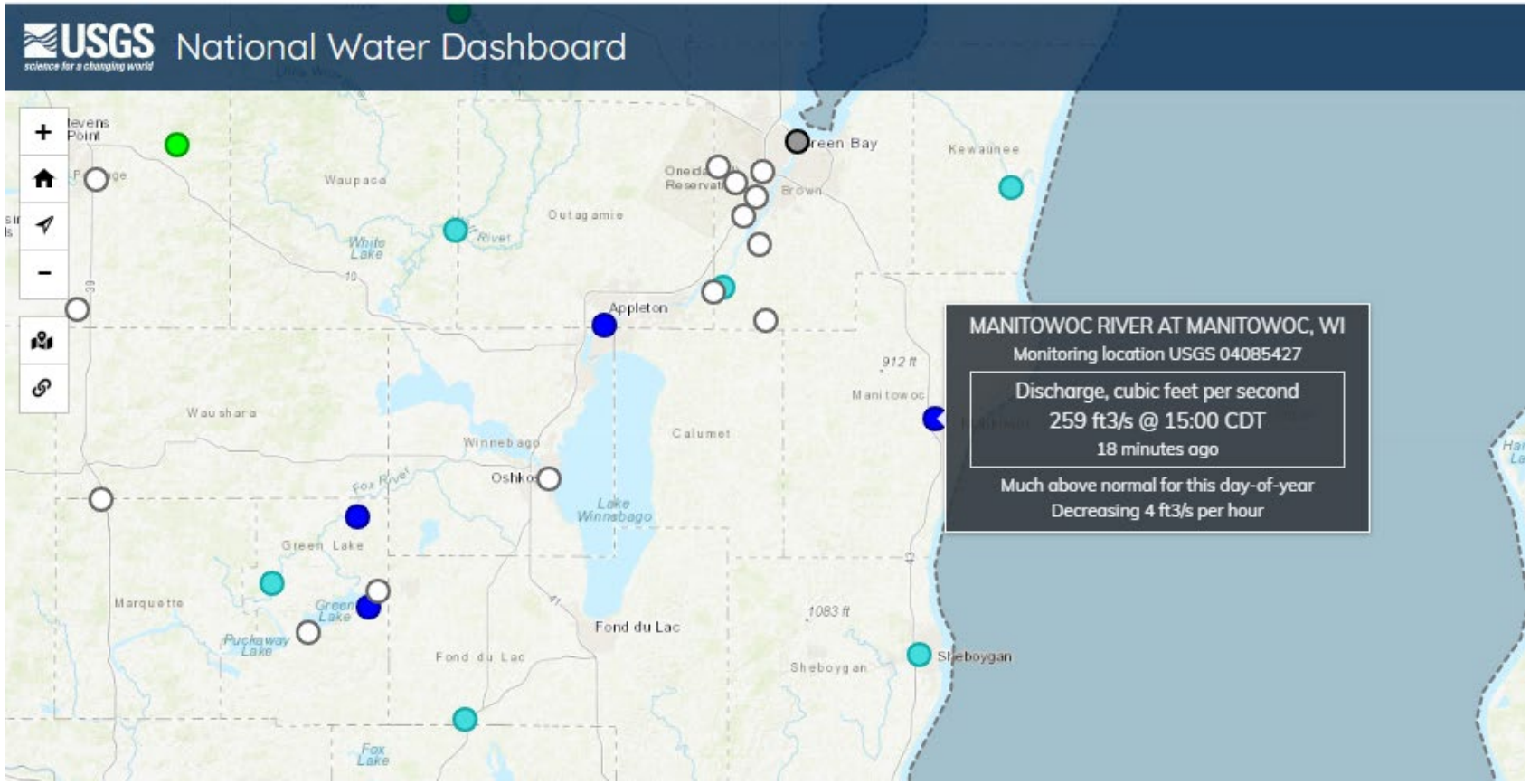
Acquiring and Wrangling Hydrologic and Water Quality Data
Flow Data

Eric Hettler

Wisconsin Department of Natural Resources

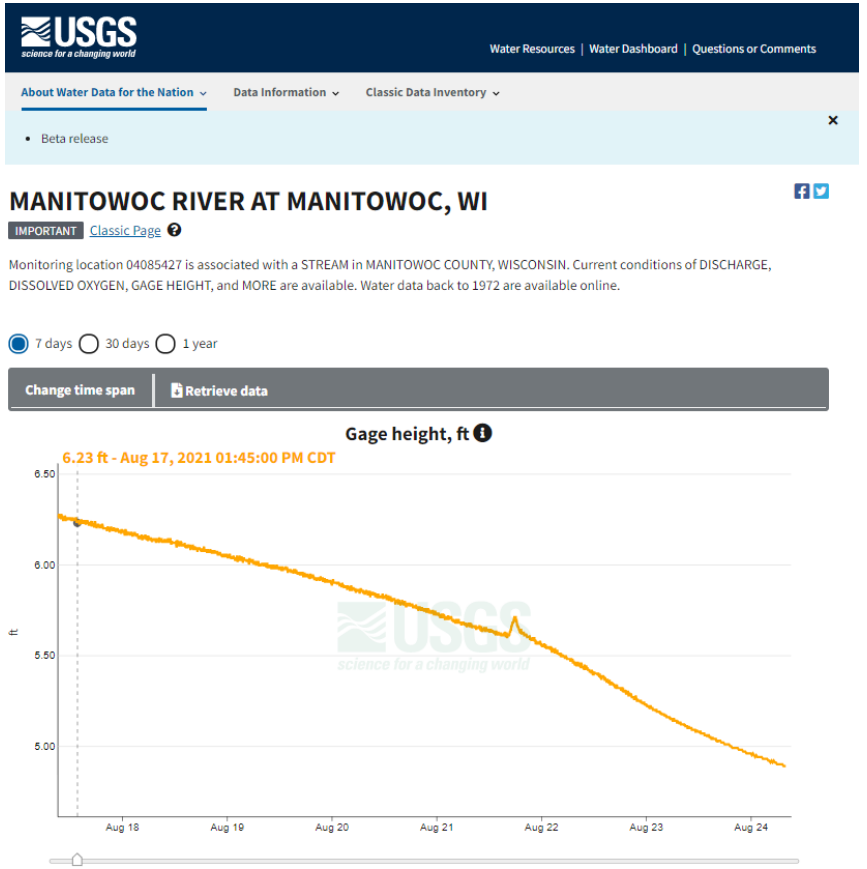
Manual Download of USGS Flow Data

USGS National Water Dashboard



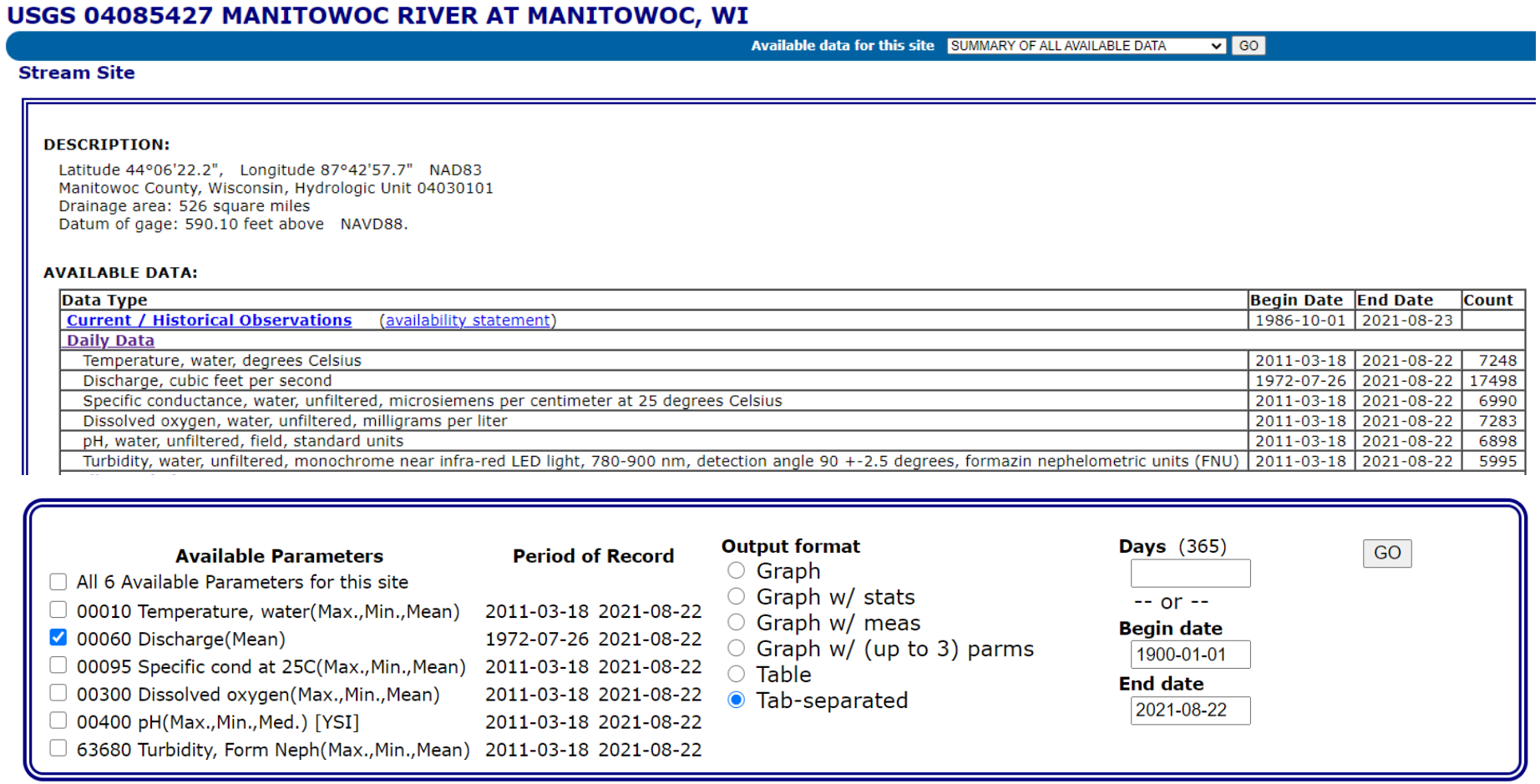
https://dashboard.waterdata.usgs.gov/app/nwd/?region=lower48&aoi=default

USGS Dashboard: Site Page



https://waterdata.usgs.gov/monitoring-location/04085427/

USGS Classic Page



USGS 04085427 MANITOWOC RIVER AT MANITOWOC, WI

Available data for this site: SUMMARY OF ALL AVAILABLE DATA — GO

Stream Site

DESCRIPTION:

Latitude 44°06'22.2", Longitude 87°42'57.7" NAD83
Manitowoc County, Wisconsin, Hydrologic Unit 04030101
Drainage area: 526 square miles
Datum of gage: 590.10 feet above NAVD88.

AVAILABLE DATA:

Data Type	Begin Date	End Date	Count
Current / Historical Observations (availability statement)	1986-10-01	2021-08-23	
Daily Data			
Temperature, water, degrees Celsius	2011-03-18	2021-08-22	7248
Discharge, cubic feet per second	1972-07-26	2021-08-22	17498
Specific conductance, water, unfiltered, microsiemens per centimeter at 25 degrees Celsius	2011-03-18	2021-08-22	6990
Dissolved oxygen, water, unfiltered, milligrams per liter	2011-03-18	2021-08-22	7283
pH, water, unfiltered, field, standard units	2011-03-18	2021-08-22	6898
Turbidity, water, unfiltered, monochrome near infra-red LED light, 780-900 nm, detection angle 90 +-2.5 degrees, formazin nephelometric units (FNU)	2011-03-18	2021-08-22	5995

Available Parameters	Period of Record
☐ All 6 Available Parameters for this site	
☐ 00010 Temperature, water(Max.,Min.,Mean)	2011-03-18 2021-08-22
☑ 00060 Discharge(Mean)	1972-07-26 2021-08-22
☐ 00095 Specific cond at 25C(Max.,Min.,Mean)	2011-03-18 2021-08-22
☐ 00300 Dissolved oxygen(Max.,Min.,Mean)	2011-03-18 2021-08-22
☐ 00400 pH(Max.,Min.,Med.) [YSI]	2011-03-18 2021-08-22
☐ 63680 Turbidity, Form Neph(Max.,Min.,Mean)	2011-03-18 2021-08-22

Output format
- ○ Graph
- ○ Graph w/ stats
- ○ Graph w/ meas
- ○ Graph w/ (up to 3) parms
- ○ Table
- ◉ Tab-separated

Days (365)

-- or --

Begin date 1900-01-01

End date 2021-08-22

GO

https://waterdata.usgs.gov/usa/nwis/uv?site_no=04085427

USGS Classic Page

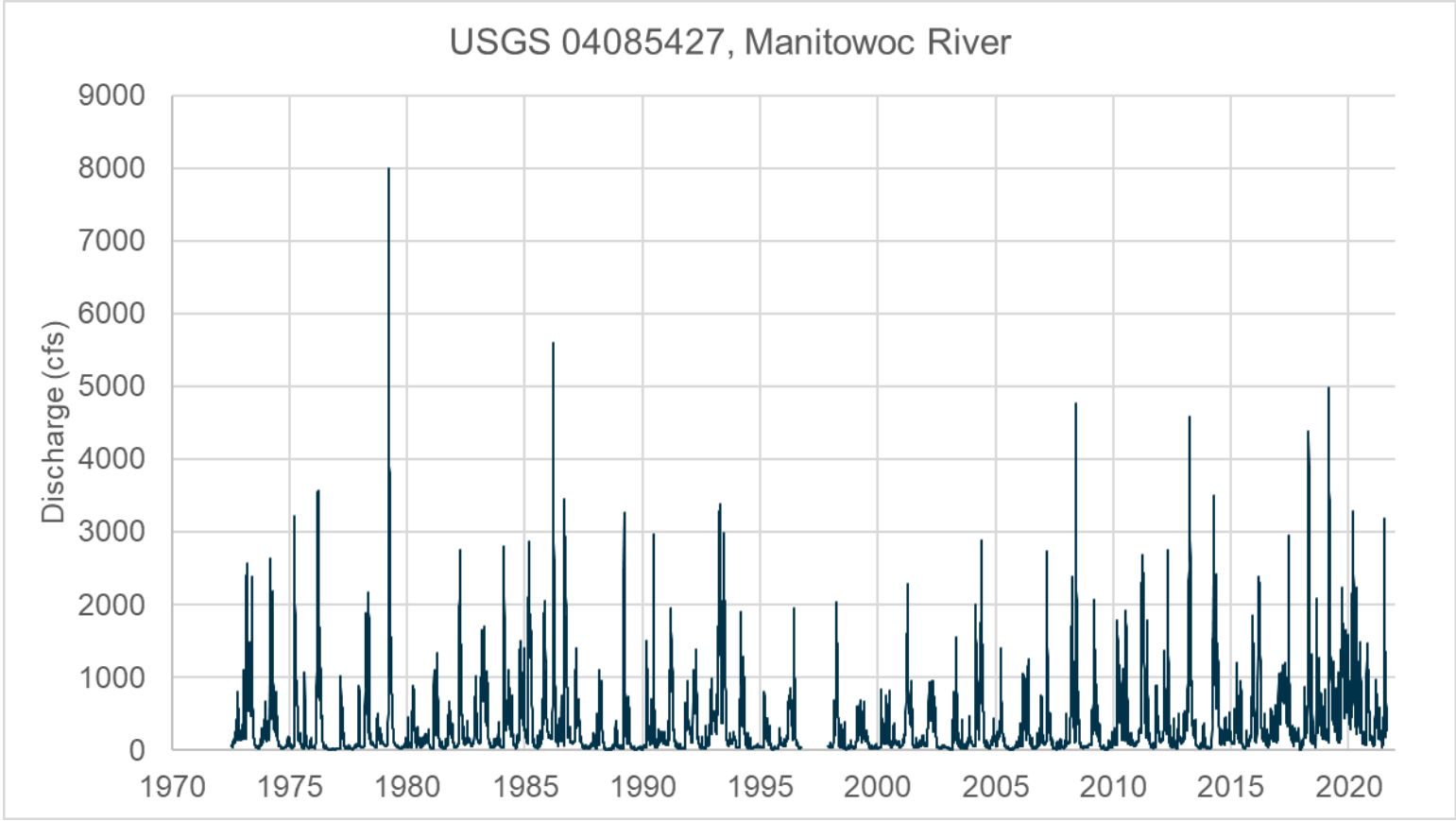
```
#
# Data provided for site 04085427
#            TS   parameter     statistic     Description
#        152896       00060     00003     Discharge, cubic feet per second (Mean)
#
# Data-value qualification codes included in this output:
#
#     A  Approved for publication -- Processing and review completed.
#     P  Provisional data subject to revision.
#     e  Value has been estimated.
#
agency_cd       site_no datetime        152896_00060_00003      152896_00060_00003_cd
5s      15s     20d     14n     10s
USGS    04085427        1972-07-26      69.0    A
USGS    04085427        1972-07-27      65.0    A
USGS    04085427        1972-07-28      57.0    A
USGS    04085427        1972-07-29      58.0    A
USGS    04085427        1972-07-30      57.0    A
USGS    04085427        1972-07-31      54.0    A
```

https://waterdata.usgs.gov/nwis/dv?cb_00060=on&format=rdb&site_no=04085427&referred_module=sw&period=&begin_date=1900-01-01&end_date=2021-08-23

Note: Format of link to tab-separated data lends itself well to data scraping

USGS Discharge Data: Import to Excel

	A	B	C	D	E	F	G
1	# ---------- WARNING ----------						
2	# Some of the data that you have obtained from this U.S. Geological Survey database						
3	# may not have received Director's approval. Any such data values are qualified						
4	# as provisional and are subject to revision. Provisional data are released on the						
5	# condition that neither the USGS nor the United States Government may be held liable						
6	# for any damages resulting from its use.						
7	#						
8	# Additional info: https://help.waterdata.usgs.gov/policies/provisional-data-statement						
9	#						
10	# File-format description: https://help.waterdata.usgs.gov/faq/about-tab-delimited-output						
11	# Automated-retrieval info: https://help.waterdata.usgs.gov/faq/automated-retrievals						
12	#						
13	# Contact: gs-w_support_nwisweb@usgs.gov						
14	# retrieved: 2021-08-24 10:07:30 EDT (caww01)						
15	#						
16	# Data for the following 1 site(s) are contained in this file						
17	# USGS 04085427 MANITOWOC RIVER AT MANITOWOC, WI						
18	# ----------						
19	#						
20	# Data provided for site 04085427						
21	# TS parameter statistic Description						
22	# 152896 00060 00003 Discharge, cubic feet per second (Mean)						
23	#						
24	# Data-value qualification codes included in this output:						
25	#						
26	# A Approved for publication -- Processing and review completed.						
27	# P Provisional data subject to revision.						
28	# e Value has been estimated.						
29	#						
30	agency_cd	site_no	datetime	152896_0(	152896_00060_00003_cd		
31	5s	15s	20d	14n	10s		
32	USGS	4085427	7/26/1972	69	A		
33	USGS	4085427	7/27/1972	65	A		
34	USGS	4085427	7/28/1972	57	A		
35	USGS	4085427	7/29/1972	58	A		
36	USGS	4085427	7/30/1972	57	A		



USGS Data Retrieval using dataRetrieval Package

Important Packages in R

```
25  library(dataRetrieval)
26  library(tidyverse)
27  library(ggplot2)
```

dataRetrieval: Simplifies process of loading hydrologic data into R environment (DeCicco & Hirsch, USGS)

ggplot2*: Creates graphics; more flexible than base R graphics (Wickham and others)

tidyverse: Collection of R packages for data wrangling and data science (Wickham and others)

*Note: ggplot2 is a part of the tidyverse package and does not need to be separately loaded if tidyverse is loaded

Site Information: readNWISsite

```
46  # set station_no to USGS code - this example uses the USGS gage for the
    Manitowoc River at Manitowoc (04085427)
47  station_no <- "04085427"
48
49  # download information about the NWIS site
50  station_info <- dataRetrieval::readNWISsite(station_no)
```

dataRetrieval::readNWISsite(sites): Returns data about a selected site from NWIS web service

station_info: Saves the data into an object that has 42 variables

---

```
55  # check station name
56  station_info$station_nm
```

➡

```
> station_info$station_nm
[1] "MANITOWOC RIVER AT MANITOWOC, WI"
```

station_info$station_nm: Prints the name of the station

Data Availability: whatNWISdata

```
65  # identify what daily data are available for station
66  daily_data_availability <- dataRetrieval::whatNWISdata(siteNumber = station_no,
    service = "dv", statCd = "00003")
```

dataRetrieval::whatNWISdata(siteNumber, service, statCd): Imports a table of available parameters, period of record, and count

---

Use R code to manipulate data and create a summary table for available daily data (*note: 00060 is parameter code for discharge)*

site_no	station_nm	parm_cd	parameter_nm	begin_date	end_date	count_nu
04085427	MANITOWOC RIVER AT MANITOWOC, WI	00010	Temperature, water, degrees Celsius	2011-03-18	2021-08-22	2414
04085427	MANITOWOC RIVER AT MANITOWOC, WI	00060	Discharge, cubic feet per second	1972-07-26	2021-08-22	17498
04085427	MANITOWOC RIVER AT MANITOWOC, WI	00095	Specific conductance, water, unfiltered, microsiemens per ce..	2011-03-18	2021-08-22	2328
04085427	MANITOWOC RIVER AT MANITOWOC, WI	00300	Dissolved oxygen, water, unfiltered, milligrams per liter	2011-03-18	2021-08-22	2424
04085427	MANITOWOC RIVER AT MANITOWOC, WI	63680	Turbidity, water, unfiltered, monochrome near infra-red LED l..	2011-03-18	2021-08-22	1998

Daily Data: readNWISdv

```
121  #select parameter for discharge ("00060")
122  parameter_code <- "00060"
123
124  #load daily flow data for the site
125  discharge_data <- dataRetrieval::readNWISdv(siteNumbers = station_no,
126                                              parameterCd = parameter_code,
127                                              startDate = start_date,
128                                              endDate = end_date)
```

dataRetreival::readNWISdv(siteNumbers, parameterCd, startDate, endDate): Reads daily data for specified sites and parameters

- siteNumbers: Station ID for the site of interest
- ParameterCd: USGS parameter code (00060 for discharge)
- startDate: Beginning of period of record
- endDate: End of period of record

Daily Data Imported from readNWISdv

Table from **readNWISdv**

	agency_cd	site_no	Date	X_00060_00003	X_00060_00003_cd
1	USGS	04085427	1972-07-26	69	A
2	USGS	04085427	1972-07-27	65	A
3	USGS	04085427	1972-07-28	57	A
4	USGS	04085427	1972-07-29	58	A
5	USGS	04085427	1972-07-30	57	A

Tab-separated table from USGS website

```
#
# Data provided for site 04085427
#            TS   parameter     statistic     Description
#        152896       00060     00003     Discharge, cubic feet per second (Mean)
#
# Data-value qualification codes included in this output:
#
#     A  Approved for publication -- Processing and review completed.
#     P  Provisional data subject to revision.
#     e  Value has been estimated.
#
agency_cd       site_no datetime        152896_00060_00003      152896_00060_00003_cd
5s      15s     20d     14n     10s
USGS    04085427        1972-07-26      69.0    A
USGS    04085427        1972-07-27      65.0    A
USGS    04085427        1972-07-28      57.0    A
USGS    04085427        1972-07-29      58.0    A
USGS    04085427        1972-07-30      57.0    A
```

Additional Functions in dataRetrieval

Instantaneous Data: readNWISuv

dataRetreival::readNWISdv(siteNumbers, parameterCd, startDate, endDate, tz): Reads instantaneous data for specified sites and parameters

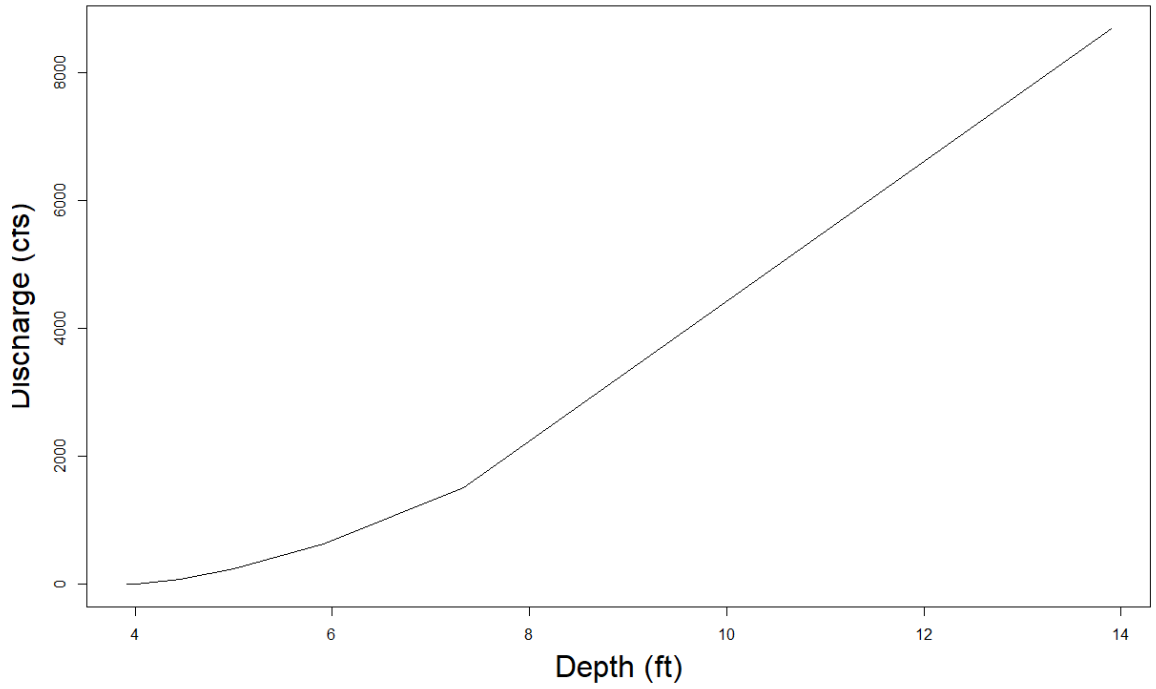
- tz: Time zone (“America/New_York, “America/Chicago”, etc.)

	agency_cd	site_no	dateTime	X_00060_00000	X_00060_00000_cd	tz_cd
1	USGS	04085427	2020-01-01 00:00:00	1500	A	America/Chicago
2	USGS	04085427	2020-01-01 00:05:00	1500	A	America/Chicago
3	USGS	04085427	2020-01-01 00:10:00	1500	A	America/Chicago
4	USGS	04085427	2020-01-01 00:15:00	1490	A	America/Chicago
5	USGS	04085427	2020-01-01 00:20:00	1500	A	America/Chicago
6	USGS	04085427	2020-01-01 00:25:00	1490	A	America/Chicago
7	USGS	04085427	2020-01-01 00:30:00	1490	A	America/Chicago
8	USGS	04085427	2020-01-01 00:35:00	1470	A	America/Chicago
9	USGS	04085427	2020-01-01 00:40:00	1480	A	America/Chicago

Rating Curve: readNWISrating

dataRetreival::readNWISrating(siteNumber): Provides rating curve (depth vs. discharge) for USGS gage

	INDEP	DEP	STOR
1	3.91	1.00	*
2	4.02	11.70	*
3	4.10	21.74	*
4	4.46	85.50	*
5	5.00	244.00	*
6	5.91	625.00	*
7	7.33	1512.00	*
8	13.90	8690.00	*

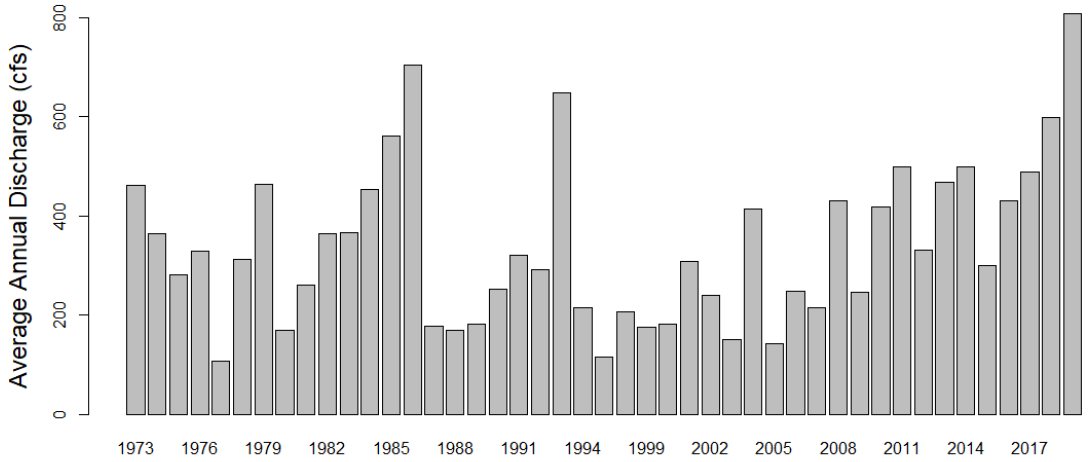


Summary Statistics: readNWISstat

dataRetreival::readNWISstat(siteNumbers, parameterCd, statReportType) : Provides summary statistics for specified periods

- statReportType: Type of report to evaluate (monthly, annual)

	agency_cd	site_no	parameter_cd	ts_id	loc_web_ds	year_nu	mean_va
1	USGS	04085427	00060	152896	*NA*	1973	462.5
2	USGS	04085427	00060	152896	*NA*	1974	365.4
3	USGS	04085427	00060	152896	*NA*	1975	281.8
4	USGS	04085427	00060	152896	*NA*	1976	329.9
5	USGS	04085427	00060	152896	*NA*	1977	108.0
6	USGS	04085427	00060	152896	*NA*	1978	311.6
7	USGS	04085427	00060	152896	*NA*	1979	463.0



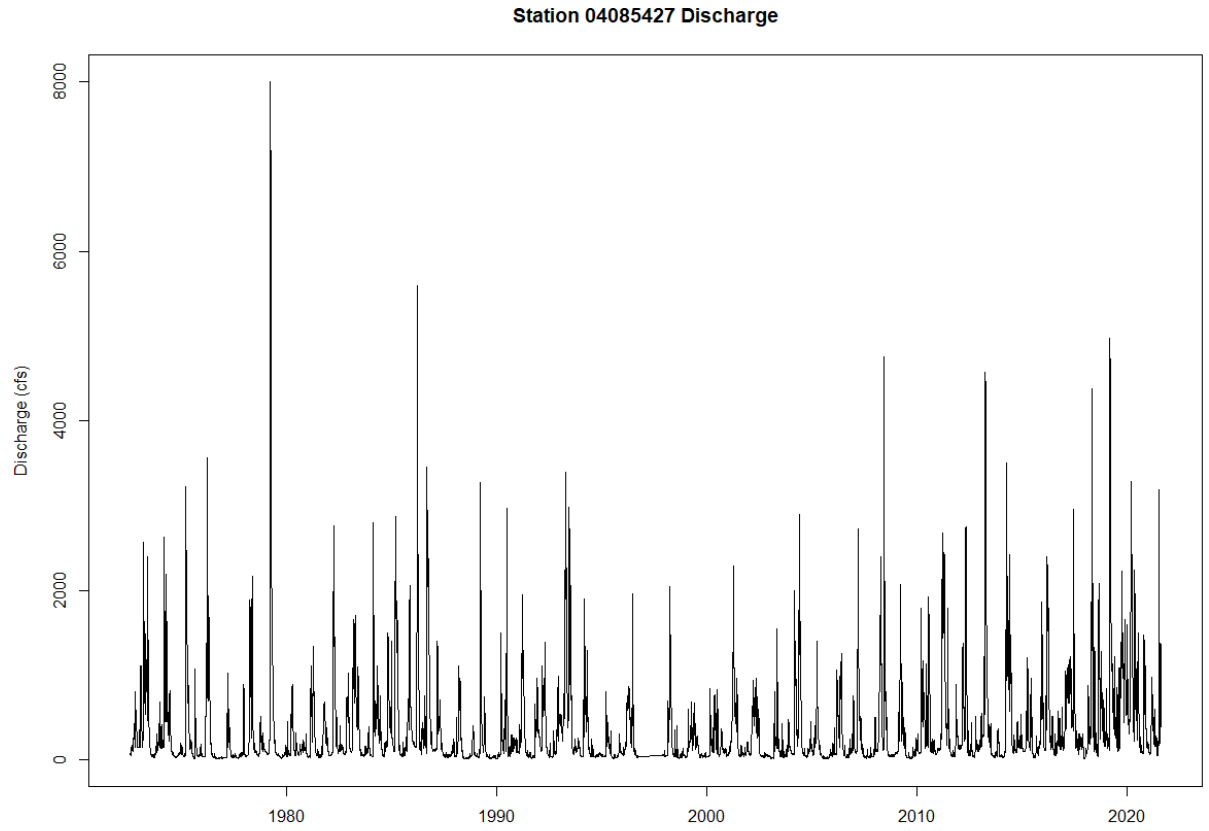
Additional dataRetrieval Functions

Visit dataRetrieval vignette in the Comprehensive R Archive Network (cran.r-project.org)

https://cran.r-project.org/web/packages/dataRetrieval/vignettes/dataRetrieval.html

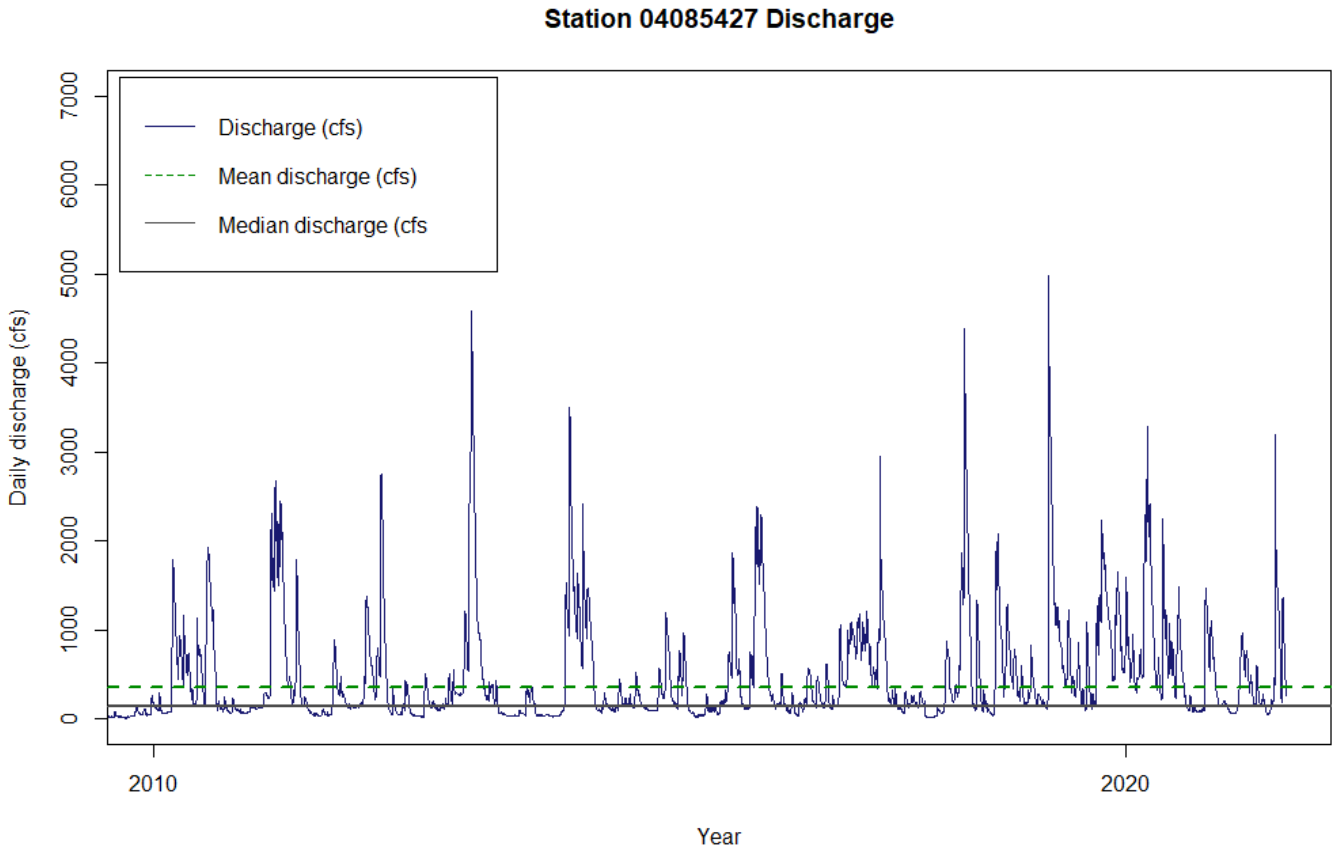
Basic Graphing of Hydrologic Data

Plot Discharge Using Base R Functions




```r
#use the base plot function to plot discharge data
plot(X_00060_00003 ~ Date,
     data = discharge_data,
     main = plot_title,
     xlab = element_blank(),
     ylab = "Discharge (cfs)",
     type = "l")
```

Plot Discharge Using Base R Functions




```r
plot(X_00060_00003 ~ Date,
     data = discharge_data,
     main = plot_title,
     xlab = "Year",
     ylab = "Daily discharge (cfs)",
     type = "l",
     col = "midnightblue",
     xlim = c(startDate, endDate),
     ylim = c(0,7000)
     )

# add lines for mean and median discharges
abline(h = median(discharge_data$X_00060_00003, na.rm = TRUE),
       col = "gray30",
       lwd = 2)
abline(h = mean(discharge_data$X_00060_00003, na.rm = TRUE),
       col = "green4",
       lty = "dashed",
       lwd = 2)

#add a legend
legend("topleft",
       legend = c("Discharge (cfs)", "Mean discharge (cfs)", "Me
       col = c("midnightblue", "green4", "gray30"),
       lty = 1:2,
       inset = 0.01)
```

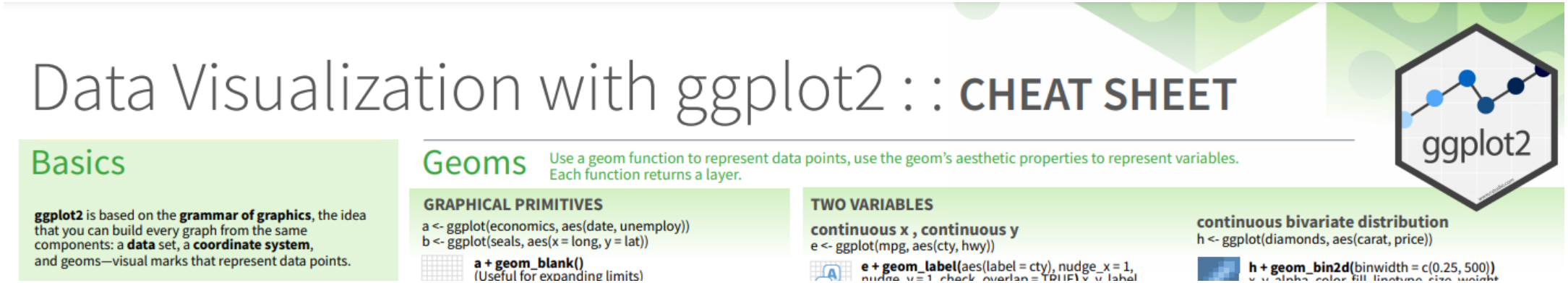
Plot Discharge Using ggplot2

```
library(ggplot2)
```
within
```
library(tidyverse)
```
https://ggplot2.tidyverse.org/

System for creating graphics

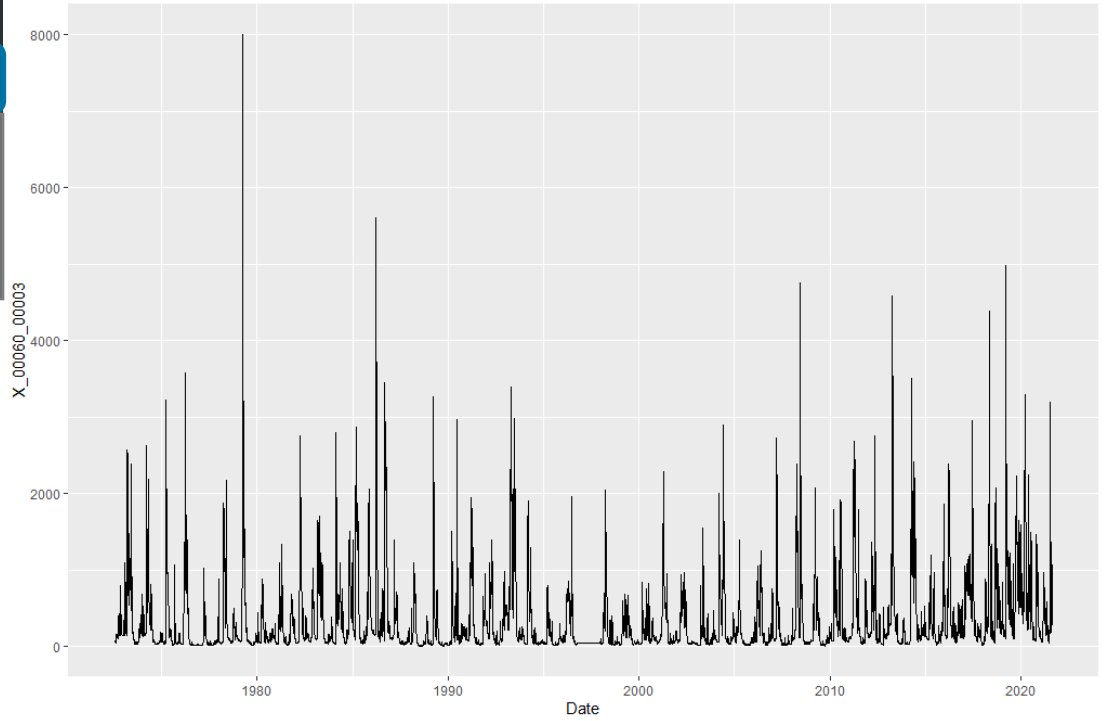
Based on “The Grammar of Graphics”

- Framework for concisely describing the components of graphics
- Layered approach using defined components to build a visualization



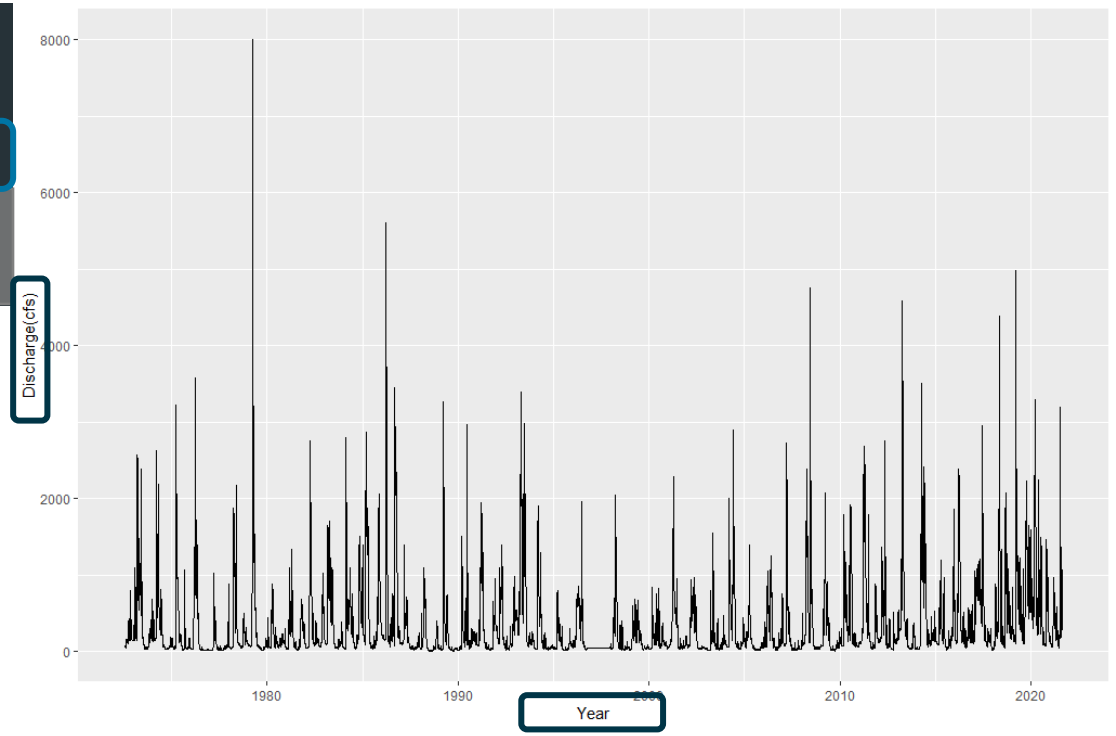
Plot Discharge Using ggplot

```
#plot the discharge data with ggplot
p <- ggplot(aes(Date, X_00060_00003), data = discharge_data) +
  geom_line() +
  xlab("Year") +
  ylab ("Discharge(cfs)") +
  xlim(c(as.Date("2010-01-01"), as.Date("2019-12-31"))) +
  ylim(c(0,6000)) +
  theme_set
```



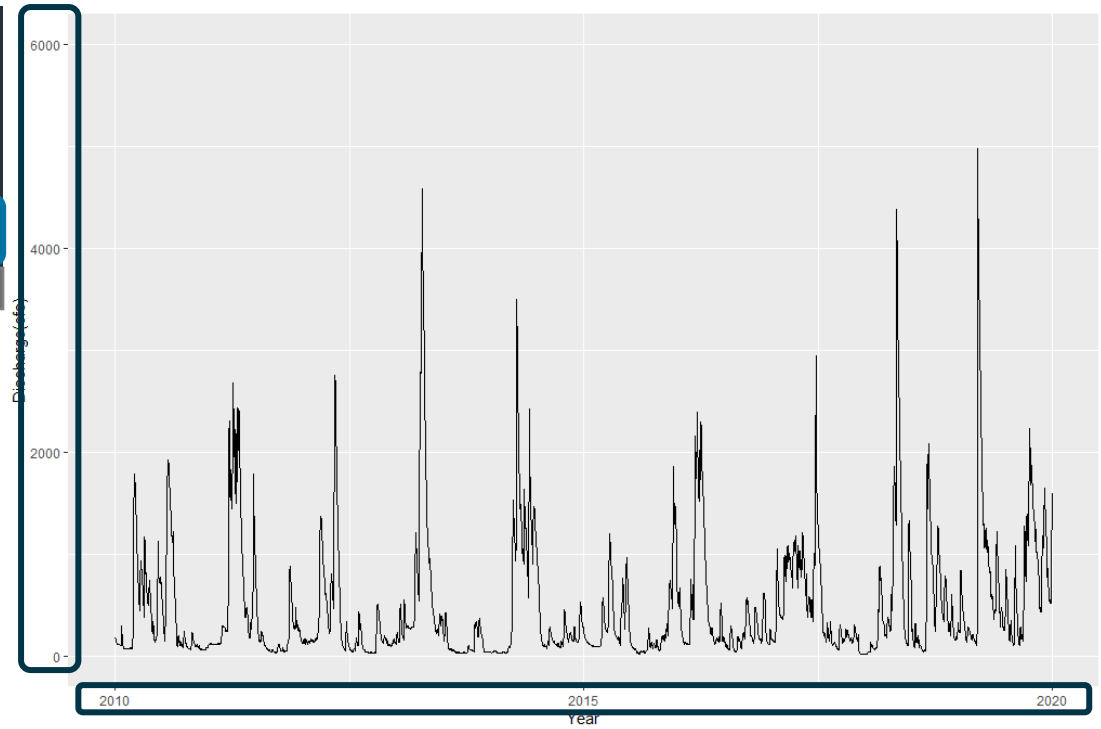
Plot Discharge Using ggplot: Axis Title

```
#plot the discharge data with ggplot
p <- ggplot(aes(Date, X_00060_00003), data = discharge_data) +
  geom_line() +
  xlab("Year") +
  ylab ("Discharge(cfs)") +
  xlim(c(as.Date("2010-01-01"), as.Date("2019-12-31"))) +
  ylim(c(0,6000)) +
  theme_set
```



Plot Discharge Using ggplot: Plot Range

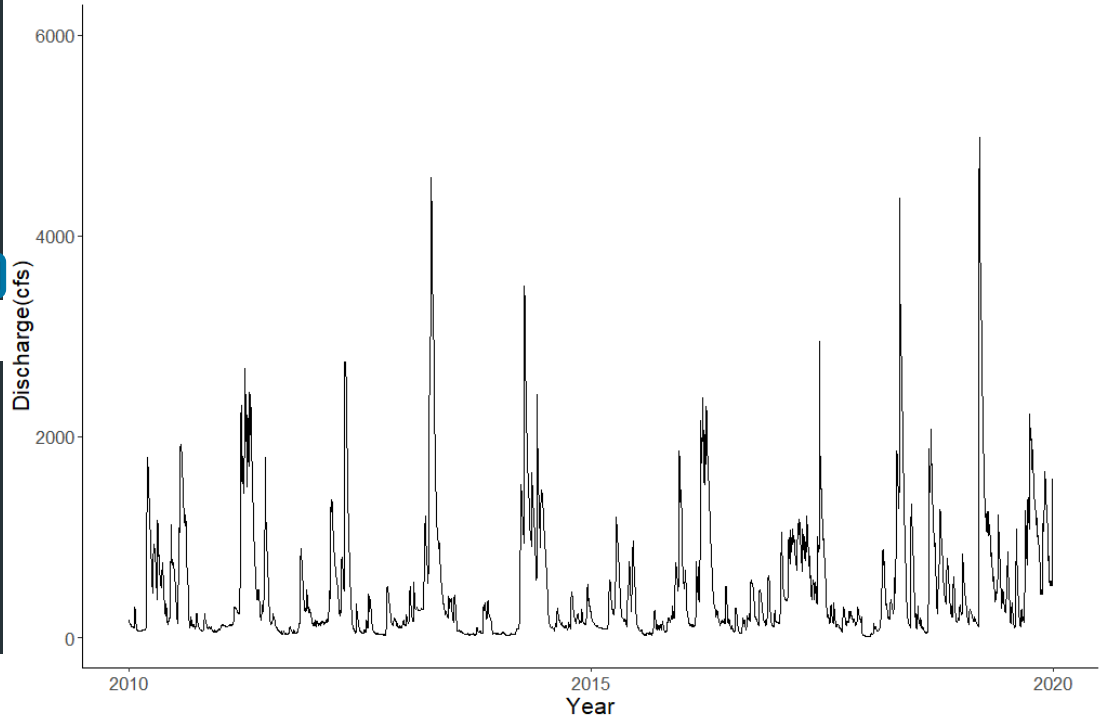
```
#plot the discharge data with ggplot
p <- ggplot(aes(Date, X_00060_00003), data = discharge_data) +
  geom_line() +
  xlab("Year") +
  ylab ("Discharge(cfs)") +
  xlim(c(as.Date("2010-01-01"), as.Date("2019-12-31"))) +
  ylim(c(0,6000)) +
  theme_set
```



Plot Discharge Using ggplot: Formatting

```
#plot the discharge data with ggplot
p <- ggplot(aes(Date, X_00060_00003), data = discharge_data) +
  geom_line() +
  xlab("Year") +
  ylab ("Discharge(cfs)") +
  xlim(c(as.Date("2010-01-01"), as.Date("2019-12-31"))) +
  ylim(c(0,6000)) +
  theme_set
```

```
#set a new theme for the figure
theme_set <-   theme_bw() +
  theme(axis.line = element_line(color = 'black'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),
        text = element_text(size = 16))
```



Discharge Data Manipulation: dplyr

```
library(dplyr)
```
within
```
library(tidyverse)
```

https://dplyr.tidyverse.org/

System for manipulating data

Essentially “The Grammar of Data Manipulation”

- Provides consistent set of ‘verbs’ to assist in data manipulation and transformation

Discharge Data Manipulation: dplyr

1. Add columns with day, month, and year

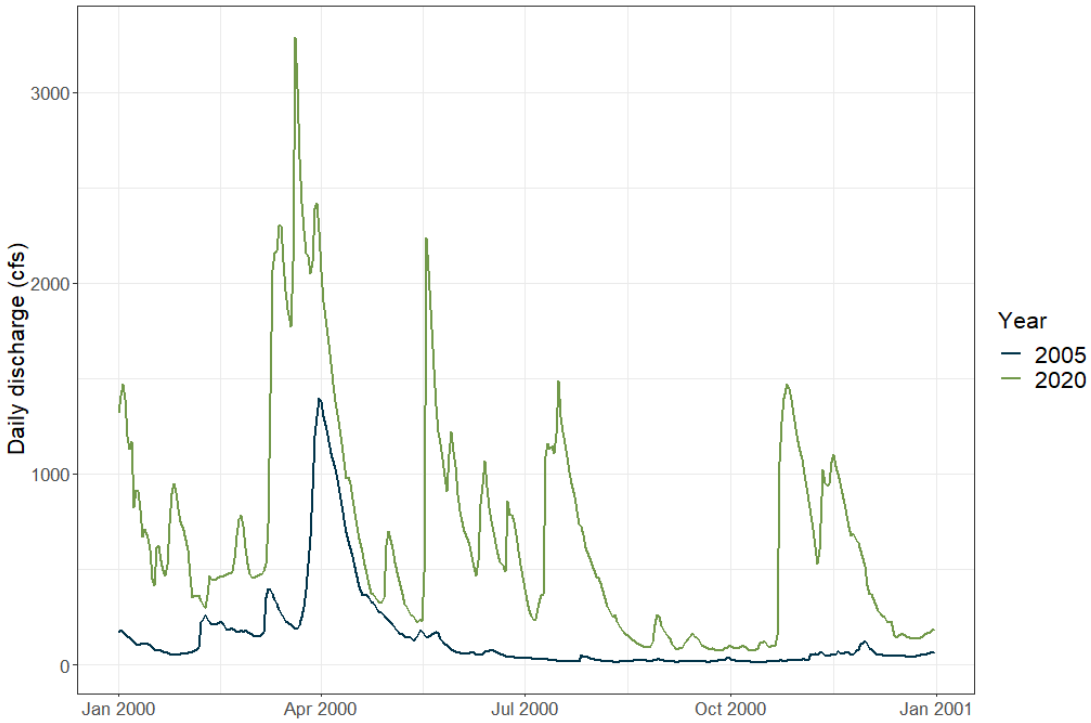
```
#add year, month, and day columns to the discharge data
discharge_date_adj <- discharge_data %>%
  mutate(d = day(Date),
         mo = month(Date),
         yr = year(Date),
         discharge = X_00060_00003) %>%
  select(yr, mo, d, discharge)
```

2. Filter for only 2005 data

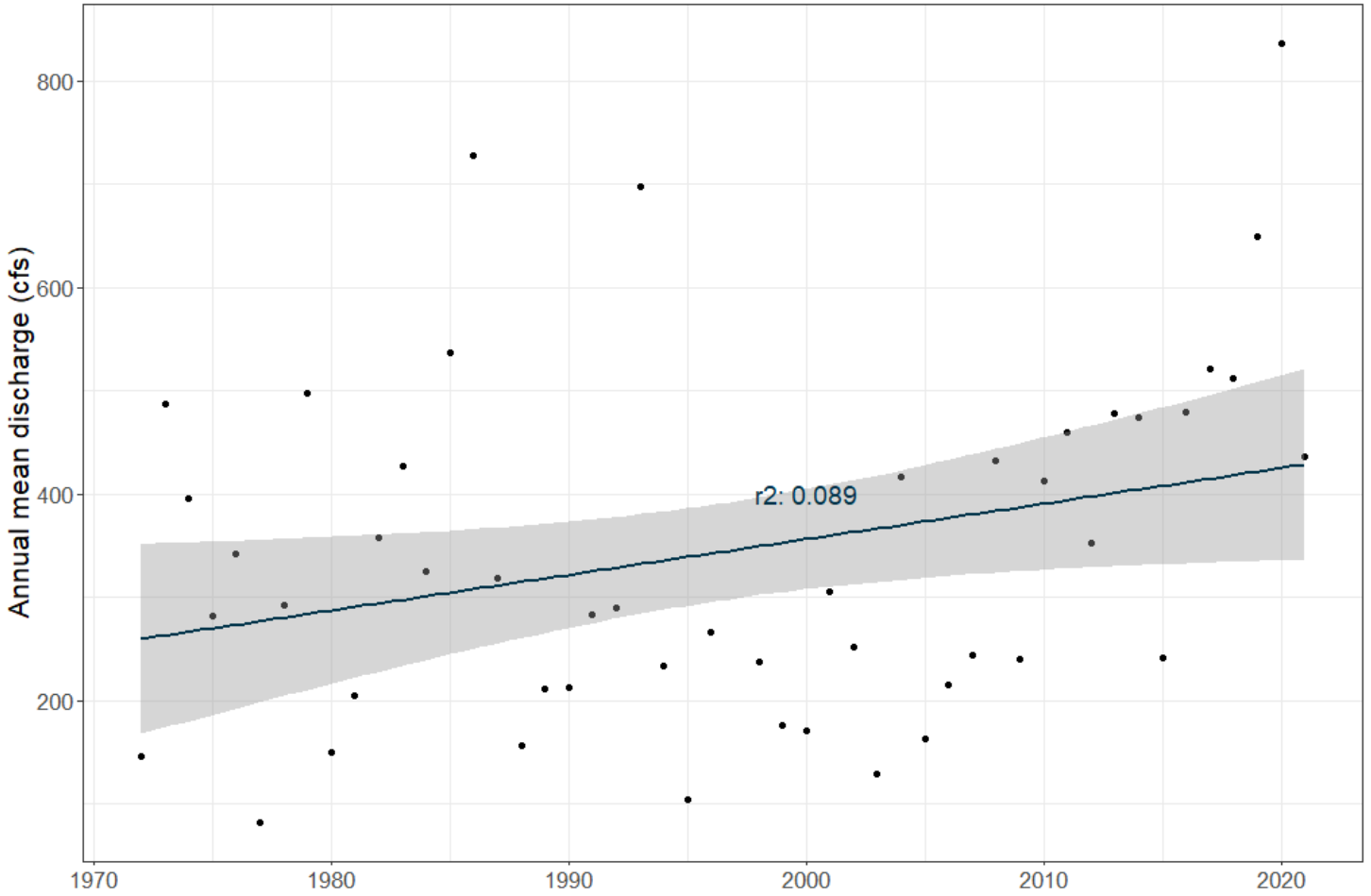
```
#filter discharge data for the first year of the input function
yr1 <- discharge_date_adj %>%
  filter(yr == 2005) %>%
  mutate(adj_date = make_date(year = 2000, month = mo, day = d)) %>%
  mutate(yr = as.character(yr)) %>%
  select(yr, adj_date, discharge)
```

3. Filter for only 2020 data

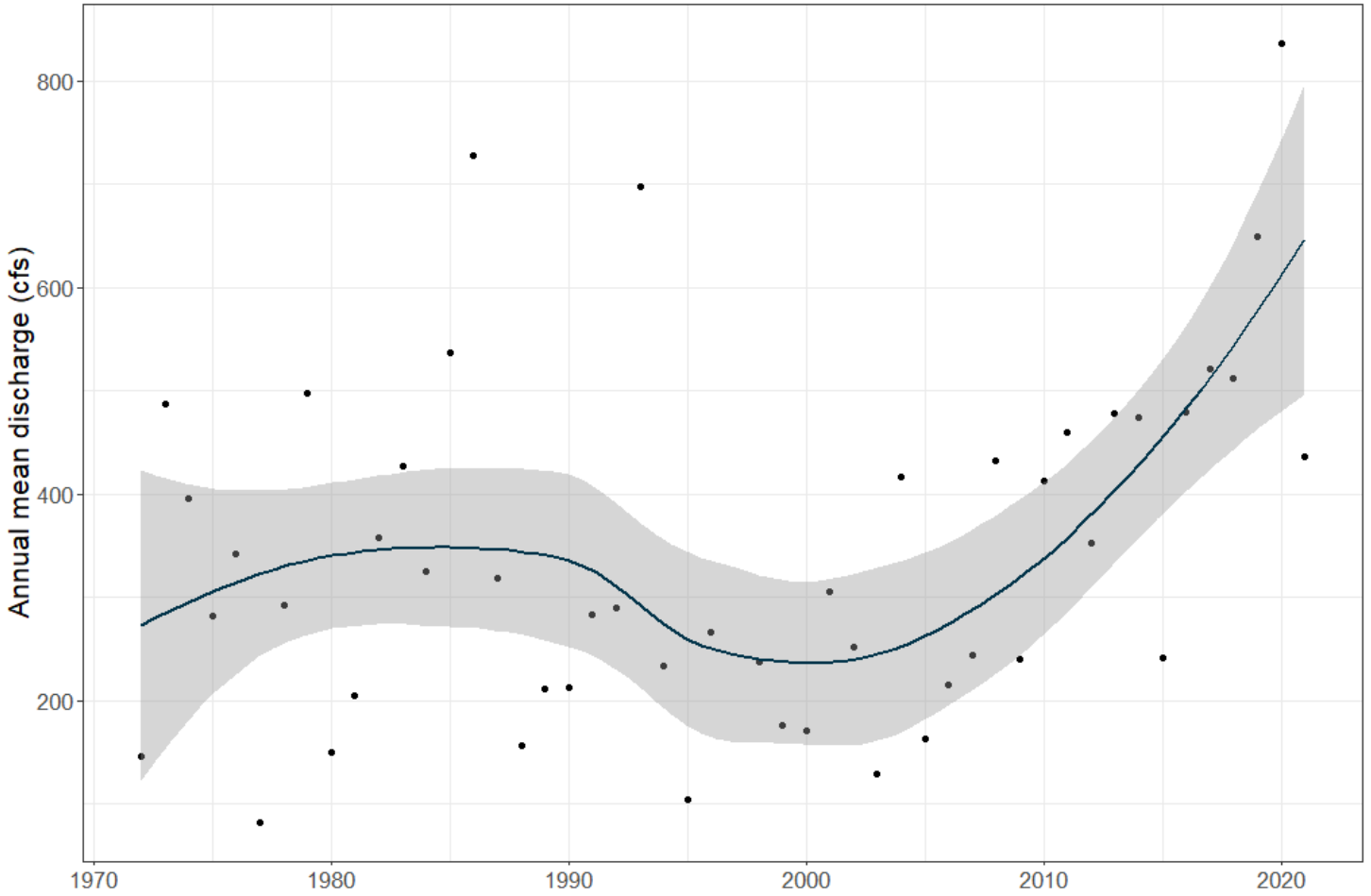
```
#filter discharge data for the second year of the input function
yr2 <- discharge_date_adj %>%
  filter(yr == 2020) %>%
  mutate(adj_date = make_date(year = 2000, month = mo, day = d)) %>%
  mutate(yr = as.character(yr)) %>%
  select(yr, adj_date, discharge)
```



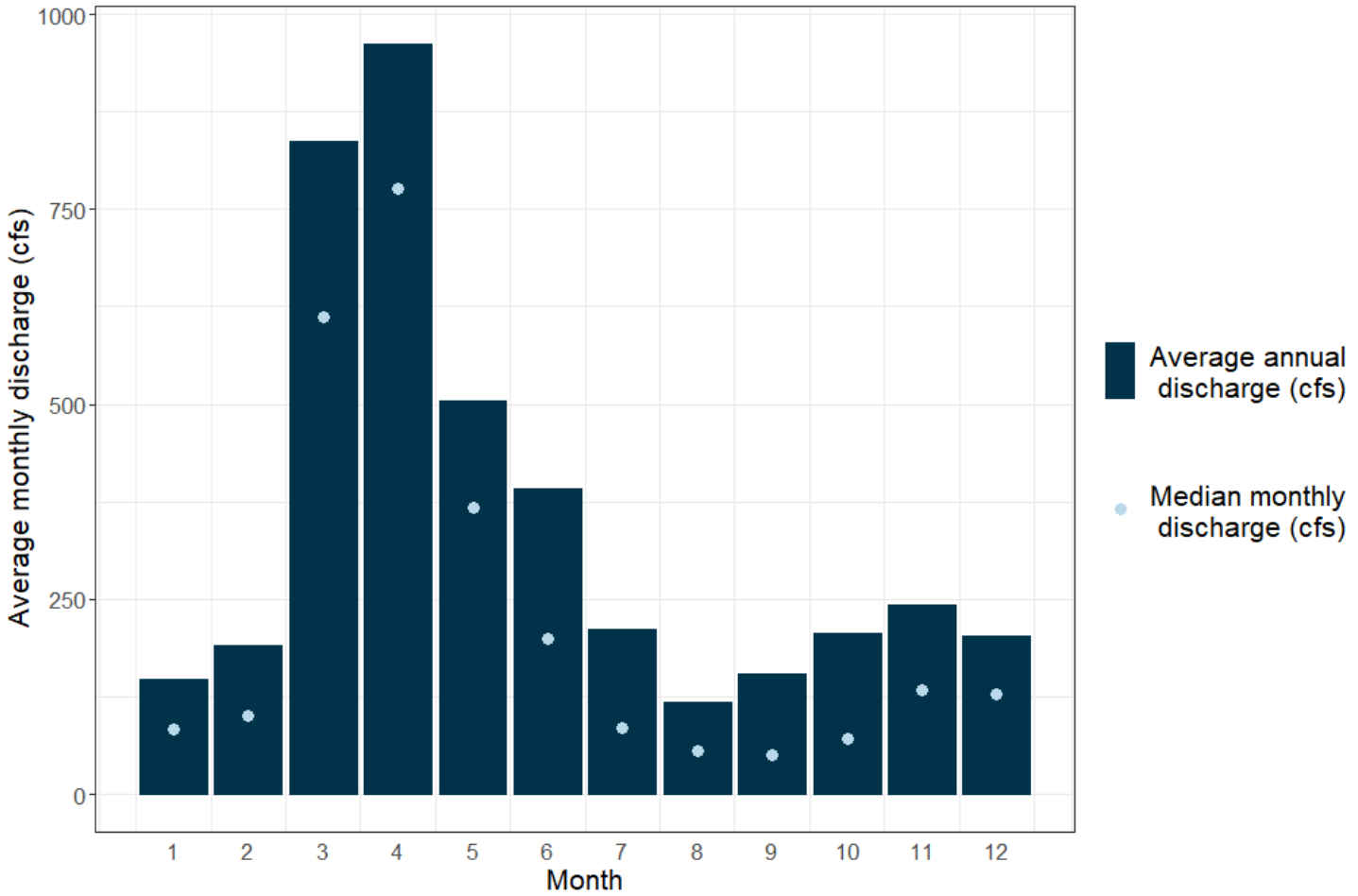
Display Average Annual Flows: Linear Reg.



Display Average Annual Flows: Smoothed



Display Average Monthly Flows



Creating Flow Duration Curves

Flow Duration Step 1: Sort Discharge Data

```r
#set a new variable equal to all discharges from discharge data
fl_dur_data <- discharge_data %>%
  select(X_00060_00003)

#sort the discharge data at the site by decreasing order
fl_dur_data <- as.data.frame(sort(fl_dur_data$X_00060_00003,
                                  decreasing = TRUE))

#change the name of the column from X_00060_00003 to discharge
colnames(fl_dur_data) <- c("discharge")
```

discharge_data

	agency_cd	site_no	Date	X_00060_00003	X_00060_00003_cd
1	USGS	04085427	1972-07-26	69	A
2	USGS	04085427	1972-07-27	65	A
3	USGS	04085427	1972-07-28	57	A
4	USGS	04085427	1972-07-29	58	A

fl_dur_data

	discharge
1	8000
2	6400
3	6200
4	5600
5	5600

Flow Duration Step 2: Calculate Exceedance

```r
#count the number of rows in the fl_dur_data dataframe
fd_rows <- nrow(fl_dur_data)

#rank the flows and calculate an exceedance probablity
fl_dur_data <- fl_dur_data %>%
  mutate(ranked = 1:fd_rows) %>%
  mutate(exceed_prob = ranked/fd_rows)
```

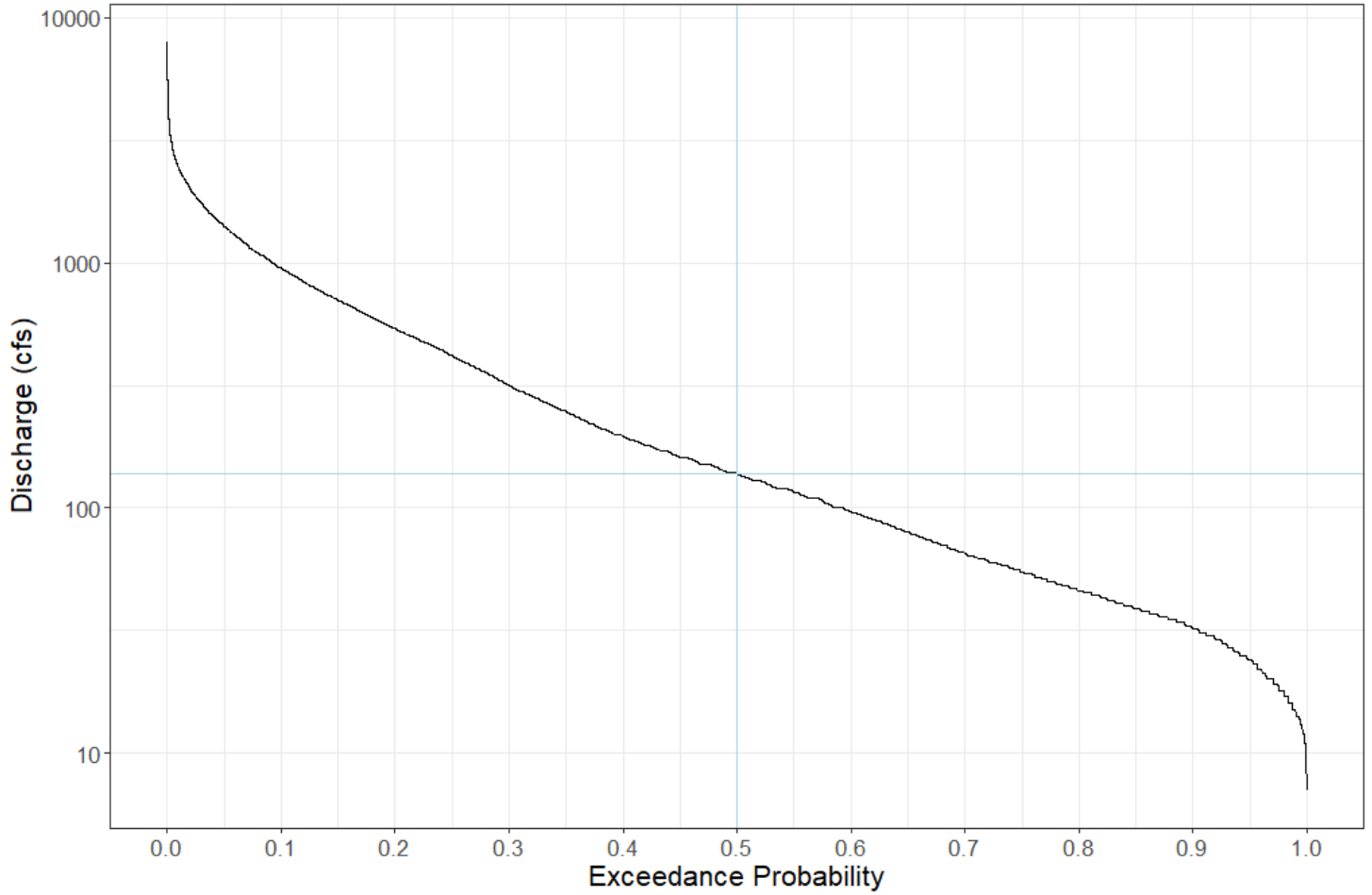
fl_dur_data

	discharge
1	8000
2	6400
3	6200
4	5600
5	5600

fl_dur_data

	discharge	ranked	exceed_prob
1	8000	1	5.714612e-05
2	6400	2	1.142922e-04
3	6200	3	1.714384e-04
4	5600	4	2.285845e-04
5	5600	5	2.857306e-04

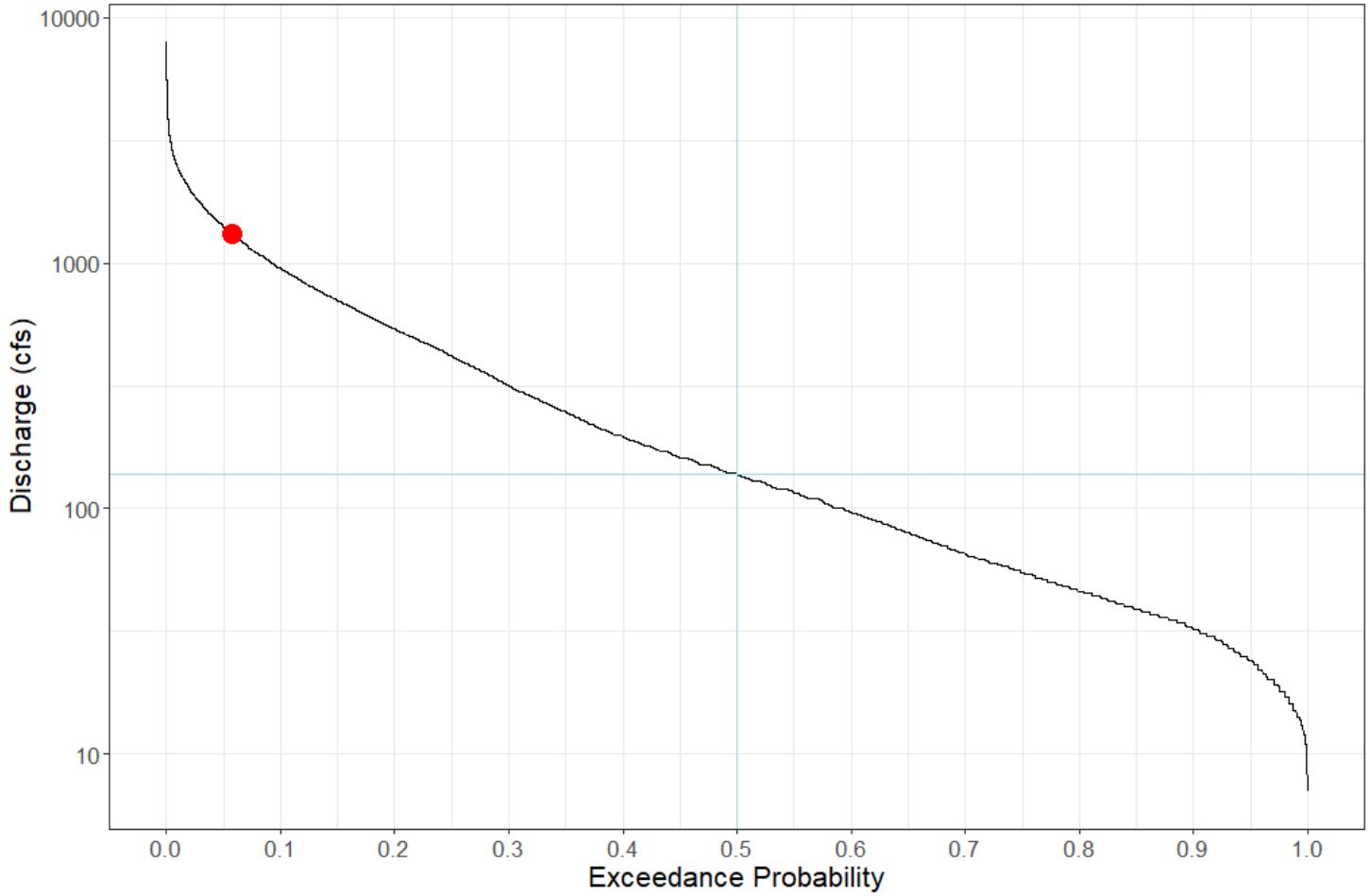
Flow Duration Step 3: Plot Curve



Flow Duration: Find Exceedance on Date

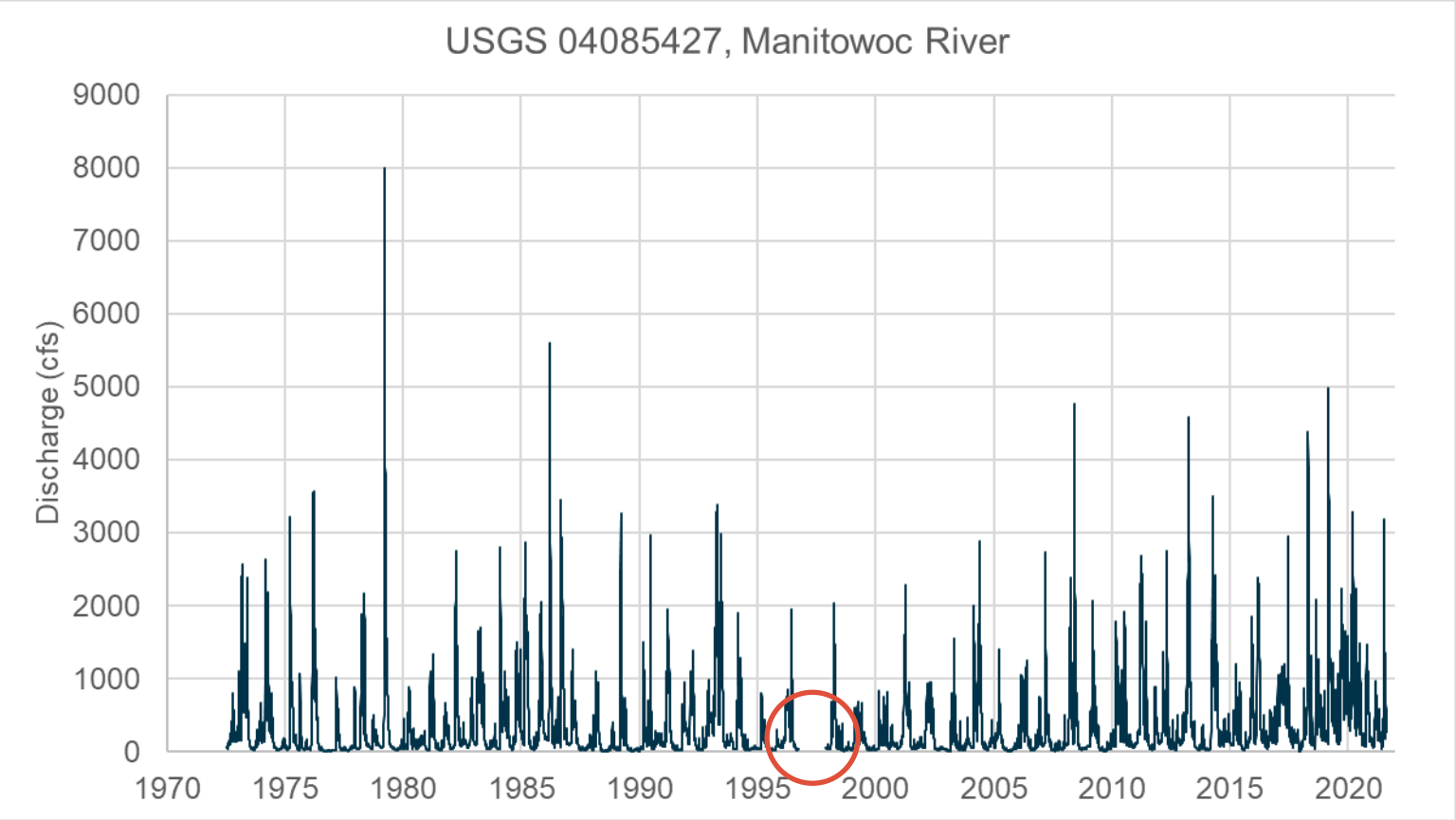
	Date	discharge
1	2020-01-01	1320

	exceed_prob	discharge
1	0.05708898	1320



Performing Baseflow Separation

Identify Missing Data: Visual Inspection



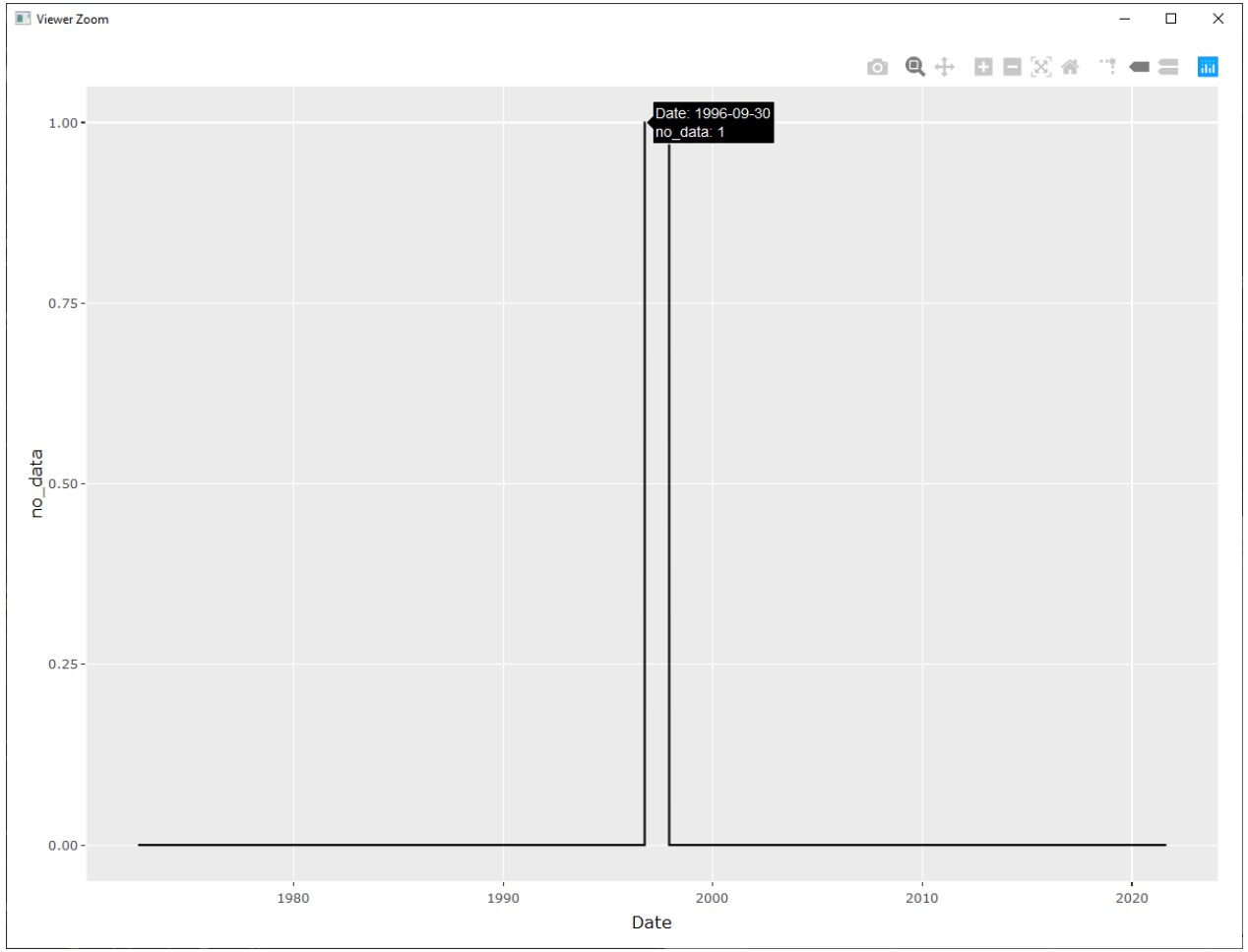
Identify Missing Data: Evaluation

Identify missing data using dplyr functions

Use plotly library for interactive exploration of data (plotly::ggplotly)

Evaluate results:

Data from 1996-09-30 and 1998-11-30 are missing



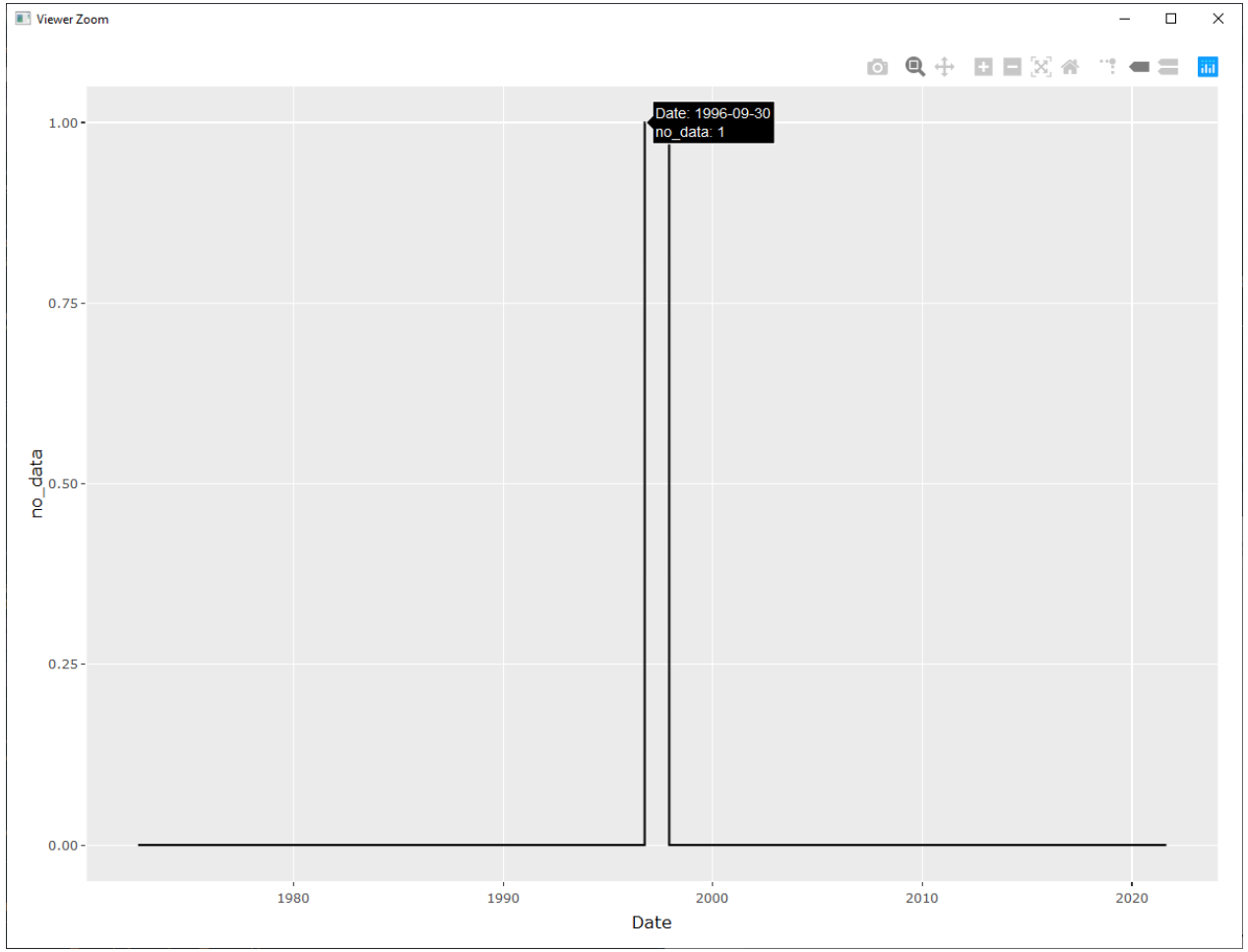
Identify Missing Data

Identify missing data using dplyr functions

Use plotly library for interactive exploration of data (plotly::ggplotly)

Evaluate results:

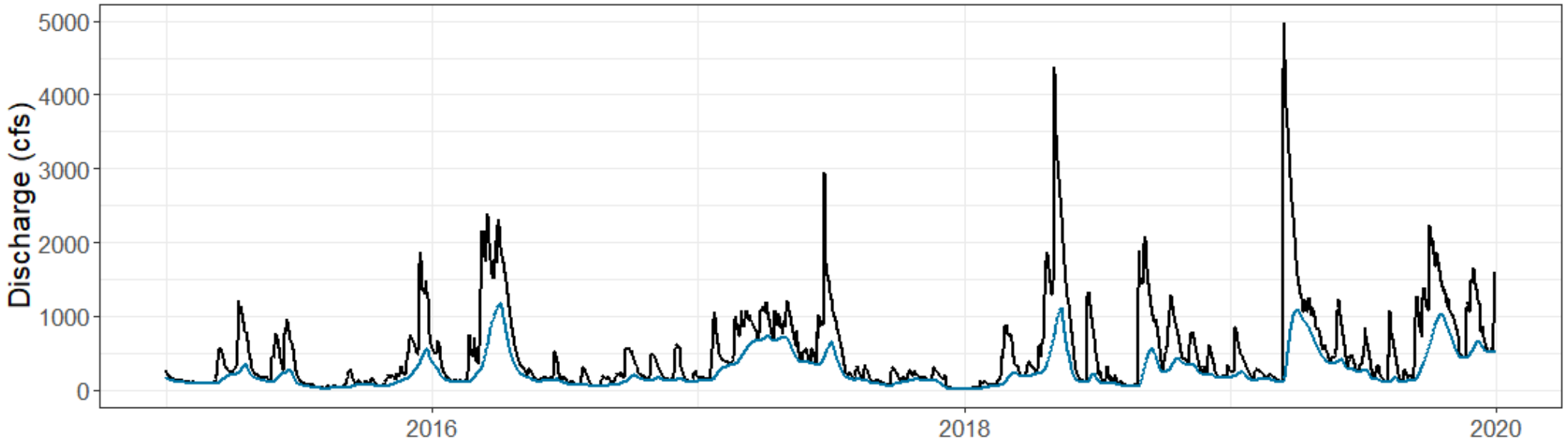
Data from 1996-09-30 and 1998-11-30 are missing



Baseflow Separation: EcoHydRology

EcoHydRology::BaseflowSeparation: Uses recursive digital filter (signal processing) to estimate baseflow (Nathan & McMahon, 1990)

Signal processing: Separates high-frequency events (runoff) from low-frequency events (baseflow)



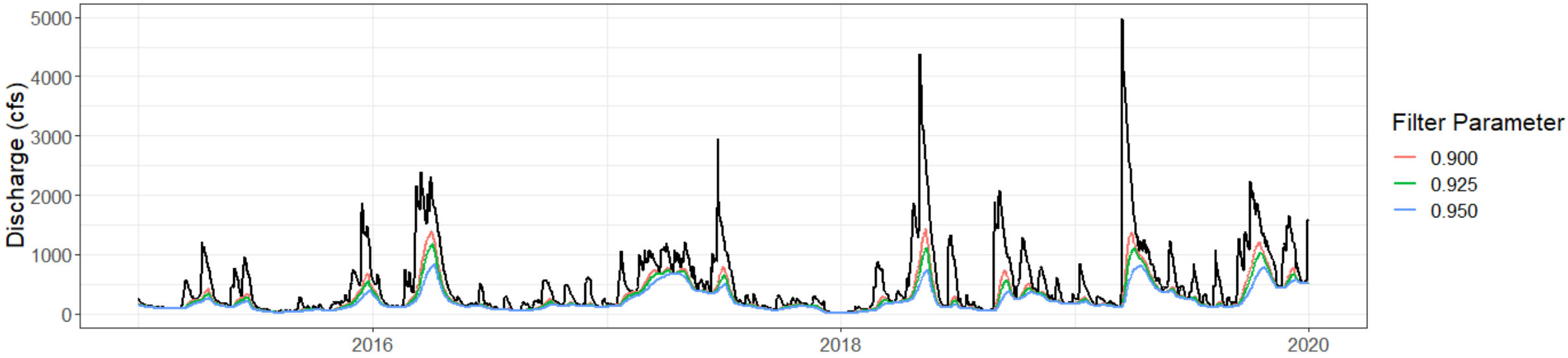
Baseflow Separation: EcoHydRology

EcoHydRology::BaseflowSeparation(streamflow, filter_parameter, passes)

streamflow: Vector of streamflow values (1 column)

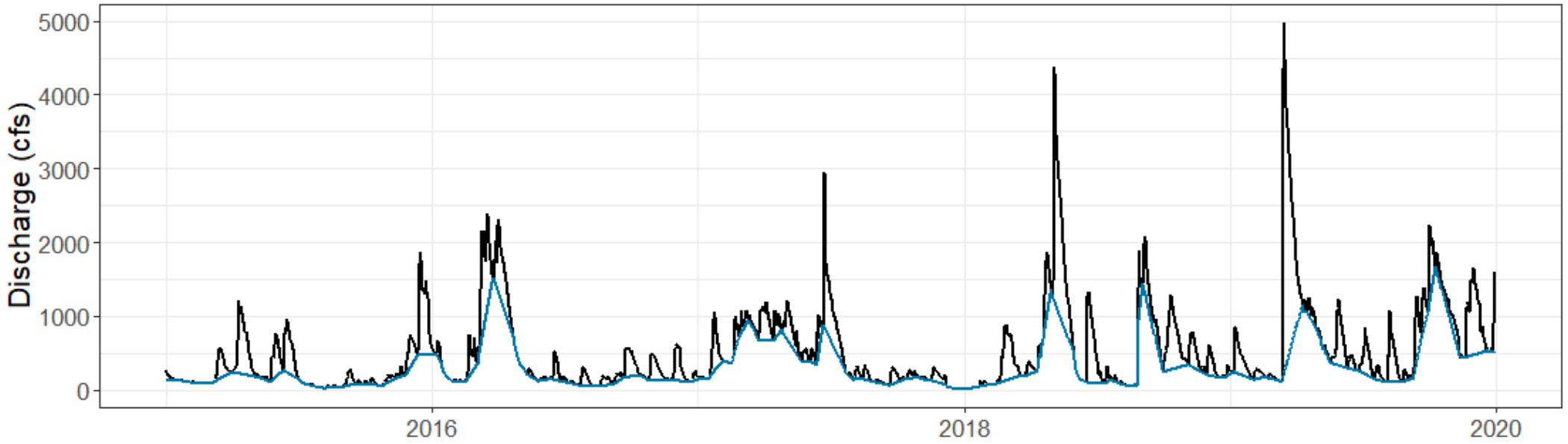
filter_parameter: Filter Parameter; ranges from 0.900 – 0.95; 0.925 is default

passes: Number of passes through data (forward – back – forward); 3 is default



Baseflow Separation: lfstat

lfstat::baseflow: Uses smoothed minima method to estimate baseflow for specified recession period (typically 5 days)



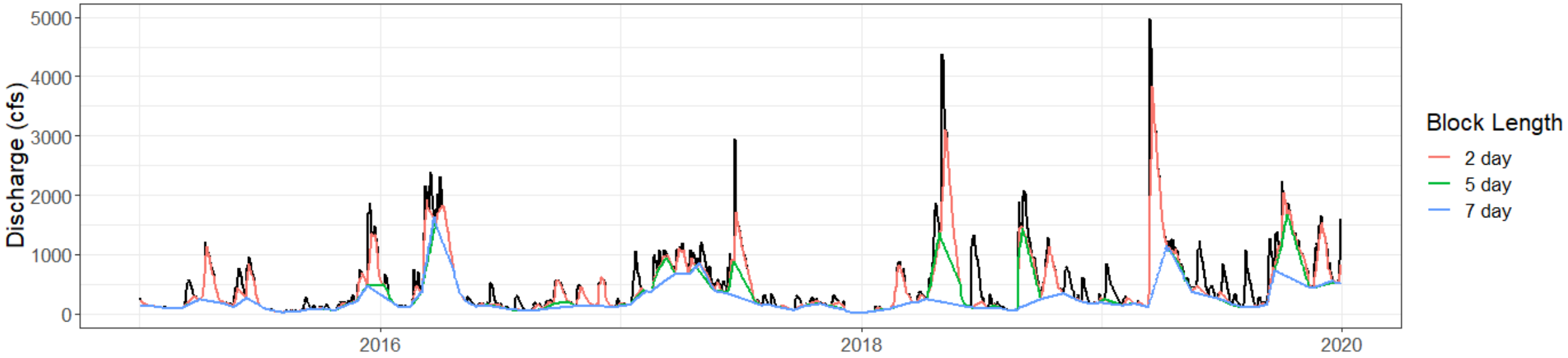
Baseflow Separation: lfstat

lfstat::baseflow(x, tp.factor, block.len)

x: Vector of streamflow values (1 column)

tp.factor: Turning point factor in days; 0.90 is default

block.length: Block length in days; 5 days is default, but depends on watershed



Baseflow Index

Baseflow Index refers to the fraction of total flow that is baseflow

$$\text{Baseflow Index} = \frac{\sum Baseflow}{\sum Total\ flow}$$

BFI Manitowoc River: 0.41 - 0.50

*Baseflow contributes ~41% - 50% of total flow in the river

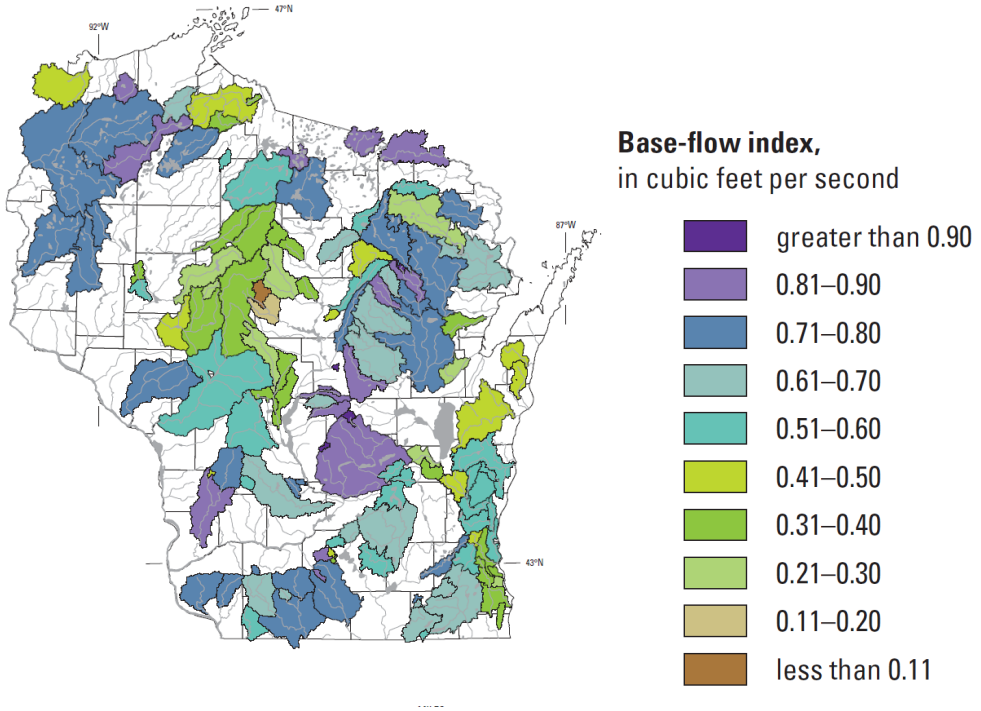


Figure Source: Gebert, W.A., Walker, J.F, and Hut, R.J., 2009, Groundwater Recharge in Wisconsin – Annual Estimates from 1970-99: United States Geological Survey Fact Sheet 2009-3092, 4 p.

Baseflow Index: Choosing inputs

Digital Filter

EcoHydRology package

filter_parameter adjusted

Filter Parameter	bfi
0.900	0.531
0.925	0.474
0.950	0.398

Default: 0.925

Smoothed Minima

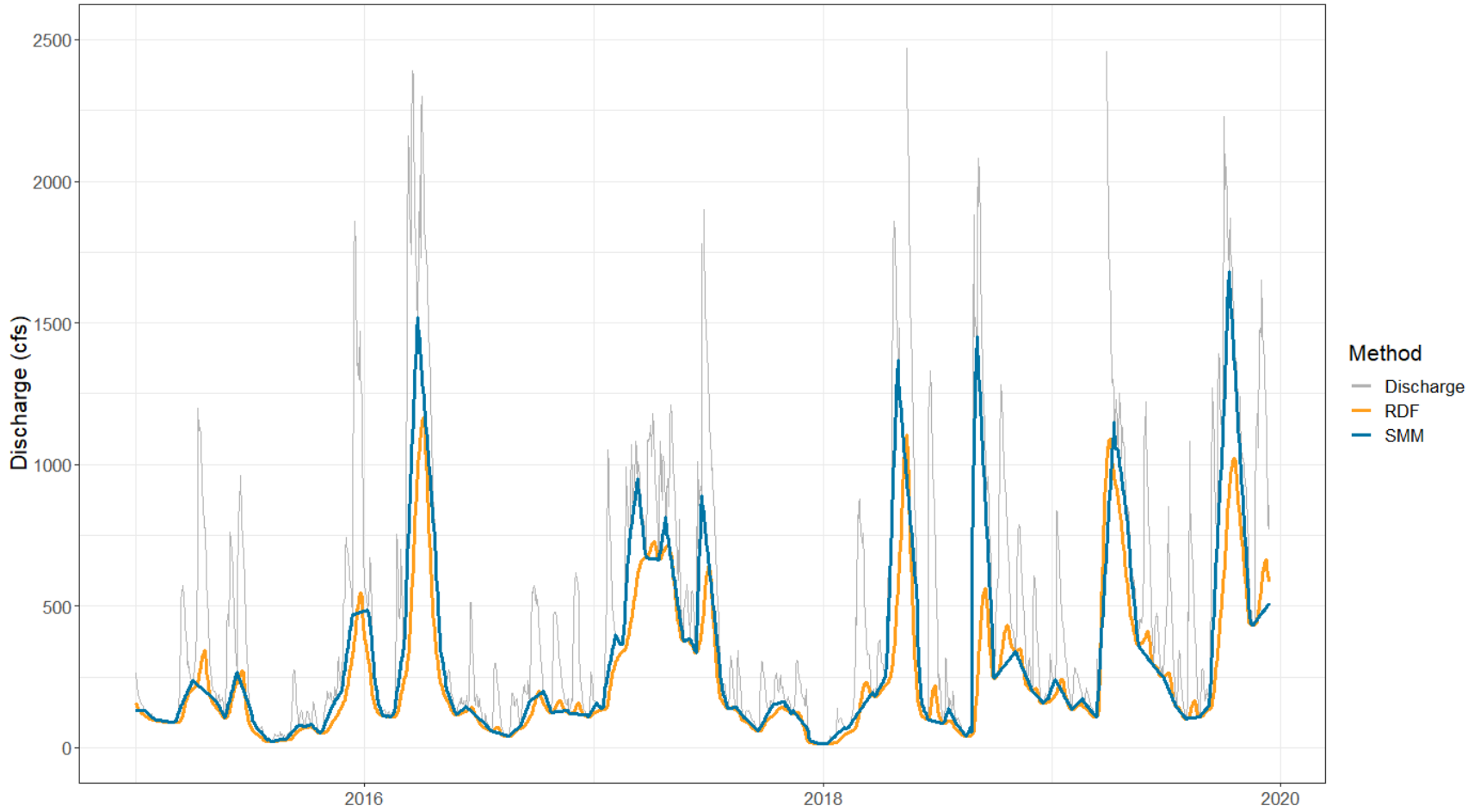
lfstat package

block.len adjusted

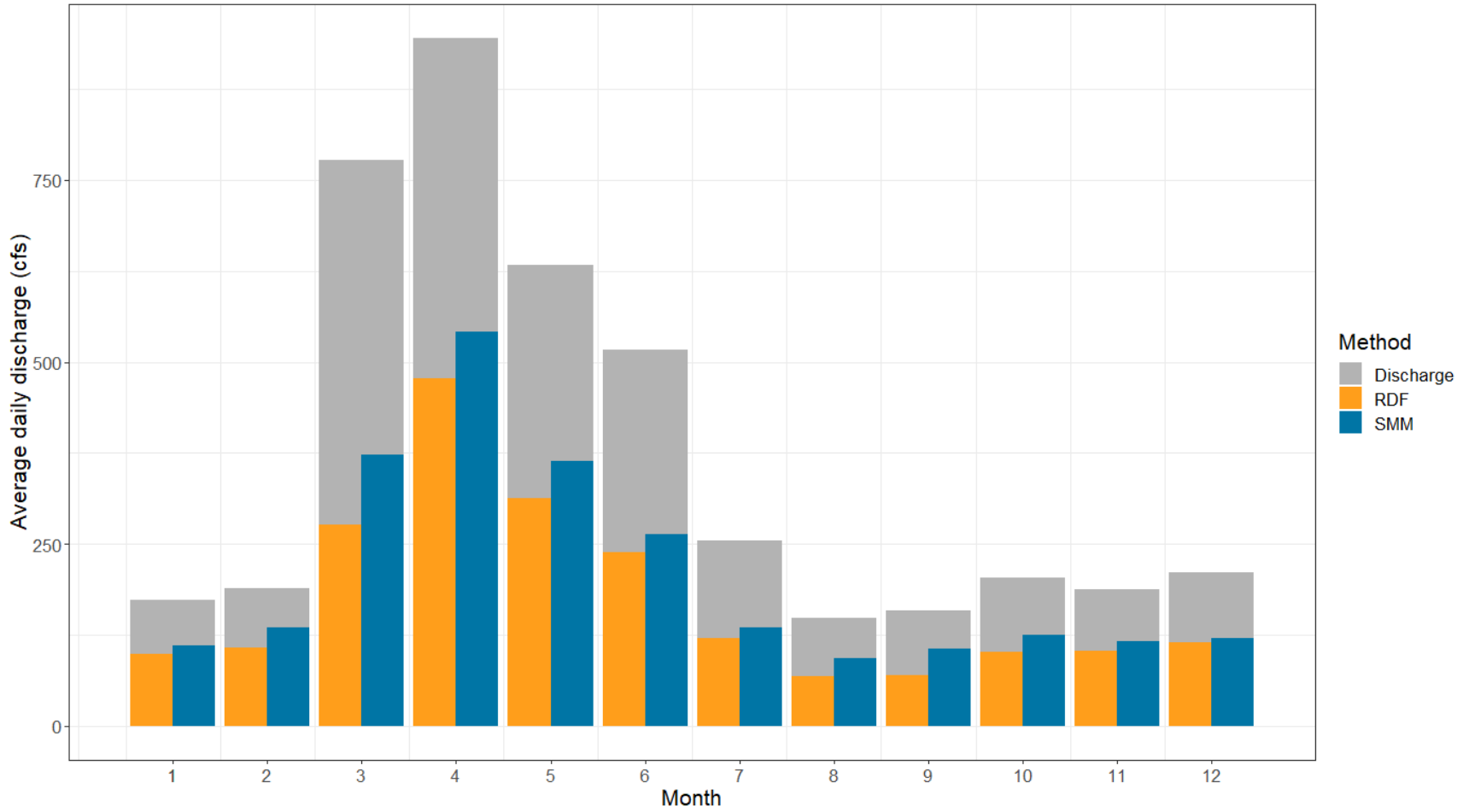
Block Length	bfi
2 day	0.839
5 day	0.562
7 day	0.49

Default: 5 days

Method Comparison: Daily Discharge



Method Comparison: Monthly Average



CONNECT WITH US

Eric Hettler

eric.hettler@wisconsin.gov

/WIDNR

@WIDNR

@WI_DNR

/WIDNRTV

"WILD WISCONSIN:
OFF THE RECORD"