Assignment 3: Physical Properties of Rivers

Haoyu Zhang

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] "D:/William/Duke/Study/EOS 722/Hydrologic_Data_Analysis/Assignments/Assignment 3"
```

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(cowplot)
```

```
##
## ************************************************************
## Note: As of version 1.0.0, cowplot does not change the
##   default ggplot2 theme anymore. To recover the previous
##   behavior, execute:
##   theme_set(theme_cowplot())
## ************************************************************
##
## Attaching package: 'cowplot'
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
theme_set(theme_classic())

MysterySiteDischarge <- readNWISdv(siteNumbers = "03431700",
                     parameterCd = "00060", # discharge (ft3/s)
                     startDate = "",
                     endDate = "")
colnames(MysterySiteDischarge)[4:5] <- c("Discharge", "Approval.Code")
head(MysterySiteDischarge)
```

```
##   agency_cd  site_no       Date Discharge Approval.Code
## 1      USGS 03431700 1964-08-01       1.0             A
## 2      USGS 03431700 1964-08-02       0.8             A
## 3      USGS 03431700 1964-08-03       0.6             A
## 4      USGS 03431700 1964-08-04       0.5             A
## 5      USGS 03431700 1964-08-05       0.5             A
## 6      USGS 03431700 1964-08-06       0.4             A
```

Analyze seasonal patterns in discharge

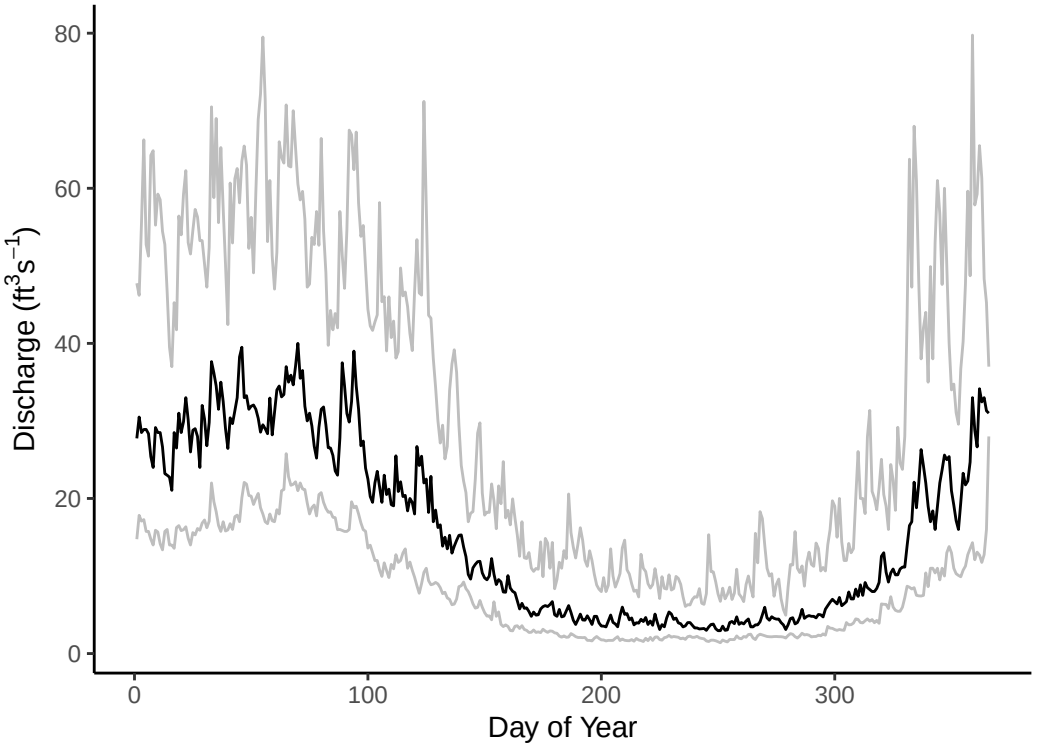
5. Add a "Year" and "Day.of.Year" column to the data frame.
6. Create a new data frame called "MysterySiteDischarge.Pattern" that has columns for Day.of.Year, median discharge for a given day of year, 75th percentile discharge for a given day of year, and 25th percentile discharge for a given day of 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 <-
  mutate(MysterySiteDischarge, Year = year(Date)) %>%
  mutate(DOY = yday(Date))

MysterySiteDischarge.Pattern <- MysterySiteDischarge %>%
  select("DOY", "Discharge") %>%
  group_by(DOY) %>%
  summarise(Q1 = quantile(Discharge, probs = 0.25),
            Median = quantile(Discharge, probs = 0.5),
            Q3 = quantile(Discharge, probs = 0.75))

ggplot(MysterySiteDischarge.Pattern, aes(x = DOY)) +
```

```
  geom_line(aes(y = Q1), col = "gray") +
  geom_line(aes(y = Median), col = "black") +
  geom_line(aes(y = Q3), col ="grey") +
  labs(x = "Day of Year", y = expression("Discharge (ft"^3*s^-1*")"))
```



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

   Discharge starts to decrease from around 100th day (March - April) and remains low throughout summer until about 300th day when discharge begins to rise. The pattern suggests that this river is most likely driven by precipitation during winter and spring, and summer appears to be the local dry season. Furthermore, there tend to be greater variations during winter than summer, suggesting the precipitation during winter might vary from year to year.

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.

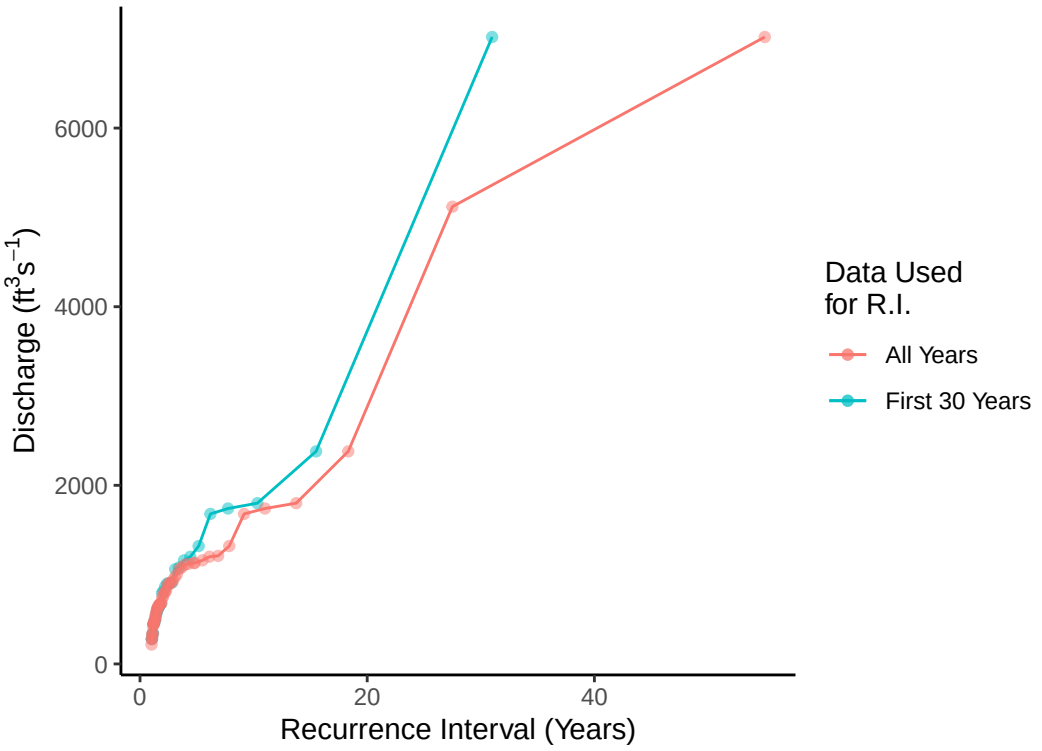
11. Create a model to predict the discharge for a 100-year flood for both sets of recurrence intervals.

```
MysterySite.Annual.30yr <- MysterySiteDischarge %>%
  filter(Year < min(Year) + 32) %>%
  group_by(Year) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
```

```
  mutate(Rank = rank(-PeakDischarge),
         RecurrenceInterval = (length(Year) + 1) / Rank,
         Probability = 1 / RecurrenceInterval)

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

library(stringr)
ggplot(MysterySite.Annual.30yr,
       aes(x = RecurrenceInterval, y = PeakDischarge, col = "First 30 Years")) +
  geom_point(alpha = 0.5) + geom_line() +
  geom_point(data = MysterySite.Annual.all, alpha = 0.5,
             aes(x = RecurrenceInterval, y = PeakDischarge, col = "All Years")) +
  geom_line(data = MysterySite.Annual.all,
             aes(x = RecurrenceInterval, y = PeakDischarge, col = "All Years")) +
  labs(x = "Recurrence Interval (Years)" , y = expression("Discharge (ft"^3*s^-1*")"),
       col = str_wrap("Data Used for R.I.", width = 10)) +
  theme(legend.margin = margin(0,0,0,0,"cm"))
```




```
model.30 <- lm(data = MysterySite.Annual.30yr, PeakDischarge ~ RecurrenceInterval)
summary(model.30)
```

```
##
## 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
```

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

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

```
##
## Call:
## lm(formula = PeakDischarge ~ RecurrenceInterval, data = MysterySite.Annual.all)
##
## 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
```

```
summary(model.all)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ RecurrenceInterval, data = MysterySite.Annual.all)
##
## 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
```

```
#model.all.log <- lm(data = MysterySite.Annual.all, PeakDischarge ~ log(RecurrenceInterval))
#summary(model.all.log)


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

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

```
model.all$coefficients[1] + model.all$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 model constructed from the first 30 years yields a higher prediction than the model used all available data. The difference may indicate that the discharge of this river is not stationary and discharge on avergae decreases over years.

Reflection

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

14. The driving factor(s) for the discharge of an unknown river could potentially be determined by examining the relationship between discharge compiled across multiple years and the day of year. In addition, quartiles (e.g. Q1, median, Q2) are helpful in reflecting the variation throughout a year as well as among different years but on the same day of year. In this case, the discharge in the unknown river seems to be driven by precipitation during winter and spring.

15. The linear regression model constructed for the relationship between discharge and recurrence interval can be used as a tool to predict the discharge of a given R.I. or vice versa. The predicted discharge for a 100-year R.I. is 20347 ft^3 s^-1 and 13225 ft^3 s^-1 according to the model using 30 years and the model using all data, respectively. The 30-year model gives a higher prediction, and the reason could be that the discharge decreases over years.

16. It is noteworthy that the relationship could be linear without logarithm-transformation, in contrast with the example from the class. In fact, for this assignment the the model without log-transformation gives a higher R-squared value (0.956 for all years; 0,955 for 30 years) than the one with transformation (0.724 and 0.699 respectively). Indeed, a logarithm curve is not shown on the graph for both sets of data, and data points appear to scatter around a straight line.

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

18. The conclusion is drawn merely based on the visualization of discharge among different days of year (The first graph).

19. The numeric predictions were produced by the two linear regression models, and the trend of discharge over years is postulated based on the difference.

20. The choice of models was made according to the statistical summary of each model.

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

    On this topic, yes, because I came to know how the data were sorted, compiled, and visualized so that I could interpret any trend shown in river discharge. Also, the hands-on data analysis is always helpful for familiarizing myself with using R to retrieve, sort, and properly display data as graphs.

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

    Generally, real-world data do no deviate greatly from my expectations, but there might be slight difference in some details, such as relationship between discharge and R.I. is not always log-transformed, as is the case with the rivers analyzed in class.