# Assignment 3: Physical Properties of Rivers

*Caroline Watson*

## OVERVIEW

This exercise accompanies the lessons in Hydrologic Data Analysis on the physical properties of rivers.

### Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A03_RiversPhysical.Rmd") prior to submission.

The completed exercise is due on 18 September 2019 at 9:00 am.

### Setup

1. Verify your working directory is set to the R project file,
2. Load the tidyverse, dataRetrieval, and cowplot packages
3. Set your ggplot theme (can be theme_classic or something else)
4. Import a data frame called "MysterySiteDischarge" from USGS gage site 03431700. Upload all discharge data for the entire period of record. Rename columns 4 and 5 as "Discharge" and "Approval.Code". DO NOT LOOK UP WHERE THIS SITE IS LOCATED.
5. Build a ggplot of discharge over the entire period of record.

```
#checking working directory
getwd()
```

```
## [1] "/Users/carolinewatson/Documents/Fall 2019/Hydrologic_Data_Analysis/Assignments"
```

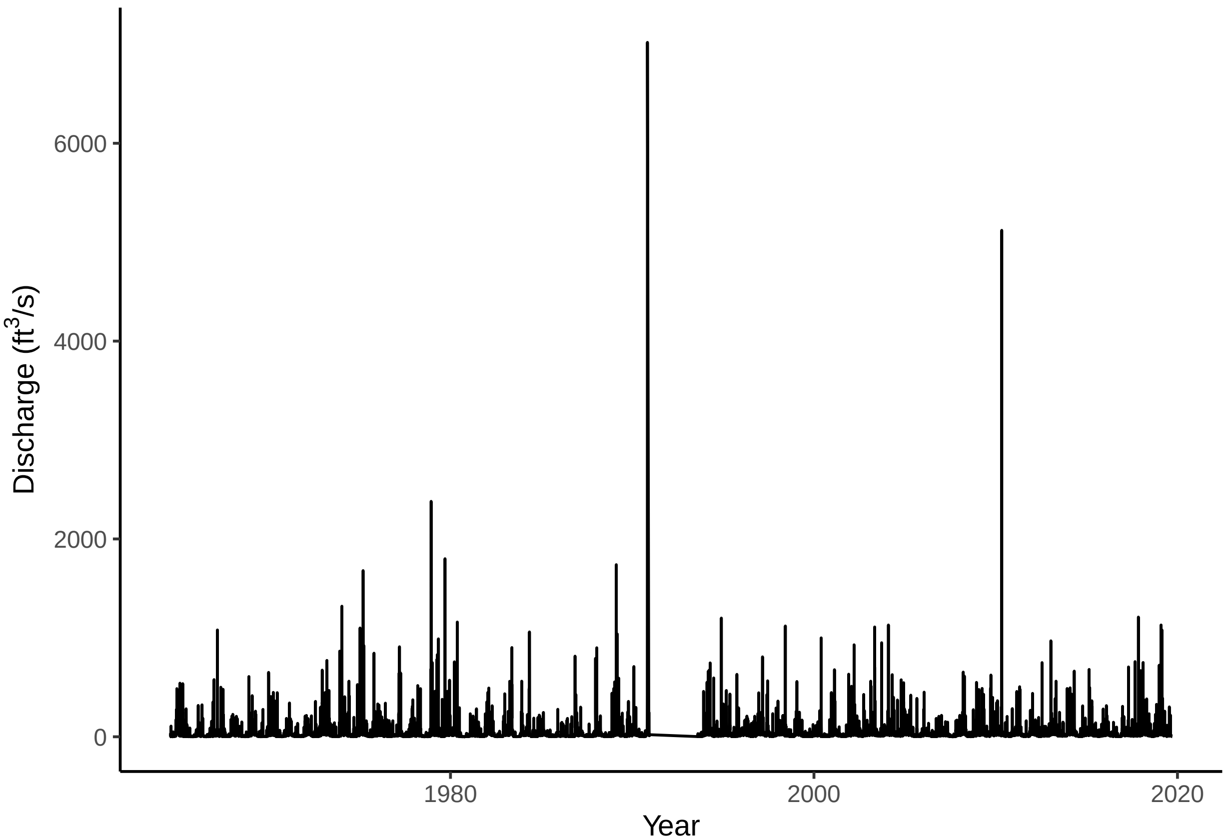
```
#loading packages
suppressMessages(library(tidyverse))
library(dataRetrieval)
library(lubridate)
library(cowplot)

#setting ggplot theme
theme_set(theme_classic())

#importing mystery site discharge dataset
MysterySiteDischarge <- readNWISdv(siteNumbers = "03431700",
                     parameterCd = "00060", # discharge (ft3/s)
                     startDate = "",
                     endDate = "")

#renaming columns 4 and 5
names(MysterySiteDischarge)[4:5] <- c("Discharge", "Approval.Code")
```

```r
#ggplot of discharge over entire sampling time
MysteryPlot <-
  ggplot(MysterySiteDischarge, aes(x = Date, y = Discharge)) +
      geom_line() +
      labs(x = "Year", y = expression("Discharge (ft"^3*"/s)"))
print(MysteryPlot)
```




## Analyze seasonal patterns in discharge

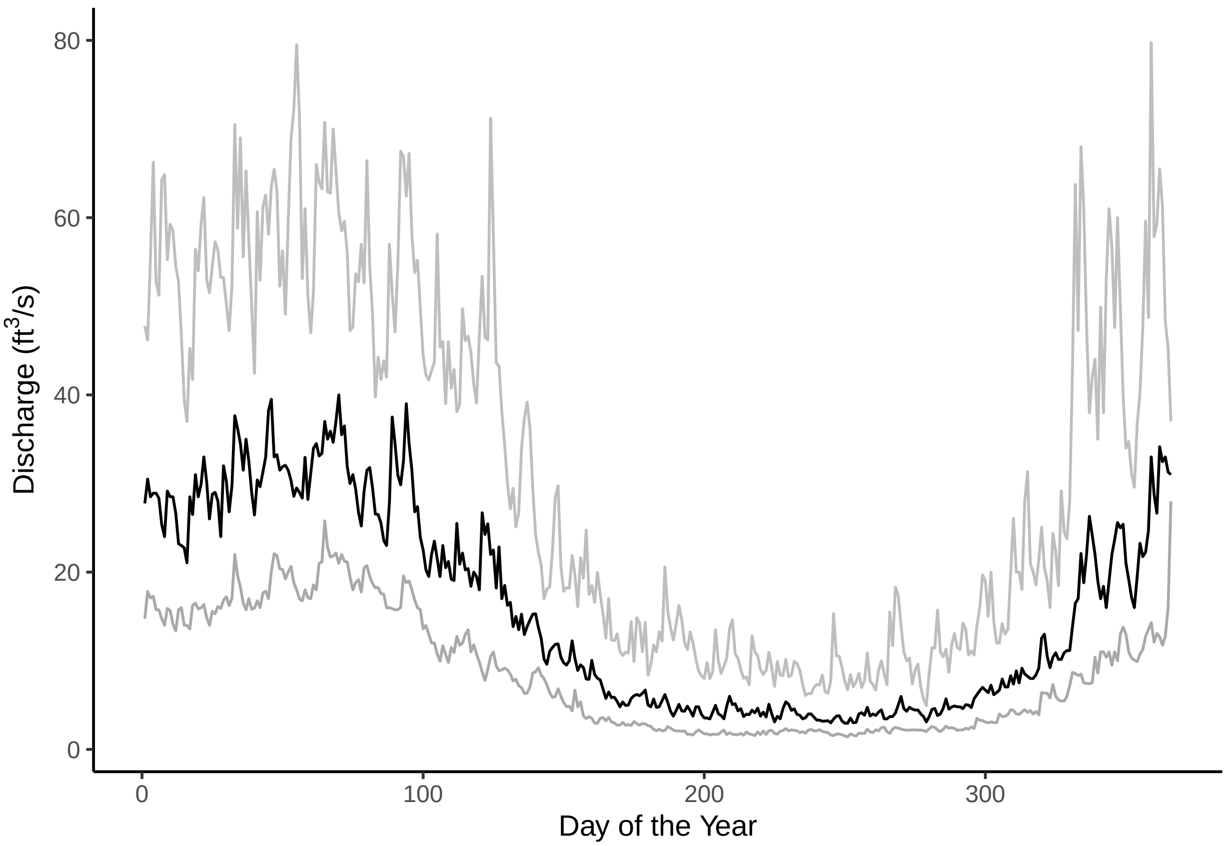
5. Add a "Year" and "Day.of.Year" column to the data frame.
6. Create a new data frame called "MysterySiteDischarge.Pattern" that has columns for Day.of.Year, median discharge for a given day of year, 75th percentile discharge for a given day of year, and 25th percentile discharge for a given day of year. Hint: the summarise function includes `quantile`, wherein you must specify `probs` as a value between 0 and 1.
7. Create a plot of median, 75th quantile, and 25th quantile discharges against day of year. Median should be black, other lines should be gray.

```r
#adding year and day of year columns to data frame
MysterySiteDischarge <- MysterySiteDischarge %>%
    mutate(Year = year(Date)) %>%
    mutate(Day.of.Year = yday(Date))

#creating new data frame
MysterySiteDischarge.Pattern <- MysterySiteDischarge %>%
  group_by(Day.of.Year) %>%
  summarise(Last.Quantile.Discharge = quantile(Discharge, probs = 0.75),
```

```r
            First.Quantile.Discharge = quantile(Discharge, probs = 0.25),
                                     MedianDischarge = median(Discharge))

#ggplot of median, 75th quantile, and 25th quantile discharges
MysterySiteDischarge.Pattern.Plot <-
  ggplot(MysterySiteDischarge.Pattern,
         aes(x = Day.of.Year)) +
  geom_line(aes(y = MedianDischarge), color = "black") +
  geom_line(aes(y = Last.Quantile.Discharge), color = "gray") +
  geom_line(aes(y = First.Quantile.Discharge), color = "dark gray") +
  labs(x = "Day of the Year", y = expression("Discharge (ft"^3*"/s)")) +
  theme(legend.position = "right")
print(MysterySiteDischarge.Pattern.Plot)
```




8. What seasonal patterns do you see? What does this tell you about precipitation patterns and climate in the watershed?

   There are clear seasonal patterns from this graph. Between the January and April (days 0 - 120), the median discharge is higher than the median discahrge from May to October (days 121 to about 300). There is another increase in discharge from November thorugh December. These same seasonal patterns are visiable in the 25th quantile and 75th quantile lines as well. From this graph of discharge by day of the year, we can see that precipitation patterns in this watershed are high in the winter and early spring months and in the late fall. This likekly means that this site is located in an aread that is not covered in snow during the winter months, so it likely is a more temperate climate area.

## Create and analyze recurrence intervals

9. Create two separate data frames for MysterySite.Annual.30yr (first 30 years of record) and MysterySite.Annual.Full (all years of record). Use a pipe to create your new data frame(s) that includes the year, the peak discharge observed in that year, a ranking of peak discharges, the recurrence interval, and the exceedende probability.

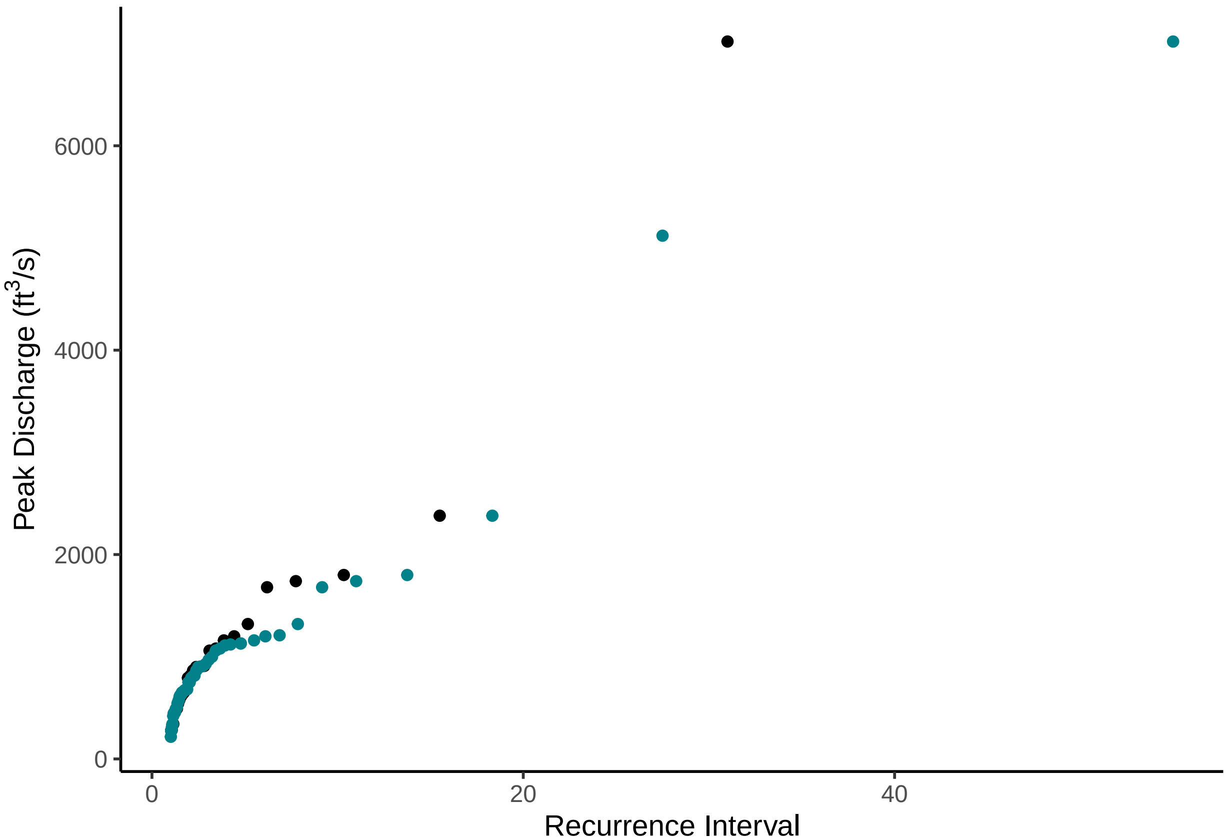
10. Create a plot that displays the discharge vs. recurrence interval relationship for the two separate data frames (one set of points includes the values computed from the first 30 years of the record and the other set of points includes the values computed for all years of the record.

11. Create a model to predict the discharge for a 100-year flood for both sets of recurrence intervals.

```r
#data frame for annual 30 year (first 30 years of record)
MysterySite.Annual.30yr <- MysterySiteDischarge %>%
  filter(Year < 1996) %>%
  group_by(Year) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
  RecurrenceInterval = (length(Year) + 1)/Rank,
  Probability = 1/RecurrenceInterval)

#data frame for annual discharge for all dates of monitoring/record
MysterySite.Annual.Full <- MysterySiteDischarge %>%
  group_by(Year) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
  RecurrenceInterval = (length(Year) + 1)/Rank,
  Probability = 1/RecurrenceInterval)

#plot of discharge vs recurrence interval for 30 year record and all data record
MysteryRecurrencePlot.Full <-
  ggplot(MysterySite.Annual.30yr, aes(x = RecurrenceInterval, y = PeakDischarge)) +
  geom_point() +
  geom_point(data = MysterySite.Annual.Full, color = "#02818a",
             aes(x = RecurrenceInterval, y = PeakDischarge)) +
  labs(x = "Recurrence Interval", y = expression("Peak Discharge (ft"^3*"/s)"))
print(MysteryRecurrencePlot.Full)
```




```r
#model to predict the discharge of a 100-year flood for both 30 years of data
Mystery.RImodel.30yr <- lm(data = MysterySite.Annual.30yr, PeakDischarge ~ log(RecurrenceInterval))
summary(Mystery.RImodel.30yr)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ log(RecurrenceInterval), data = MysterySite.Annual.30yr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -974.12 -337.65   34.84  232.57 2908.00
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -69.87     185.73  -0.376     0.71
## log(RecurrenceInterval)  1217.79     147.16   8.275 5.26e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 673.9 on 28 degrees of freedom
## Multiple R-squared:  0.7098, Adjusted R-squared:  0.6994
## F-statistic: 68.48 on 1 and 28 DF,  p-value: 5.261e-09
```

```r
Mystery.RImodel.30yr$coefficients
```

```
##             (Intercept) log(RecurrenceInterval)
##               -69.87363              1217.79015
```

```r
#Predicted discharge of a 100 year flood
Mystery.RImodel.30yr$coefficients[1] + Mystery.RImodel.30yr$coefficients[2]*log(100) #5538.257 cfs
```

```
## (Intercept)
##    5538.257
```

```r
#model to predict the discharge of a 100 year flood for the whole record of data
Mystery.RImodel.Full <- lm(data = MysterySite.Annual.Full, PeakDischarge ~ log(RecurrenceInterval))
summary(Mystery.RImodel.Full)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ log(RecurrenceInterval), data = MysterySite.Annual.Full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -955.95 -236.29   41.91  210.67 2805.35
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.001    116.322  -0.017    0.986
## log(RecurrenceInterval) 1052.234     88.834  11.845   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 578.3 on 52 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.7244
## F-statistic: 140.3 on 1 and 52 DF,  p-value: < 2.2e-16
```

```r
Mystery.RImodel.Full$coefficients
```

```
##             (Intercept) log(RecurrenceInterval)
##               -2.000535             1052.234166
```

```r
#Predicted discharge of a 100 year flood for the whole record of data
Mystery.RImodel.Full$coefficients[1] + Mystery.RImodel.Full$coefficients[2]*log(100)
```

```
## (Intercept)
##    4843.717
```

```r
#4843 cfs discharge for 100 year flood with all the data
```

12. How did the recurrence interval plots and predictions of a 100-year flood differ among the two data frames? What does this tell you about the stationarity of discharge in this river?

    Discharge for the recurrence interval plots shows that as recurrence interval increases in both the first 30 years of data (black points) and the full set of data (blue/green points), the peak discharge increaes exponentially. Both data frames had fairly similar trends in the data for the recurrence interval plotted against peak discharge. This graph shows that stationarity of discharge in this river is fairly consistent. The data points with all records of data and the data points with the first 30 years of data represent stationarity in terms of discharge at this river. The few points that have very high flows (two 30 year data points and one all year data points) could be attributed to an extreme weather event, like a hurricane. Discharge for the 100-year flood with the first 30 years of data predicts that the discharge will be about 5538 cfs. For the 100-year flood using all data, discharge was predicted to be 4843 cfs. Although the discharge values are not that close, they are similar enough to assume that discharge in this river is stationary. If the data points were more spread out, particularly at the lower peak discharge levels when the recurrence interval is smaller, then the river discharge may not be representing staitonarity.

## Reflection

13. What are 2-3 conclusions or summary points about river discharge you learned through your analysis?

    Summary points about river discharge that I learned from this analysis is that discharge can be seasonally dependent and have seasonal patterns which are easy to see and understand when looking at graphs. I also learned that peak discharge and recurrence intervals can be exponentially related. This is likely because the lower the peak discharge, the more likely that rain even will happen, which will result in a lower reucrrence interval.

14. What data, visualizations, and/or models supported your conclusions from 13?

    The data that support point about seasonal dependency is the graph with the median, 25th percentile, and 75th percentile discharge of this mystery site river. This graph shows that there are seasonal patterns, particularly since there is very high discharge during the winter months. For the recurrence interval, the visualization that helps with understanding that summary point is the last graph in the lab. This graph shows the recurrence interval and the peak discharge at that recurrence interval. This visualization supports my conclusion from 13.

15. Did hands-on data analysis impact your learning about discharge relative to a theory-based lesson? If so, how?

    Hands-on data analysis impacted my learning about discharge in that it helped me visualize and understand the graphs that we were producing. It helped to use real world data as well since that gave me a better understanding of how discharge can vary seasonally as well as climactically.

16. How did the real-world data compare with your expectations from theory?

    The real-world data was what I expected to see based off the theory. Although I thought some of the real-world data was messier than I expected (or remembered it was going to be!), it was easy to work with. Having to figure out what to do with missing days in the dataset reminded me that I was working with real world data.