Assignment 3: Physical Properties of Rivers

Rachel Bash

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.

```
getwd()
```

```
## [1] "C:/Users/19524/Documents/DUKE/Hydrologic Data Analytics/Hydrologic_Data_Analysis/Assignments"
```

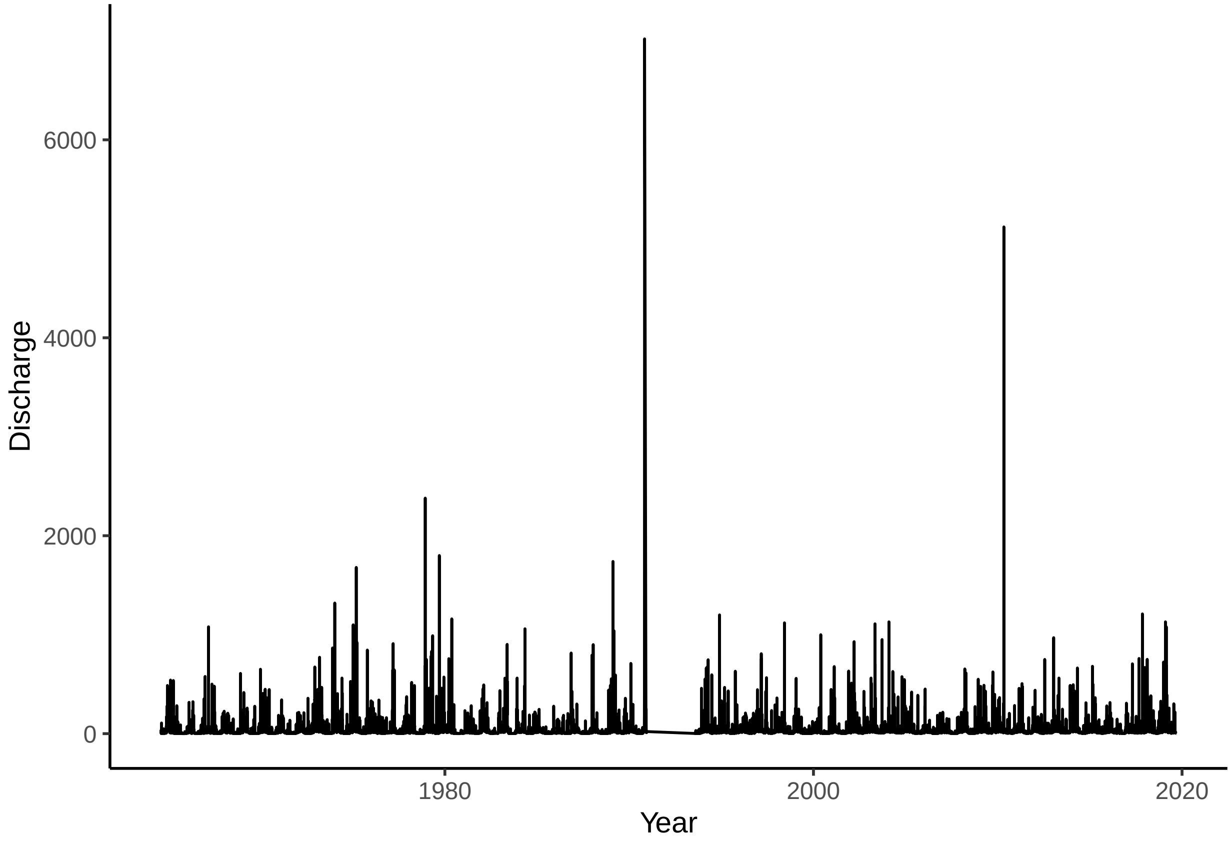
```
library(tidyverse)
library(dataRetrieval)
library(lubridate)

theme_set(theme_classic())

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

names(MysterySiteDischarge)[4:5] <- c("Discharge", "Approval.Code")

#build ggplot
MysteryPlot.all <- ggplot(MysterySiteDischarge, aes(x=Date, y=Discharge)) +
  geom_line() +
  xlab("Year")
print(MysteryPlot.all)
```



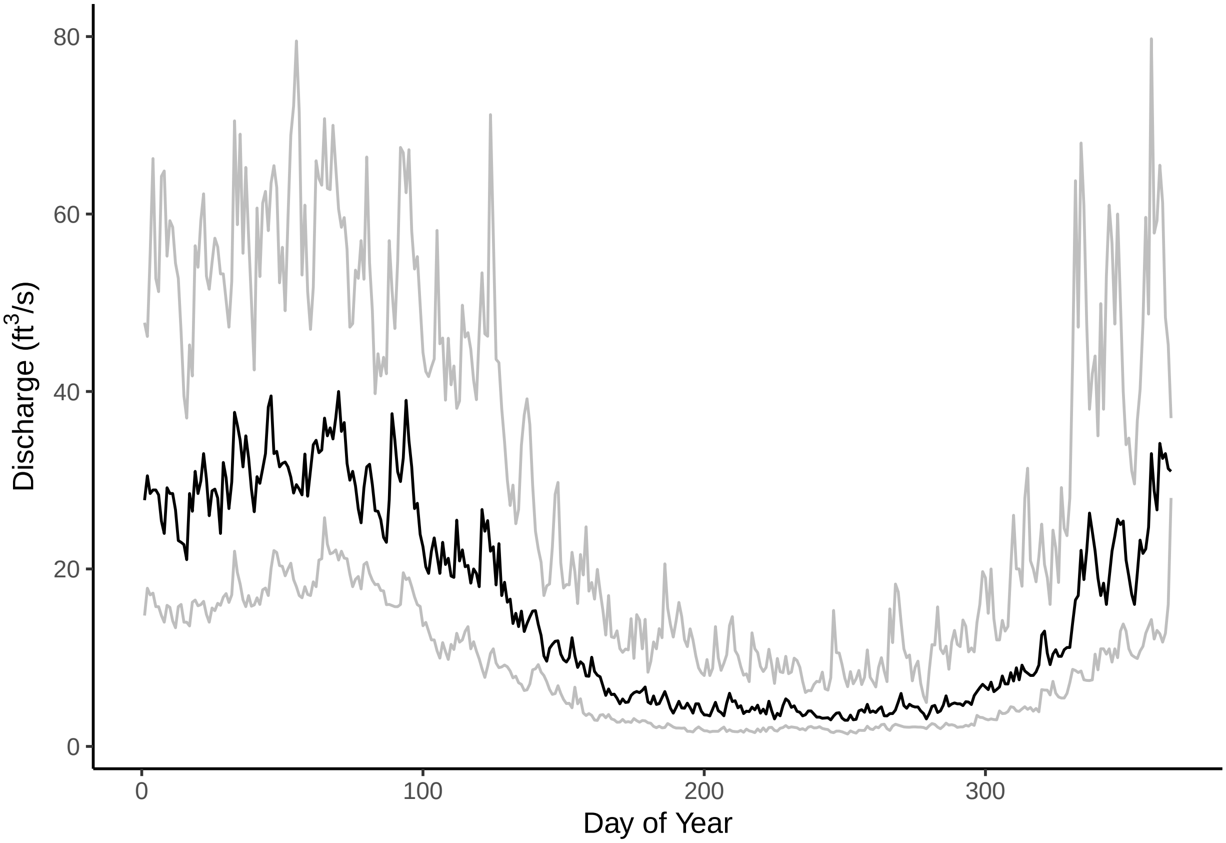
Analyze seasonal patterns in discharge

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.

```
MysterySiteDischarge <- MysterySiteDischarge %>%
  mutate(Year = year(Date)) %>%
  mutate(Day.of.Year = yday(Date))

MysterySiteDischarge.Pattern <- MysterySiteDischarge %>%
  group_by(Day.of.Year) %>%
  summarise(Median = median(Discharge),
            percent75 = quantile(Discharge, probs = 0.75),
            percent25 = quantile(Discharge, probs = 0.25))


Mystery.Pattern.Plot <-
  ggplot(MysterySiteDischarge.Pattern, aes(x = Day.of.Year)) +
  geom_line(aes(y = Median)) +
  geom_line(aes(y = percent75), color = "gray") +
  geom_line(aes(y = percent25), color = "gray") +
  labs(x = "Day of Year", y = expression("Discharge (ft"^3*"/s)"))
print(Mystery.Pattern.Plot)
```



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

   The plot indicates that there is a clear pattern of a rainy season or high water(perhaps from snowpack) from about November to May and then a dry season with little rainfall/snowpack runoff from June through October. The water level stays low and fairly consistent during the summer months(middle of the year), indicating that it is probably hot and dry. There is higher variability (both in the median and in the upper and lower quartiles) in the beginning and end of the year due to the rain/snowpack runoff, which means that the river is sensitive to the weather and fluctuates regularly during that time.

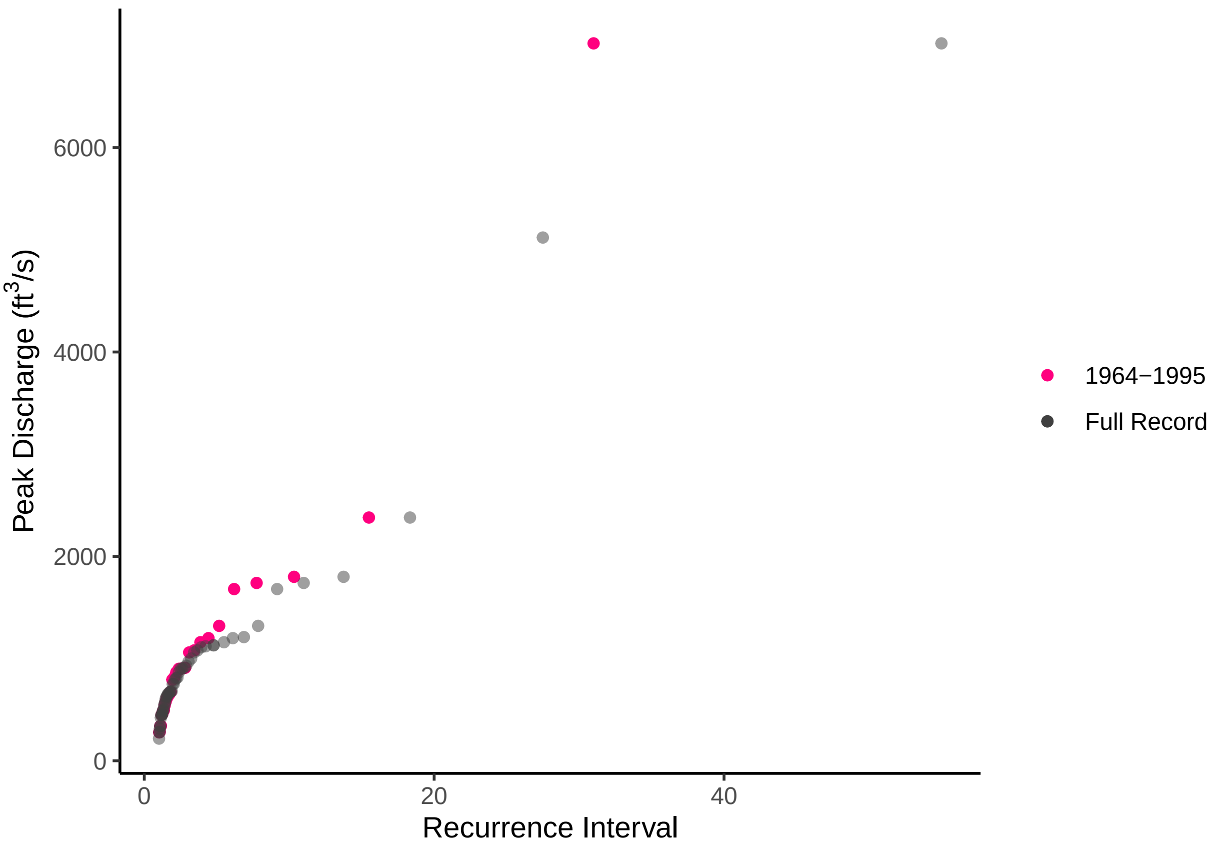
Create and analyze recurrence intervals

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
#creating data frames
MysterySite.Annual.30yr <- MysterySiteDischarge %>%
  filter(Year<"1995-07-31") %>%
  group_by(Year) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
         RecurrenceInterval = (length(Year) + 1)/Rank,
         Probability = 1/RecurrenceInterval)

MysterySite.Annual.Full <- MysterySiteDischarge %>%
  group_by(Year) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
         RecurrenceInterval = (length(Year) + 1)/Rank,
         Probability = 1/RecurrenceInterval)

#plot
MysteryRecurrence.Plot <-
  ggplot() +
  geom_point(data = MysterySite.Annual.30yr,
             aes(x = RecurrenceInterval, y = PeakDischarge, color = "1964-1995")) +
  geom_point(data = MysterySite.Annual.Full,
             aes(x = RecurrenceInterval, y = PeakDischarge,
                 color="Full Record"), alpha=0.5) +
  theme(legend.title=element_blank()) +
  scale_color_manual(name="", values = c("#FF007F", "#404040")) +
  labs(x= "Recurrence Interval", y = expression("Peak Discharge (ft"^3*"/s)"))
print(MysteryRecurrence.Plot)
```



```
#log models to predict 100-year flood
MysteryModel.30yr <- lm(data = MysterySite.Annual.30yr, PeakDischarge ~ log(RecurrenceInterval))
MysteryModel.Full <- lm(data = MysterySite.Annual.Full, PeakDischarge ~ log(RecurrenceInterval))
summary(MysteryModel.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
```

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

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

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

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

```r
#log models didn't fit well and gave weird results for the 100-year flood
#discharge values, so I decided to see what it looks like without the log.
MysteryModel.30yr <- lm(data = MysterySite.Annual.30yr, PeakDischarge ~
                          RecurrenceInterval)
MysteryModel.Full <- lm(data = MysterySite.Annual.Full, PeakDischarge ~
                          RecurrenceInterval)
summary(MysteryModel.30yr)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ RecurrenceInterval, data = MysterySite.Annual.30yr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -986.64  -42.32   48.50  134.76  538.55
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         251.824     58.153    4.33 0.000172 ***
## RecurrenceInterval  200.956      8.092   24.83  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 260.7 on 28 degrees of freedom
## Multiple R-squared:  0.9566, Adjusted R-squared:  0.955
## F-statistic: 616.7 on 1 and 28 DF,  p-value: < 2.2e-16
```

```
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.30yr$coefficients[1] +MysteryModel.30yr$coefficients[2]*100
```

```
## (Intercept)
##    20347.41
```

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

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

```
#models without the log fit much better and give 100-year flood discharge
#values that make much more sense
```

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 first 30 years of data actually have a higher proportion of large flood events, and the points on the plot show that the recurrence interval for peak discharges is smaller in the 30 year window than in the full record. This means that there were higher peak discharges in the first 30 years than in the second half of the full record data frame. I decided to use a linear model instead of a log model for two reasons. First, the linear model had a much better fit (higher R^2 value). Secondly, the linear model predicted a 100-year flood discharge that made much more sense to me than the log model. Using the 30-year linear model, the 100-year flood discharge value is 20347 cfs, and the full record model has a 100-year discharge value of 13226 cfs. The 30 year model gives a higher 100-year flood prediction, which matches the conclusions that were made from the plot. Because the two models predict very different 100-year flood intervals, I can conclude that this river system does NOT reflect stationarity principles. Rather, it seems that peak discharge is becoming more subdued as time goes on.

Reflection

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

    The river discharge is low in the summer months and higher and more variable in the winter months (November through April), likely due to rainy/snowy winters and hot, dry summers. The plot I created shows that high peak discharges happen more often (have a lower recurrence interval) in the first 30 years of data than in the full record, indicating non-stationarity of the river. I predict that this river is in a location that has been receiving less water over time, as peak discharge events are occuring less often. This may indicate that this river is in an area where droughts could be a possibility.

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

    The first plot of the discharge data by year supports the conclusion that high peak discharges are occuring less frequently in the later part of the record (save for the two major discharge events in 1990 and 2010). The plot showing seasonality of the river is a clear visualization that there is little rainfall in the warmer months of the year. Lastly, the recurrence interval plot shows that high peak discharges happened more frequently (steeper curve) in the 30-year data frame than in the full record data frame.

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

    Yes. When comparing the conlusions of this assignment to the river systems we looked at in class, it is clear that that climate change can have different effects on different river systems. Also, even though I have learned about recurrence intervals in previous classes, creating the plot myself makes me understand it a lot better.

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

    The recurrence interval to discharge curve isn't as neat as theory would suggest. The two major discharge events in this data set certainly skew the data and makes the curve a little wonky, creating problems when trying to make a log-based model. The log model does not fit the data well due to the rare, extreme discharge events. Therefore, it made more sense to employ a linear model in this case.