freq, →different numbers
When comparing cross site, different levels of PRECISION matter.
Hopefully changes in methodology are already corrected in the databases.
Precision, do you accept read machine values with 9 decimal places? Does it have any meaning?

What is missing in future StreamChemDB?
Other constituents: Sediments, particulates, Isotopes, Temperatures,Toxics and pathogens (urban systems), Tissue-heavy metals, SUVA

Intent of the data: useful to help the interpretation of the data
CUAHSI: 1 sample can have several analytical methods but 1 sampling method

Important information: Analytical methods (instruments, labs, name of the analyzer, ...), Sample collection methods, Sample preparation -filtering, storage, Sample type: grab, autosampler, composite...
Hierarchy of the sample (core and subcores after a long time)

Important to be able to add new analytical methods in StreamChemDB. NEMI? People should be able to look at NEMI and decide by themselves which method they are using. People should be submitting new methods to NEMI. CUAHSI does not use NEMI methods -they are open.
https://www.nemi.gov/apex/f?p=237:1:3935898730158217

ODM captures space and time data
Time step: if the value is an average which is the time between the first measure and the last
Time spacing: how often do you have an average

Space: Surface, 4 cm below the surface, 4 cm above the streambed…
Space 2: data linked to an spatial model-this sampling point is located here (x, y), but may represent this reach, this stream, this watershed… Synoptic data.
Spatial resolution –LIDAR
Disturbances: How do you define disturbances? Date of the treatments? Pulse vs. press disturbances?
Could include a disturbance layer in the geographic system, Site DB

How to cite and track citations? DOI: CUAHSI information model is published

Discussion:
Give options/guidance to sites so they can enhance
Decide on format/metadata/template;
Have a defined coordinator;
Partner non LTER FS sites with LTER FS sites;
Connect with other efforts: CUASHI, CZO, NEON,
Combine Hydro/Climb/Chem (LNO scheduled to do facelift on Hydro/climDB 2013 from Yang)
Encourage LNO to connect with other efforts, propose a production workshop

Get feedback from users during the process (both data providers and researchers)
Need to support a full time person

Push or pull systems (do not follow Ecotrends example), Pull system takes funds;
Suggest a Flux calculator;
Followup -ASM workshop session next year, get LTER Working group formed, another workshop

NEXT STEPS:

1. **LNO Production workshop -** 3-4 days Don leader-proposal due by mid-November
   Idea: SiteDB click on 1 site, useful for hydrological data, chemistry data, climatic data…
   Machine to machine
   How to upload data -can be desktop based or web based
   web services: machine to machine
   machine to people (people=client)
   Censored data not included in most of the plots CUAHSI –visualization

2. **Geographic framework –**Theresa V heading to LTERMaps workshop Nov.
   What would we like to appear in the different layers of a map + queries?
   Gages-what is collected in each gage? Which analytes are being analyzed?
   Which sites have chemistry, gaged stations?
   SiteDB: + interactive map http://www.lternet.edu/map/

3. **Follow up workshops -next LNO call in December for cross site; ASM Sept 2012**
   CUAHSI willing to collaborate in those workshops
   Create products= stuff, data sets, pubs and highlight outcomes,address "So what" Big Picture

4. **Continue efforts on StreamChemDB**
   -Standardize web page for inputs and outputs, to make it easier for sites to input additional analytes and new sites to collaborate, programmer with EFR/ARRA project will begin to work on this winter;
   -Revise controlled vocab relative to other groups, define methods, Effie's funding has ended, need additional funding for this;
   -New synthesis products - multiple types of flux calculations for comparison in cross site paper, Alba will be applying for Post Doc funding from LNO, additional sites that want to collaborate on this could ideally add their own data to database through web interface;
   -Find additional funding to help sites add data or to bring in new sites and link to other groups;