Data expedition, Spring 2020
Faculty sponsor: **Dr. Kateri Salk** (kateri.salk@duke.edu)**,**Nicholas School
Graduate Student: **Nicholas Bruns** (neb8@duke.edu), 3rd year PhD student, river ecology
Course: EOS.323, Landscape Hydrology for undergraduates (instructor: Salk)

Spatially explicit surface water quality analysis in river networks: linking public water quality data to watersheds and network flowlines

The below document gives an overview of the module, including both its core data products and 3 course sessions. All data are given in RMarkdwon files or served through interactive web portal. Accompanying this main document are 4 items:

- *slides (sessions 2 and 3)*
- *RMarkdown files used in class sessions (session 1 and 2)*

Summary

This data expeditions module used 3, full course sessions to introduce undergraduate hydrology students with minimal programming background to:

- public water data (water quantity and chemistry)
- spatial analysis of water data
- 2 core, spatial datasets produced by the USGS that enable spatial analysis
- the programming language R
- R based tools for water data
- spatial analysis and maps in R

Motivation

Water science now has an extensive, standardized ecosystem for accessing and analyzing water data. The USGS has led this initiative through 4 actions:

1) publishing its own high quality public data,
2) joining data from other federal, state, local, tribal, and private sector entities, publishing "harmonized" public datasets casted into USGS standards
3) developing R interfaces for public data
4) developing other R tools for working with its published data

Our module aimed to introduce students to this ecosystem. We also emphasized spatial analysis because many core problems in water science require a spatial perspective. For example, managers aiming to improve water quality in a lake may aim to reduce the nutrients in the river entering that lake. These nutrients entered the waterway somewhere in the upstream watershed, for instance from an outdated septic system. Reducing nutrients entering the lake therefore, replacing that outdated septic system, which in turn requires locating its specific location.

Sessions 1 and 2 worked through RMarkdown files. Rather than using virtual machines, all students installed and developed an R and data environment on their own machine. Course sessions began with overview lecture, proceeded with walking through and running code chunks together, and finished with activities to assess comprehension. We achieved our comprehension

assessment by giving students questions that required modification and running of existing code. We found that this "template based", or "cut and paste" based approach was successful, especially on our second session, as it allowed students with minimal previous coding experience to gain exposure to the potential of programming, public data, and programming based spatial analysis. Our intended approach, however, was designed for, and indeed required, in-person circulation by the instructors to trouble-shoot installation differences. Therefore, when our third planned session occurred in the surprise remote conditions of spring 2020, we shifted from using R as planned to using a pre-built tool for interacting with water chemistry data. Specifically, we used a wonderful visualization dashboard (that actually began as an IID, Data+ project):

Ecosystem Data Visualization at the Hubbard Brooke long term ecological research station.

Our final sessions emphasized the "so what"—what would we want to do with data?

Topic list by session

Session 1 (intro day)

- Introduction to R
- Installation checks on student machines
- Spatial analysis and map making in R,
- Watershed boundaries and the "Watershed Boundary" dataset

Evaluation: Students began with a specific spatial location, were tasked to produce an interactive map of the nested watersheds it belongs to (figure 1), and then answer questions on the limitations in the watershed boundary dataset. What are assumptions it makes when describing space, and what are their strengths and weaknesses?

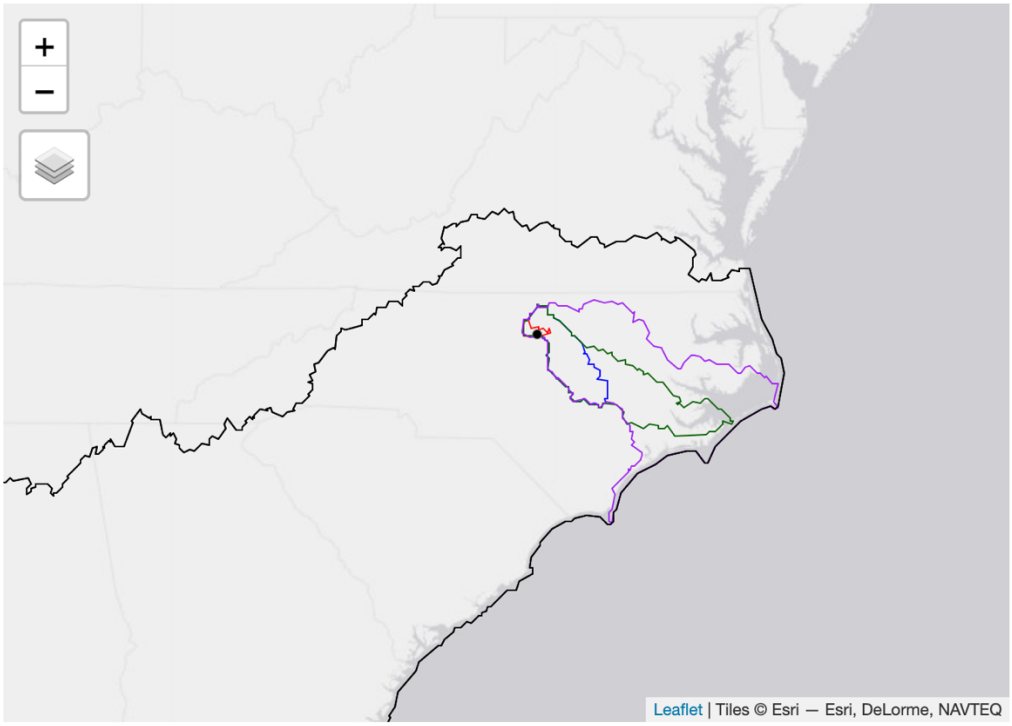

Figure 1: screenshot of assignment outcome in session 1, map of nested watershed that a specific location belongs too. Shown above is Duke campus.

Session 2 (river network day)

- River networks, and how network structure impacts time series of water quantity
- The "NHD" dataset of river networks
- Comparison of NHD network data to water quantity time series

Evaluation: students began with a water quantity time series at a location (USGS data), were tasked to visualize the upstream drainage network draining to that point (NHD data, figure 2), and then had to answer qualitative questions comparing hydrographs to their upstream river networks (NHD)

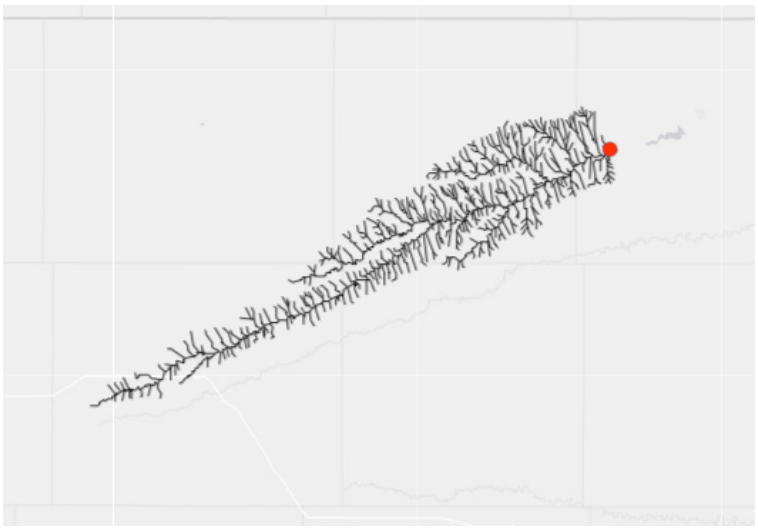

Figure 2: screenshot of assignment outcome in session 2, visualization of the river network draining to a USGS measurement gage.

Session 3 (water chemistry day):

- inferring process from river chemistry time series

Evaluation: students were tasked to make observations of features in time series produced from a large-scale ecosystem manipulation, and present to the group hypothesized processes responsible for time series features.

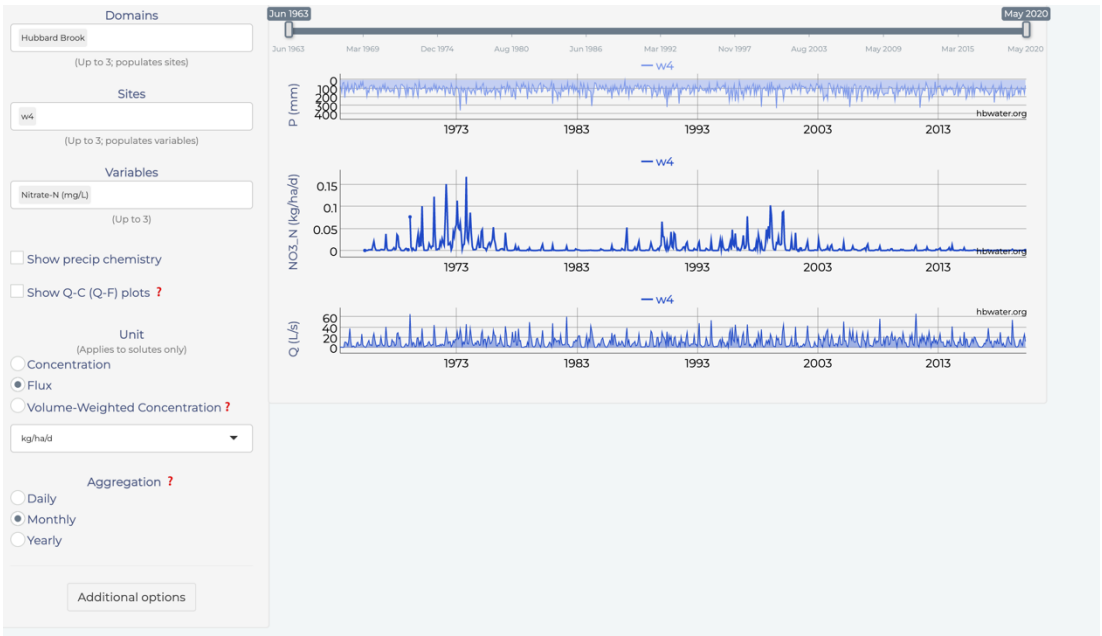

Figure 3: screenshot of dashboard interface, here showing a visualization of long-term nitrate concentrations in water leaving a small watershed. This final lesson used skills

and tools from the first two sessions to make inferences on hydrological and ecological processes.

Core datasets

- ***National Hydrography Dataset + (NHD+):*** NHD+ divides all US river reaches in .5-2 KM "NHD reaches" which have spatial attributes, including the catchment area that drains to that reach, the NHD reach upstream, and the NHD reach. Any river sample with a geo-coordinate can be linked to this data, which allows network aware analyses and network traversal (e.g., river distance between 2 points.) This dataset has 30 attributes for each reach, across 33,944 reaches.

- ***Watershed Boundary Dataset (WBD):*** The WBD is a collection of geospatial layers of the US divided into watershed polygons. Watersheds are nested, hierarchical structures, and these layers have dimensions corresponding to the layer of nesting. The finest scale degree of nesting divides the country into 101,292 polygons (e.g., Stoney Creek in Hillsborough), the coarsest, 22 (e.g. Columbia River), and we will worked with a middle sized resolution (e.g. Eno River), which divides the country into 2,321 polygons

- ***Water Quality Portal (WQP):*** a central repository and access point built by the USGS and EPA for all water quality data measured by federal, state, local, and tribal organizations. This dataset is queried by location and constituent, which in turn determine the dimensions of the returned data.