Assignment 3: Physical Properties of Rivers

Walker Grimshaw

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.

```
# keep warnings from appearing in knitted pdf
knitr::opts_chunk$set(warning = FALSE)

getwd()
```

```
## [1] "C:/Users/walke/OneDrive/Documents/Duke/Courses/Fall_2019/Hydrologic_Data_Analysis/Assignments"
```

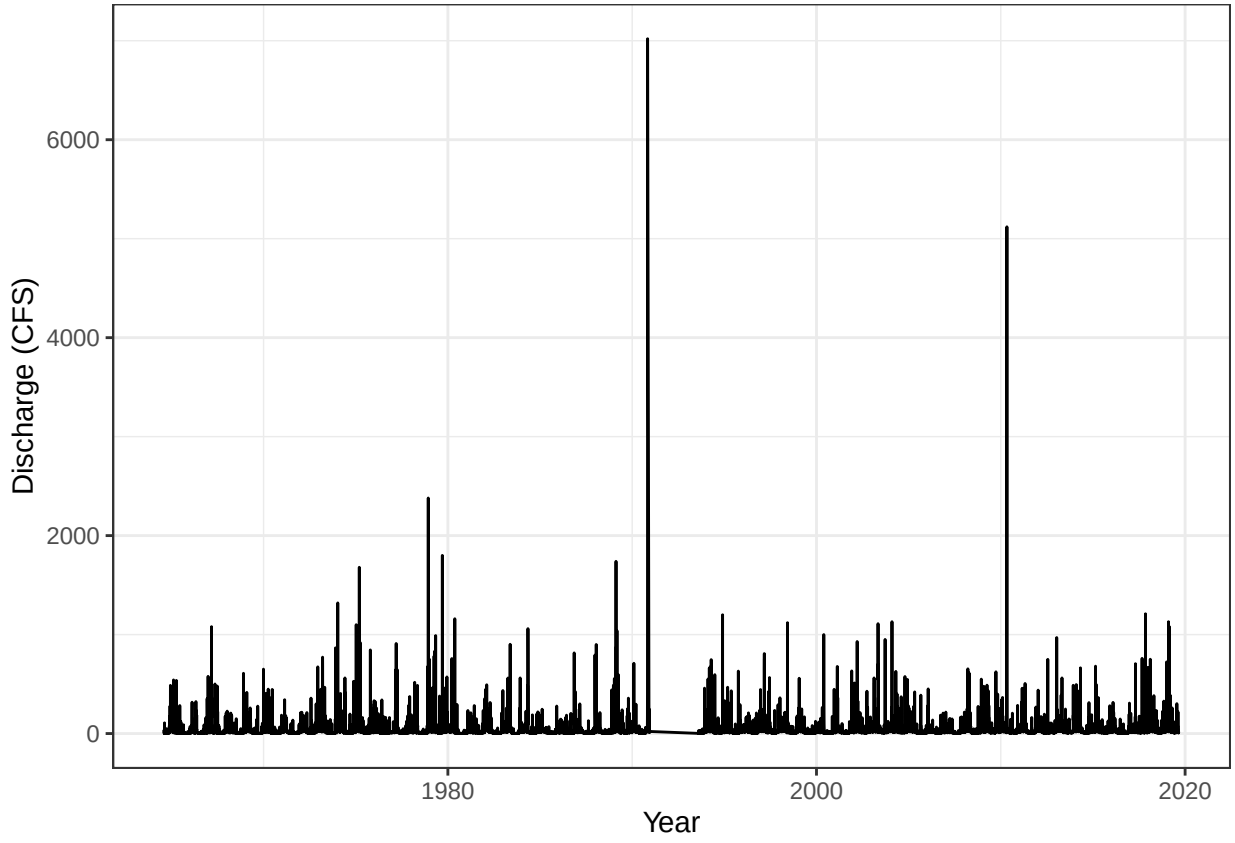
```
library(tidyverse)
library(dataRetrieval)
library(cowplot)
library(lubridate)

theme_set(theme_bw())

## Load data
MysterySiteDischarge <- readNWISdv(siteNumbers = "03431700",
                                   parameterCd = "00060", # discharge (ft3/s)
                                   startDate = "",
                                   endDate = "")
## Change column names
names(MysterySiteDischarge)[4:5] <- c("Discharge", "Approval.Code")

MysterySiteDischarge.plot <-
  ggplot(data = MysterySiteDischarge, aes(x = Date, y = Discharge)) +
  geom_line() +
```

```
  labs(x = "Year", y = "Discharge (CFS)")
print(MysterySiteDischarge.plot)
```



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.

```
## Add year and day columns
MysterySiteDischarge <- MysterySiteDischarge %>%
  mutate(Year = year(Date)) %>%
  mutate(Day.of.Year = yday(Date))

MysterySiteDischarge.Pattern <- MysterySiteDischarge %>%
  group_by(Day.of.Year) %>%
  summarize(Median = median(Discharge),
            Q75 = quantile(Discharge, p = 0.75),
            Q25 = quantile(Discharge, p = 0.25))

## Plot
ggplot(data = MysterySiteDischarge.Pattern,
```

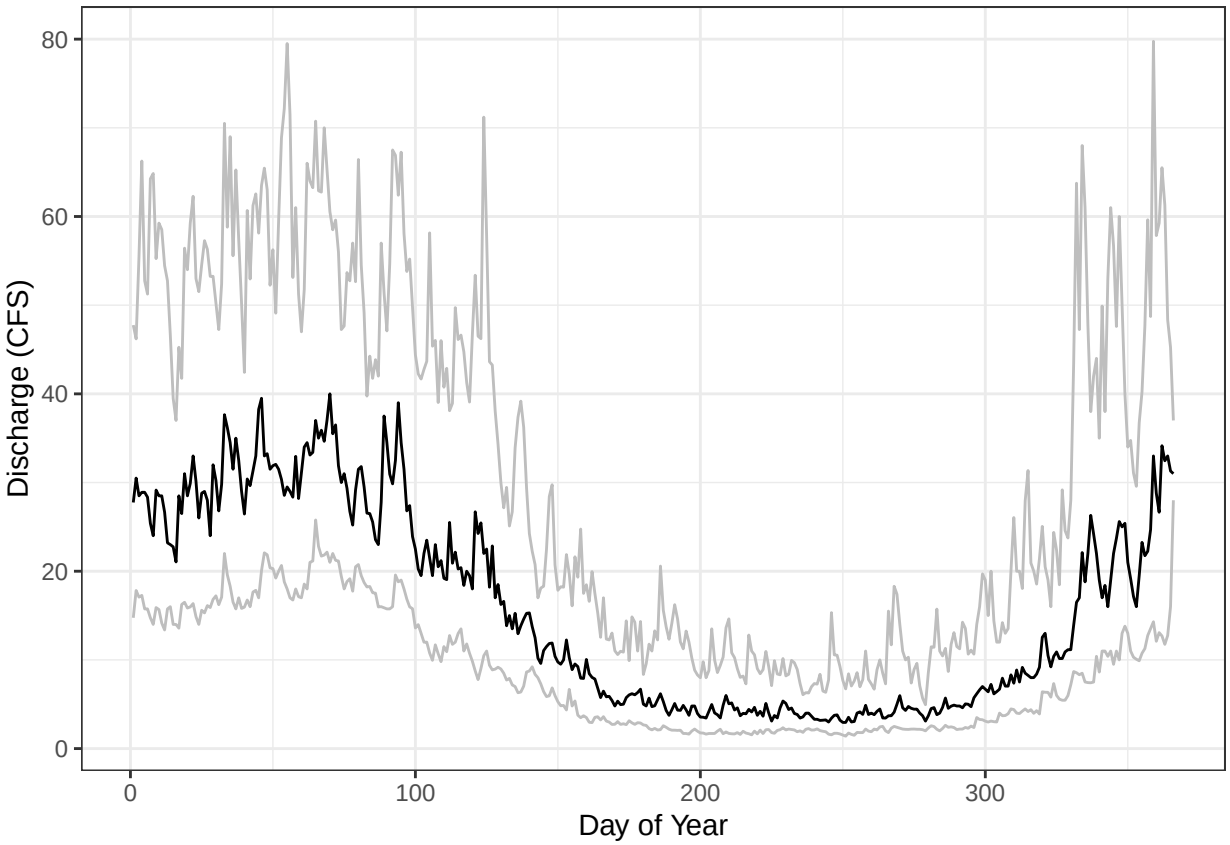


Figure 1: Discharge for the mystery site, showing the median discharge in black and the 25th and 75th quantiles in gray.

```
      aes(x = Day.of.Year)) +
  geom_line(aes(y = Median)) +
  geom_line(aes(y = Q25), color = "gray") +
  geom_line(aes(y = Q75), color = "gray") +
  labs(y = "Discharge (CFS)", x = "Day of Year")
```

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

   Discharge is consistently low from roughly day 150 to 325, or June through mid-November. In the winter through spring, discharge is higher, peaking around mid-March. This likely means precipitation is reliable in the winter and scarce in the summer. The peak is early enough in the year that there is likely little snowfall in the watershed and mostly rainfall.

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 exceedence 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

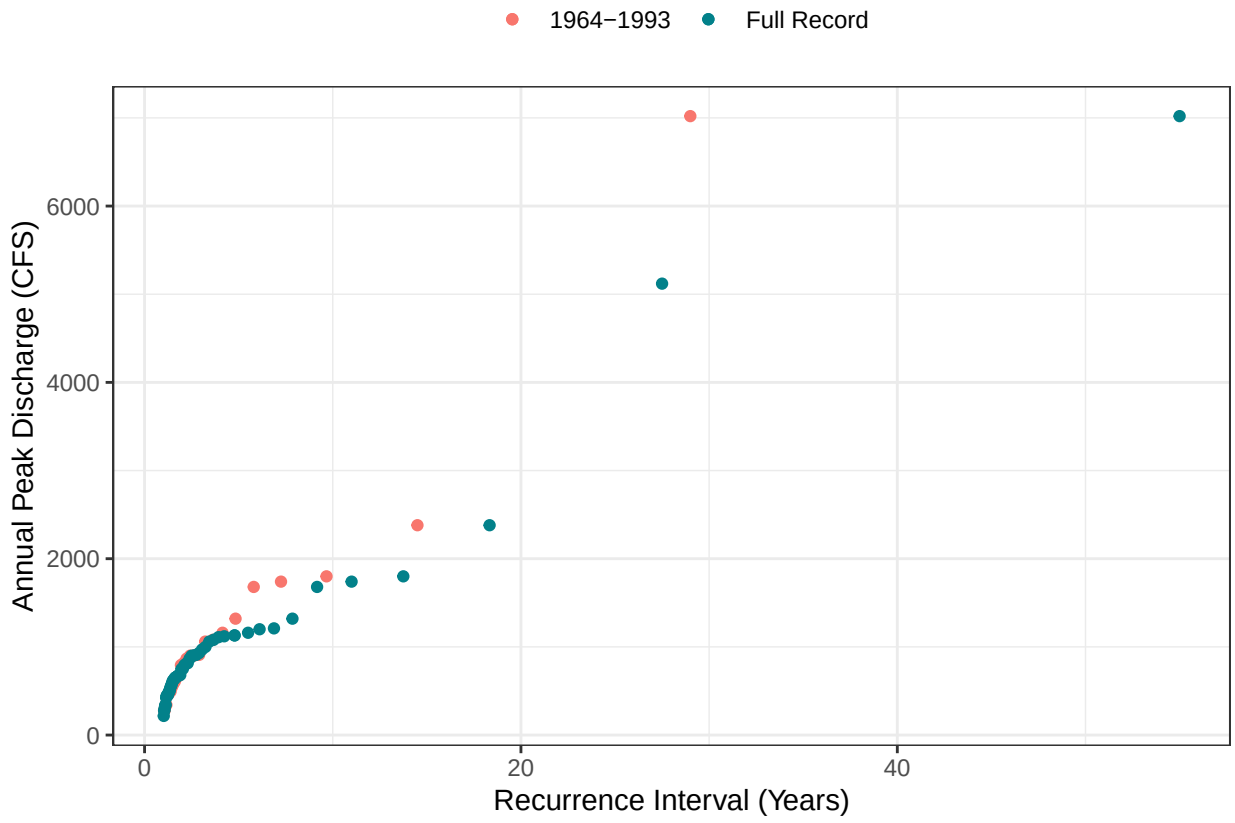
other set of points includes the values computed for all years of the record.

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

```
## First 30 years exceedence data frame, 1964-1993
MysterySite.Annual.30yr <-
  MysterySiteDischarge %>%
  filter(Year < 1994) %>%
  group_by(Year) %>%
  summarize(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
         # rank normally gives largest rank to largest number
         RecurrenceInterval = (length(Year) + 1)/Rank,
         Probability = 1/RecurrenceInterval)

## Full exceedence data frame
MysterySite.Annual.Full <-
  MysterySiteDischarge %>%
  group_by(Year) %>%
  summarize(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
         # rank normally gives largest rank to largest number
         RecurrenceInterval = (length(Year) + 1)/Rank,
         Probability = 1/RecurrenceInterval)

## Plot of Discharge vs Recurrence Interval
ggplot(mapping = aes(x = RecurrenceInterval)) +
  geom_point(data = MysterySite.Annual.30yr,
             aes(y = PeakDischarge, color = "1964-1993")) +
  geom_point(data = MysterySite.Annual.Full,
             aes(y = PeakDischarge, color = "Full Record")) +
  labs(x = "Recurrence Interval (Years)",
       y = "Annual Peak Discharge (CFS)") +
  scale_color_manual(name = "", values = c("#F8766D", "#02818a")) +
  theme(legend.position = "top")
```




```
## Recurrence Interval Models, logarithmic

## First 30
MysteryModel.30.log <- lm(data = MysterySite.Annual.30yr,
                      PeakDischarge ~ log(RecurrenceInterval))
summary(MysteryModel.30.log)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ log(RecurrenceInterval), data = MysterySite.Annual.30yr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -982.36 -366.95   42.38  247.09 2838.21
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -107.5      197.3  -0.545     0.59
## log(RecurrenceInterval)   1273.8      157.1   8.107 1.38e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 689.4 on 26 degrees of freedom
## Multiple R-squared:  0.7166, Adjusted R-squared:  0.7057
## F-statistic: 65.73 on 1 and 26 DF,  p-value: 1.377e-08
```

```r
MysteryModel.30.log$coefficients[1] +
  MysteryModel.30.log$coefficients[2]*log(100)
```

```
## (Intercept)
##    5758.621
```

```r
# low R squared so try without logarithm
MysteryModel.30 <- lm(data = MysterySite.Annual.30yr,
                      PeakDischarge ~ RecurrenceInterval)
summary(MysteryModel.30)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ RecurrenceInterval, data = MysterySite.Annual.30yr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -973.89  -30.29   49.21  126.61  524.10
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           211.88      60.98   3.474  0.00181 **
## RecurrenceInterval    216.69       8.77  24.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 261.7 on 26 degrees of freedom
## Multiple R-squared:  0.9592, Adjusted R-squared:  0.9576
## F-statistic: 610.5 on 1 and 26 DF,  p-value: < 2.2e-16
```

```r
# better R squared, but does it makes physical sense?
MysteryModel.30$coefficients[1] +
  MysteryModel.30$coefficients[2]*100
```

```
## (Intercept)
##     21880.9
```

```r
## Full Record
MysteryModel.Full.log <- lm(data = MysterySite.Annual.Full,
                      PeakDischarge ~ log(RecurrenceInterval))
summary(MysteryModel.Full.log)
```

```
##
## 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
```

```
MysteryModel.Full.log$coefficients[1] +
  MysteryModel.Full.log$coefficients[2]*log(100)
```

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

```
MysteryModel.Full <- lm(data = MysterySite.Annual.Full,
                        PeakDischarge ~ RecurrenceInterval)
summary(MysteryModel.Full)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ RecurrenceInterval, data = MysterySite.Annual.Full)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -441.3 -113.2   18.1  106.1 1181.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         415.781     36.224   11.48  7.2e-16 ***
## RecurrenceInterval  128.100      3.795   33.76  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 232.3 on 52 degrees of freedom
## Multiple R-squared:  0.9564, Adjusted R-squared:  0.9555
## F-statistic:  1139 on 1 and 52 DF,  p-value: < 2.2e-16
```

```
MysteryModel.Full$coefficients[1] +
  MysteryModel.Full$coefficients[2]*100
```

```
## (Intercept)
##    13225.76
```

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?

    The 30 year log model predicts a 100-year flood of 5,759 cfs while the full log model predicts a 100-year flood of 4844 cfs. The 30 year linear model predicts a 100-year flood of 21,880 cfs while the full linear model predicts a 100-year flood of 13,226 cfs. The linear models have higher R-squared values that fit the data better, but physically, the logarithmic models make more sense. Either way, the models indicate that using the full record predicts a 100-year flood of lesser magnitude than using only the first 30 years. The river discharge is likely non-stationary, with decreasing flow over time.

Reflection

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

    The seasonality of river discharge depends greatly on the geographic location of the river and the climate of that area. Relatedly, the discharge pattern of a river may change over time due

to climate change or other long-term climatic factors. It is difficult to predict the magnitude of large, rare flood events from historical records as well.

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

    The class lessons in concert with the analysis here, especially the discharge over time figures, show how discharge changes with season and over time. The seasonal statistical flow figure from this analysis especially showed the seasonal pattern of the river in question. The recurrence interval plot and associated models however show how difficult it is to predict rare floods and even what model would fit this historical data best to be used for future predictions.

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

    As with the lake lessons, I think struggling through an actual river discharge analysis and trying to think about where this mystery river is were helpful for a deeper and more analytical approach to studying discharge. The combination made it so I could neither just plug numbers into formulae without thinking about their real world meaning nor just memorize a few concepts of river hydrology.

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

    I assumed a logarthimic relationship would always provide the best model for annual peak discharge from recurrence interval. However, for the data from this site, a linear model was a better predictor of the annual peak discharge because of the two storm events in 1990 and 2010 that threw off the logarithmic model.