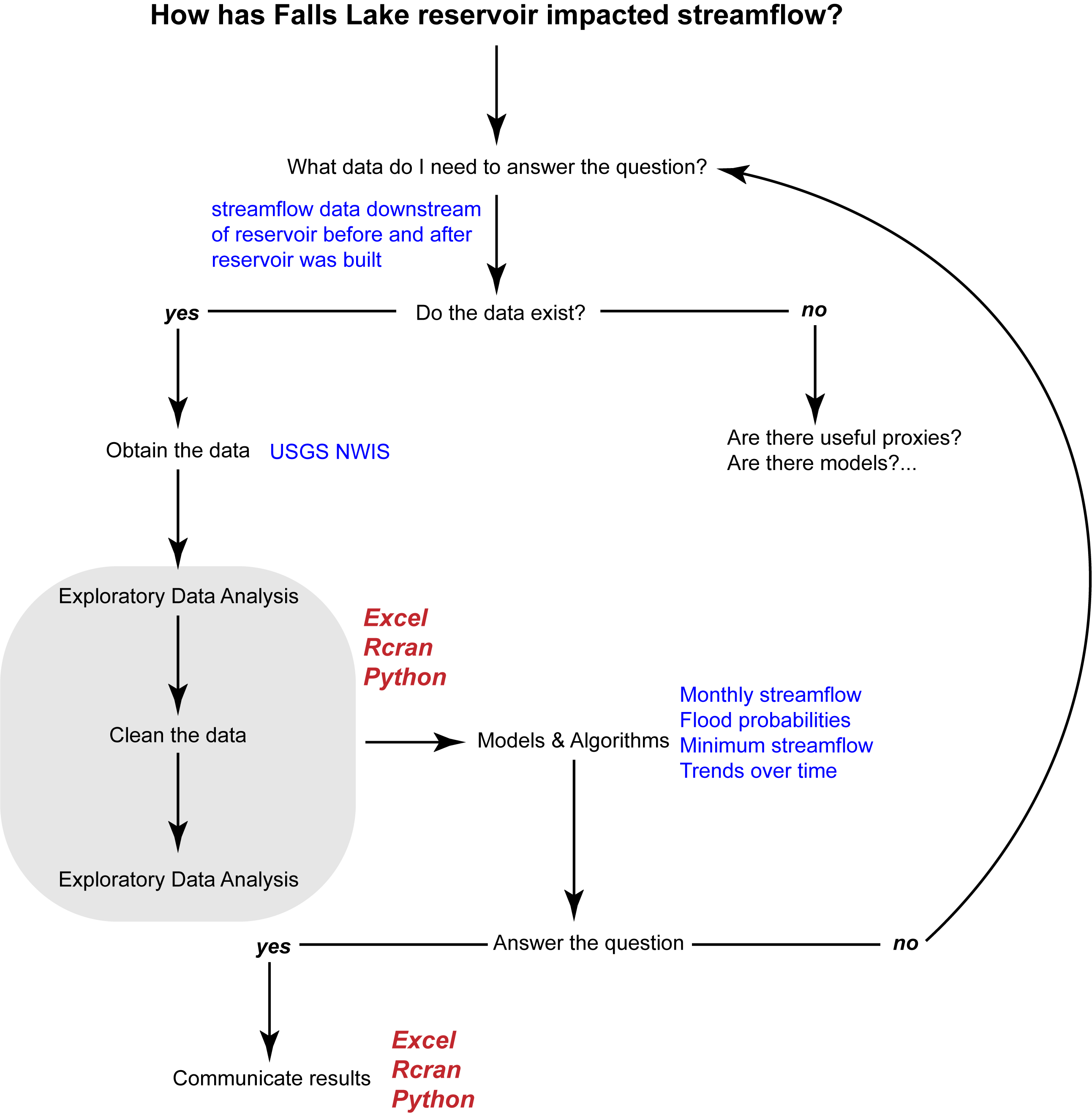
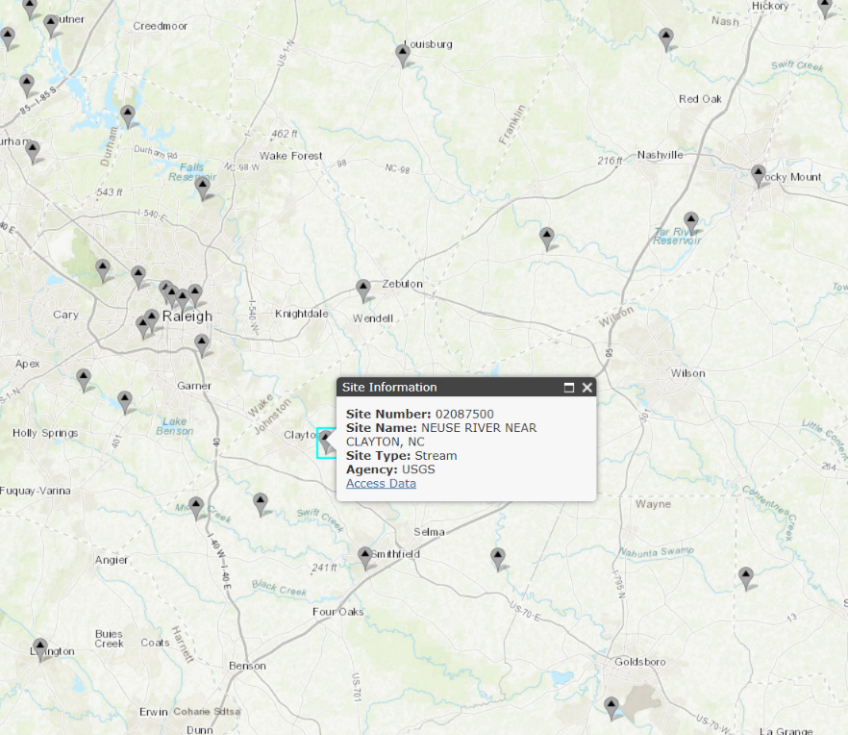
Unit 1: Streamflow Data



Excel Exercise:

# What data do I need? & Do the data exist

1. Find daily discharge data for a site downstream of Falls Lake Reservoir
   1. USGS NWIS: Mapper.
      1. Zoom into Falls Lake Reservoir and manually trace downstream. Noting the site number and site name.

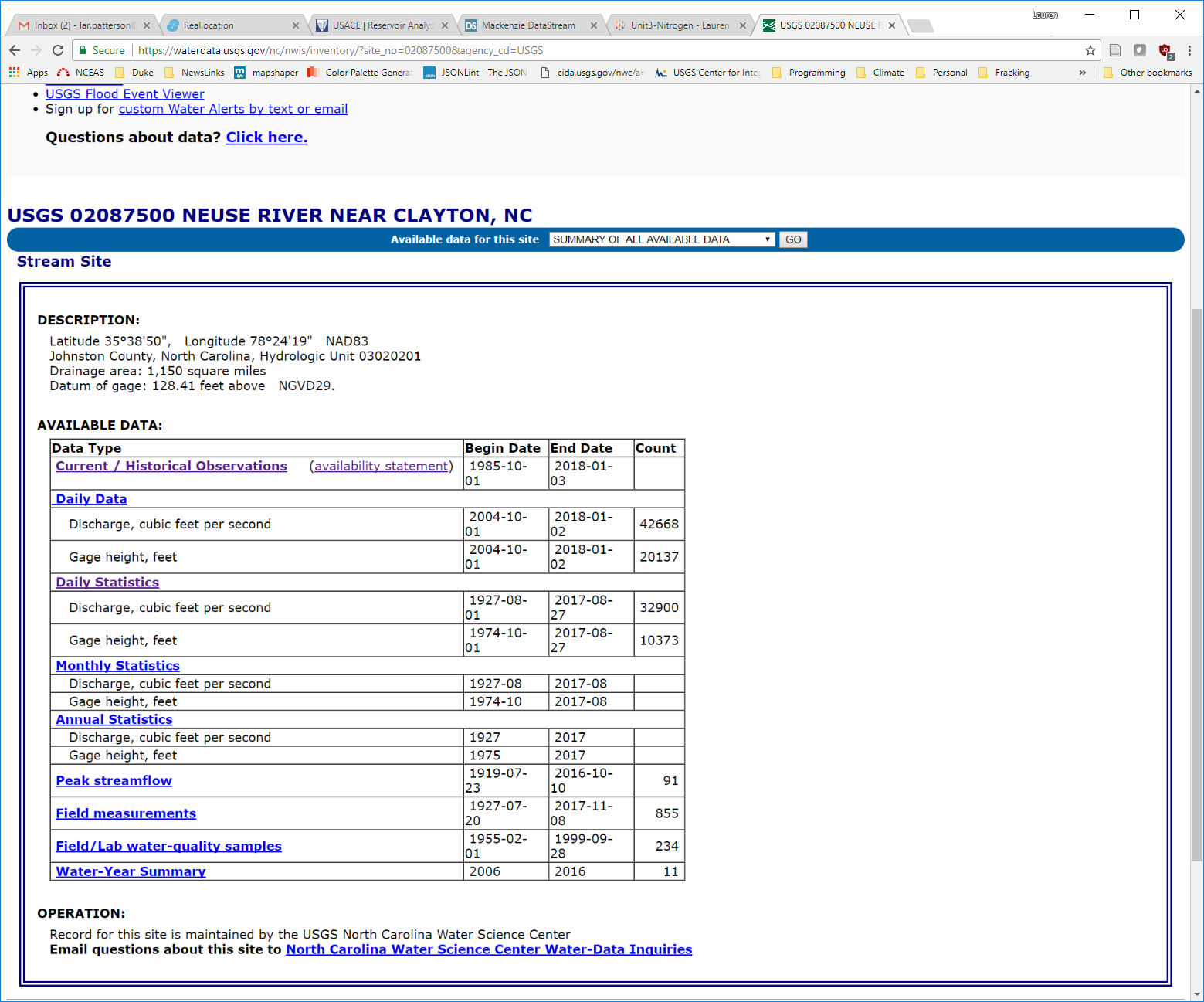


**Figure:** Snapshot of NWIS Mapper. Falls Lake is in the upper left corner. Clayton is the next site downstream of the reservoir. As are Smithfield and Goldsboro. Notice that as you move downstream additional streams join the Neuse River, complicating the signal from Falls Lake.

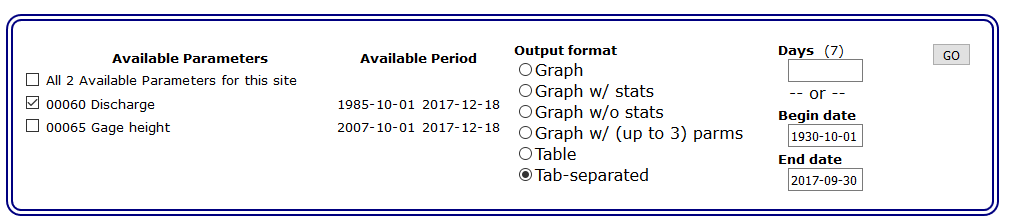
1. Find the daily discharge data for the Neuse River near Clayton, NC using one of these methods
   1. Click on the “Access Data” Button on NWIS Mapper
   2. USGS NWIS 🡪 select NC, Daily Data, Discharge, search for “Clayton” or the Site Number (02087500)
   3. Google USGS Neuse River near Clayton

# Obtain the data

1. Download daily discharge data from 10-01-1930 to 9-30-2017 (October to September is the Water Year).
   1. Notice that if you are simply looking at Current / Historical Observations the data are only listed as beginning in 1985. The definition of “Current / Historical Observations” varies between sites. Instead, select “SUMMARY OF ALL AVAILABLE DATA” and you will see daily discharge data are available back until 1927.



* 1. Select parameter of interest (Discharge)
  2. Select output format (tab-separated)
  3. Select dates

[](https://waterdata.usgs.gov/nwis/dv?cb_00060=on&format=rdb&site_no=02087500&referred_module=sw&period=&begin_date=1930-10-01&end_date=2017-10-01)

1. Notice the metadata provided
   1. Ask students about structure of the product. Metadata lines start with #, actual data are in columns with headers in line XX, followed by a line indicating data types, and then the actual data. (Another issue will be that Excel will treat the field names as integers and dates as strings…)
2. Open up excel
   1. Walk through terminology
      1. Workbook
      2. Worksheet
      3. Row
      4. Column
      5. Cell
      6. Data Types
      7. Menu options
3. Excel has a number of tricks for accomplishing the same task. Here we are going to move data from the website into excel using two methods. Our end goal is to get the desired data into columns.
4. Copy and paste *all* the data into excel
   1. To select all: Go to top left corner of website, highlight a little corner. Press CTRL+SHIFT+END or CTRL-A
   2. Paste (CTRL+V) 🡪 Notice the data are lumped into a single column. There are two ways to address this problem.
      1. Highlight the column (click on first row where data starts and press CTRL+SHIFT+Down Arrow)
         1. In menu 🡪 Data🡪Text to Columns🡪Delimited🡪Space🡪Finish
            1. Notice the data are in columns now, but the column headers don’t match the data fields until 2004 when minimum and maximum discharge were collected. Need to be careful for these types of errors using this method. I recommend the previous method.
      2. Let’s delete this version and try pasting in the data using another method (highlight column letters, Right Click, Delete)

OR

* 1. In Excel Right click, paste special, text
     1. Notice the data are in different columns
  2. Name this worksheet “Raw”
     1. Double click the worksheet tab and type in “Raw”
  3. Save spreadsheet

1. There is another way to directly import data into excel without copy and paste.
   1. Data🡪Get External Data🡪From Web🡪[Paste the URL]
   2. <https://nwis.waterdata.usgs.gov/nc/nwis/uv/?cb_00060=on&format=rdb&site_no=02087500&period=&begin_date=1930-10-01&end_date=2017-09-30>

# Exploratory Data Analysis

## Plot Streamflow Data (are there obvious gaps or outliers)

1. Copy worksheet and paste as a new sheet. Name this tab “EDA”.
   1. Right click 🡪 Move or Copy 🡪 Select Create a Copy 🡪 (move to end) 🡪 OK
   2. Remove metadata rows
      1. Highlight row numbers 🡪 Right Click 🡪 Delete
      2. Delete the second row (not data but 5s, 15s, etc.)
2. Do *not* change anything in the Raw worksheet moving forward.
3. We are interested in only the *mean* discharge.
   1. Delete all columns except site\_no, datetime, and mean discharge
   2. Change the “datetime” column name to “Date”
   3. Change the discharge column name to: “Mean Flow (cfs)”
      1. Always include the units in the column name. The unit information is located in the metadata in the “Raw” tab.
4. Create a scatter plot of discharge over time.
   1. Highlight on B1 and C1 🡪 Press CTRL+Shift+Down Arrow 🡪 In Menu Select Insert 🡪 hover over scatter plot options and select the appropriate plot given the data.
      1. Should be Straight Lines because a lot of data based on measurements
      2. Notice the x-axis needs formatting. Double click the x-axis. Notice the Bounds go from 0 to 50,000. This represents the number of days starting January 1st 1900 to December 31st 2036. You can adjust the bounds to match your date range or you can try a different chart.
   2. In the Menu 🡪 Design 🡪 Change Chart Type 🡪 Line
      1. Notice you can get a preview of what the charts will look like
   3. Make the graph pretty/meaningful
      1. Change the title to something meaningful, such as “Neuse Streamflow near Clayton, NC”
      2. Add a y-axis label
         1. Menu 🡪 Design 🡪 Add Chart Element 🡪 Axis Titles 🡪 Primary Vertical
         2. Provide a meaningful label, such as “Streamflow (cfs)”
      3. Change y-axis labels
         1. Double click on the y-axis labels
            1. Axis Options🡪Bounds 🡪 Change Maximum to 23000

Notice the minimum changed to -2000, change to 0.

Display units in thousands

* + 1. Delete gridlines
    2. You can play around in Design / Format to change colors, font-size, borders, etc.
       1. Try changing the line width – make it smaller to get more details

1. You remembered that scientists like the metric system and you need to convert the data from cubic feet per second to cubic meters per second.
   1. Add a header in column D: “Mean Flow (cms)”
   2. Find the unit conversion rate (0.028316847)
   3. Convert cfs to cms in the following row
      1. D2=C2\*conversion rate
      2. Highlight D2 and double click on the bottom right corner button
   4. Scroll down to make sure the data filled correctly. Notice you can’t see the labels for your columns anymore.
      1. CTRL + HOME
      2. Highlight Row 1, Menu 🡪 View 🡪Freeze Panes 🡪 Freeze Top Row
         1. Scroll down again.
2. Replot the graph
   1. Click on the graph 🡪 grab the highlighted blue box and drag it over to column D
   2. Change the y-axis label and bounds
   3. Reformat as needed…
3. Your general manager looks at the chart but he doesn’t like the metric system. Add a new column to convert to millions of gallons per day. Make a new plot and show side by side.
   1. 1 CFS = 0.53817 MGD

## Summarize and Plot Streamflow Data

1. Know your data. Whenever you get data it’s good practice to look at and understand the raw data to make sure it looks reasonable and you understand something about its quality and content. One of the first things to assess in the metadata is how confident you are in the quality of the data being used in your analysis. Is the data fit for purpose?
2. How confident are we in the data?
   1. We deleted the confidence markers from the original data. Let’s add those back in.
   2. Column F 🡪 add a header: “Confidence”
      1. You can copy and paste because we haven’t moved things around – but a useful trick when you are matching data together by a unique identifier is “VLOOKUP”.
         1. Vlookup(unique identifier,

highlight table you want to get data from with first column being the unique identifier,

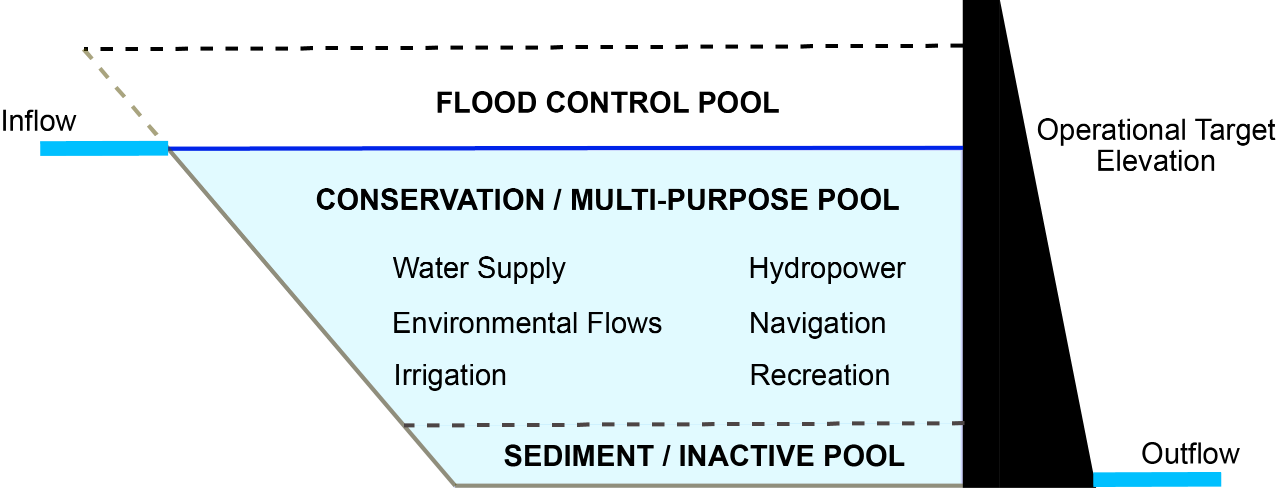
number of columns to the data you want to grab,

FALSE)

* + 1. In F2 type: =VLOOKUP(B2,Raw!$C$34:$I$31810,7,FALSE)
       1. Dollar signs lock the table in place – otherwise the range would move up or down depending on the cell.
       2. Double click the bottom corner to grab information for the entire table
  1. How many total observations are there?
     1. Below the charts, add a label (cell with text in it) and then use the COUNT() function in the adjacent cell
  2. How many, and what percentage of, observations were in each qualification code?
     1. How do you know what categories to use? Go to the raw data🡪Data🡪Filter
     2. Look at column I to see what variables are present and compare to metadata in row 28-30
        1. Look at raw data to get the qualification code
        2. Use the “COUNTIF” function
           1. =COUNTIF(“data range”, Value)
           2. Add $ and copy/paste, changing the value
  3. Now we interpret the results. How confident are we in the quality of the data at this location?

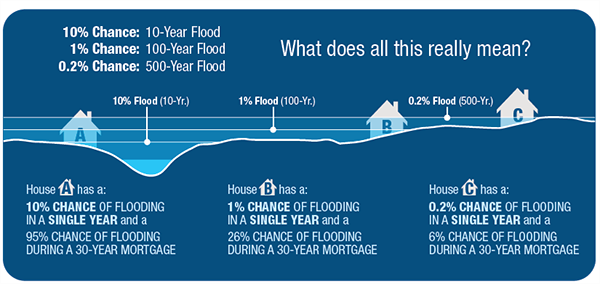
1. Summary Statistics
   1. Create the table skeleton with labels
      1. Min(range)
      2. PERCENTILE.INC(range,0.1)…0.25, 0.75, 0.90 (copy and paste with $)
         1. INC means 🡪 Inclusive, where the percentile value must be between 0 and 1.
         2. EXC means 🡪 Exclusive, where the percentile must be between 1/n and 1-1/n, where n is the number of elements in the array.
      3. Median(range)
      4. Average(range)
      5. Max(range)
   2. Did the summary statistics change with the construction and filling of Falls Lake Reservoir (1980-1983)?
      1. A quick way to do this is search for 1980 and select 12/31/1979 and up with CTRL+SHIFT+Up Arrow. Then search for 1984 and select from 1/1/1984 and down. Use $ signs and repeat the above.
      2. What do you notice? Particularly with regards to min and max streamflow. Does that make sense given what you know about reservoirs?
2. Is there seasonal variation in streamflow?
   1. Parse out the Date to include Year and Month and Water Year
      1. Insert three columns. On Column C 🡪 Right Click 🡪 Insert
      2. Give them appropriate column names
         1. =YEAR()
            1. Fix the format from date to number
         2. =MONTH()
         3. Water year runs from October to September. How can we adjust the year column to account for this information?
            1. Use If()
            2. Scroll downwards to make sure your answer worked
   2. Create Table Skeleton of average streamflow by month over the different time periods
      1. AVERAGEIF(Value Range, Value, Average Range)
   3. Plot for the period of record, prior to 1980, and after 1984
      1. What do you observe?
   4. Repeat the above using SUMIF for annual streamflow based on the water year.

# Has the reservoir impacted the 100 year flood frequency?



**Figure:** Reservoirs should moderate downstream flows. There is a flood control pool to hold flood waters that can be released slowly over time. There is also a conservation pool that holds water that can be released downstream during drier conditions to meet minimum streamflow requirements.

Flood insurance policy is built around the concept of the 100-year flood event. The housing industry has been working to explain what that risk actually means to homeowners with 30 year mortgages.



Reservoirs decrease the likelihood of downstream flooding, but that often means development occurs in areas that would have been frequently flooded prior to the reservoir. We’ve seen examples of this just his year with Hurricane Harvey.

## Understand what analyses are needed to answer the question.

1. Research how to calculate the 100 year flood return interval. We will use Leopold’s (1994) flood frequency curve and the Weibull equation to calculate the recurrence interval.
   1. Return Interval: (n+1)/m where n is the number of years of data and m is the rank of the year from largest (1) to smallest (n).
      1. \*\*\*The accuracy of a return interval is highly impacted by the length of the time series.
2. Calculate the maximum annual streamflow
   1. Pivot table
   2. Highlight your EDA table 🡪 Menu 🡪 Insert 🡪 PivotTable 🡪 New Worksheet🡪 OK
      1. Rename worksheet “Flood”
   3. Click on the Pivot Table
      1. Rows = Water Year
      2. Values = Mean Discharge (cms)
         1. Change value from sum to max (double click in the values box and go to value field settings)
   4. Copy and paste the pivot table (right click + s + v) nearby. Change column names to something meaningful. Add header: “Period of Record (1931-2017)”
3. Sort the data from max to min based on the Peak Flow (cms).
   1. Menu 🡪 Sort 🡪 Max Value 🡪 Largest to smallest
4. Add a new column and rank the flood events starting from 1 downward
   1. Enter 1, 2, 3. Then highlight cells and double click bottom right corner to finish ranking

OR

* 1. Use the RANK.EQ() function

1. Calculate the Recurrence Interval
   1. RI = (n+1)/m; where n is the number of years of data and m is the rank
      1. How many years of data do we have?
2. Calculate the Annual Exceedance Probability for each return interval
   1. Pe = 1/RI
3. Calculate the probability of the 100 and 500 year flood occurring over the next 30 years as a binomial distribution
   1. Pe = 1 – [1-(1/T)]n, where T is the return period (e.g. 100 years) and n is the number of years of interest.
4. Plot your recurrence interval on the x-axis and your max discharge on the y-axis
   1. Place the x-axis on a log scale and add minor tick marks
   2. Add a regression and find the one with the best fit (excluding polynomial and moving average).
      1. Right click on the data points and select add trendline
   3. Using the best-fit regression you just calculated – estimate the discharge for the 100-year, 500, and 1000 year events
   4. Add those events to the chart as markers.
      1. Select Data 🡪 Add 🡪 RI Year goes to X Values, Y Values are the best fit POR equation

## Calculate Return Period for 100 year Flood Using Data Prior to 1980

1. Repeat the above analysis only using data prior to 1980 to calculate the return interval
   1. How many fewer years of data are used?
2. Plot the new dataset on top of the original plot
   1. Menu 🡪 Design 🡪 Select Data 🡪 Add
   2. Add the estimated 100 and 500 year points for each plot based on the regression
3. How big is the difference between the 100 and 500 year estimates?
   1. Percent is calculated relative to the smaller record (my choice)
4. Calculate the discharge for different return periods and exceedance probabilities
5. Plot annual discharge and calculate the number of times the 100 year flood was surpassed for both the POR and prior to 1980.
   1. Plot the estimates on the same chart. How much do they differ? Are you surprised by the results?
   2. Look at the first plot you did of streamflow on the EDA spreadsheet to look at the distribution of peak events.
      1. These events are all hurricanes. How does it change your understanding of why Falls Lake doesn’t seem to impact flood frequency?
      2. What happens to your answer if you remove those three points?

# How has Falls Lake impacted minimum flows?

The passing of the Clean Water Act in 1972 and the Endangered Species Act in 1973 has resulted in many reservoirs having downstream flow requirements they need to meet for either water quality purposes or to protect downstream species. For example, at the Clayton gauge, minimum flow requirements have ranged from 184 to 404 cfs since 1983. Here we want to see if Falls Lake has raised minimum flows.

## Understand what analyses are needed to answer the question.

There are many ways to approach low flow to understand how minimum streamflow has changed since Falls Lake was constructed. We will look at a common metric known as 7Q10. 7Q10 is the lowest average discharge over a one week period with a recurrence interval of 10 years. This means there is only a 10% probability that there will be lower flows than the 7Q10 threshold in any given year. This metric can be modified to look at the lowest average discharge over a 7 month period, 7 season period, etc. To get more practice with pivot tables and if statements, we will calculate this metric using the 7 month period. To do this we need to construct a rolling average, which we can do using a series of pivot tables…

1. Create a new tab named 7Q10
2. Create a pivot table to get the average daily discharge by year and month
   1. Year and Month go into the rows. We use year instead of Water Year to ensure the data are being read in the correct order. If we use water year, the wrong September and Octobers are matched together.
   2. Mean flows is the value. Use the sum of the monthly flows (since taking lowest 7 month average, small variability in the number of days in each month is ok.
   3. Copy and paste the pivot table. What we now want to do is convert the column “Row Labels” into three columns with the date, year, and month.
      1. Add column (Year – not water year)…
         1. Create if statement
            1. =IF(Row Labels > 12, Row Labels, Above Cell)

This creates a column showing the year. Label this column “Year”

Copy and paste as values

* + 1. Relabel Row Labels as “Month” and sort the entire table.
       1. Delete month values that are years (bottom)
    2. Resort by year and month

1. Add a column to calculate 7 month average
   1. Go to the 7th cell and take the average of that streamflow and the 6 preceding months)
   2. Double click bottom corner to copy down
2. Extract the minimum value for each year
   1. Hint: use a pivot table to get the minimum rolling average per year.
3. Calculate the Rank, RI and probability
   1. Use the formulas in the flood lesson
   2. Remember to sort in the opposite direction!
4. Plot the 7Q flow with the Probability factor and flows
   1. Fit with best fitting regression
   2. Estimate the 7Q10 (this will be where 10% of flows are smaller).
   3. Add the 7Q10 point to the graph
5. How many months were below 7Q10 in the POR? Was it close to 10%?
   1. Use countif
      1. When countif need to use “ “, For example, =countif(range, “<=240”)
6. How many months per year?
   1. Add a column to your Total Monthly Streamflow Table
      1. Use if statement to pull out those months that are less than the 7Q10
      2. Use a dummy variable such as “NA” for those that are above the 7Q10.
         1. Copy and paste special
         2. Then do find and replace (CTRL + H) and leave the Replace with blank.
      3. Plot up the monthly streamflow and overlay with those below the threshold.
         1. Add a date column to Total Monthly Streamflow
         2. Use DATE(year, month, day) to create a date. Since we don’t have a day value, use 15.
      4. Plot y axis in log form, add labels and legend
   2. To find the number of Q710 events per year you can:
      1. (1) use the countif function
      2. (2) expand your pivot table to include the Below 7Q10 column and count the number of occurrences: PivotTableTools 🡪 Change Data Source…
   3. Create a column Plot – what do you observe?
      1. What percent of months were below 7Q10 prior to Falls Lake? After Falls Lake?

## Continued Practice

1. On your own – calculate the 7Q10 prior to Falls Lake and after Falls Lake

# How has streamflow changed over time? Before and after Falls Lake?

Water security is becoming increasingly important as population and water demand continue to grow. This is especially true with changing climate conditions that introduce new variability into our expectations of water supply. Briefly, we want to know whether the average annual streamflow has changed over time.

1. Create a new spreadsheet and name it “Trends”
2. Create a Table of Year, Total Streamflow, Count
   1. Copy and paste the entire Water Year column from the EDA tab.
      1. Menu 🡪 Data 🡪Remove Duplicates
   2. Use SUMIF and COUNTIF to calculate the number of observations per year and the annual streamflow.
   3. Remove those years with less than 90% of data
      1. Use IF to calculate and flag rows
3. Plot Data and add a linear trendline
4. Go to File 🡪 Options 🡪 Add-ins 🡪 Analysis Toolpack
   1. Menu 🡪 Data 🡪 Data Analysis 🡪 Regression
   2. Run the regression analysis on the data
      1. Turn on all the plots
      2. Is the trend significant?
5. Repeat for 1930-1980 and for 1984-2017
   1. What do you observe?
   2. Are any of the trends obvious?

## More Practice

1. If there are not annual trends, are there seasonal ones? What about February and August?
   1. Grab all February values
      1. Go to the working spreadsheet and filter by month
      2. Averageif the filtered data
   2. Repeat the above analysis
      1. What do you observe?

This type of analysis is very limited in excel. In fact, we cannot run the appropriate non-parametrical statistical tests in excel. Furthermore, there is often a seasonality to streamflow that may be important to pull out. It would be time consuming to copy and paste and repeat this analysis for each month in excel. This is where statistical programs and coding really become important tools to be able to use. Plus, maybe you want to know if there are changes observed at other stream gauges. For example, on the Neuse at Smithfield or Goldsboro?