Introduction to the EGRET package

Robert Hirsch[1] and Laura De Cicco[1]

[1] *United States Geological Survey*

June 3, 2013

Contents

1 Introduction to Exploration and Graphics for RivEr Trends (EGRET)

For information on getting started in R, downloading and installing the package, see Appendix 1: (A).

Exploration and Graphics for RivEr Trends (EGRET): An R-package for the analysis of long-term changes in water quality and streamflow. EGRET includes statistics and graphics for streamflow history, water quality trends, and the modeling algorithm Weighted Regressions on Time, Discharge, and Season (WRTDS).

Please see the official EGRET manual: (link to download) **for more information on the EGRET package.**

The best way to learn about the WRTDS approach and to see examples of its application to multiple large data sets is to read two journal articles. Both are available, for free, from the journals in which they were published.

The first relates to nitrate and total phosphorus data for 9 rivers draining to Chesapeake Bay. The URL is (2): `http://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.2010.00482.x/full`

The second is an application to nitrate data for 8 monitoring sites on the Mississippi River or

its major tributaries (3). The URL is: http://pubs.acs.org/doi/abs/10.1021/es201221s

This vignette assumes that the user understands the concepts underlying WRTDS. Thus, reading at least the first of these papers is necessary for understanding. The method has been enhanced beyond what was published there. The enhancement is that it now properly handles censored data by using survival regression rather than ordinary regression. The details of that are in a manuscript currently in process by Doug Moyer and Bob Hirsch.

This vignette will walk through the major functions provided by the EGRET package. The package dataRetrieval is required for importing data in a EGRET-friendly format. The dataRetrieval package, along with installation instructions can be found at:
https://github.com/USGS-R/dataRetrieval
Installing dataRetrieval will provide a vignette similar to this document, with complete working examples of the main dataRetrieval functions.

This vignette is divided into four sections: EGRET Dataframes, Flow History, WRTDS Analysis, and WRTDS Results. This document assumes the reader is familiar with the dataRetrieval package. The examples will follow an analysis of nitrate on the Choptank River at Greensboro, MD. Further details can be found in the user guide that can be found on gitHub:
https://github.com/USGS-R/EGRET/raw/Documentation/EGRET%2Bmanual_4.doc

2 EGRET Dataframes and Units

The EGRET package uses 3 default dataframes throughout the calculations, analysis, and graphing. These dataframes are Daily (2.1), Sample (2.2), and INFO (2.3). EGRET uses entirely SI units to store the data, but for purposes of output, it can report results in a wide variety of units, which will be discussed in (2.4). To start our exploration, the packages must be installed (check the appendix for detailed instructions (A)), then opened:

```
> library(dataRetrieval)
> library(EGRET)
```

2.1 Daily

The Daily dataframe initially is populated with columns generated by the dataRetrieval package (Table 1). After running the WRTDS calculations (as will be described in 5), additional columns are inserted (Table 2).

Table 1: Daily dataframe

ColumnName	Type	Description	Units
Date	Date	Date	date
Q	number	Discharge in cms	cms
Julian	number	Number of days since January 1, 1850	days
Month	integer	Month of the year [1-12]	months
Day	integer	Day of the year [1-366]	days
DecYear	number	Decimal year	years
MonthSeq	integer	Number of months since January 1, 1850	months
Qualifier	string	Qualifing code	character
i	integer	Index	days
LogQ	number	Natural logarithm of Q	numeric
Q7	number	7 day running average of Q	cms
Q30	number	30 running average of Q	cms

Table 2: Daily dataframe, post-WRTDS

ColumnName	Type	Description	Units
yHat	number	The WRTDS estimate of the log of concentration	numeric
SE	number	The WRTDS estimate of the standard error of yHat	numeric
ConcDay	number	The WRTDS estimate of concentration	mg/L
FluxDay	number	The WRTDS estimate of flux	kg/day
FNConc	number	Flow normalized estimate of concentration	mg/L
FNFlux	number	Flow Normalized estimate of flux	kg/day

2.2 Sample

The Sample dataframe initially is populated with columns generated by the dataRetrieval package (Table 3). After running the WRTDS calculations (as will be described in 5), additional columns are inserted (Table 4):

Table 3: Sample dataframe

ColumnName	Type	Description	Units
Date	Date	Date	date
ConcLow	number	Lower limit of concentration	mg/L
ConcHigh	number	Upper limit of concentration	mg/L
Uncen	integer	Uncensored data (1=true, 0=false)	integer
ConcAve	number	Average concentration	mg/L
Julian	number	Number of days since January 1, 1850	days
Month	integer	Month of the year [1-12]	months
Day	integer	Day of the year [1-366]	days
DecYear	number	Decimal year	years
MonthSeq	integer	Number of months since January 1, 1850	months
SinDY	number	Sine of DecYear	numeric
CosDY	number	Cosine of DecYear	numeric
Q [a]	number	Discharge	cms
LogQ [b]	number	Natural logarithm of flow	numeric

[a] Populated after calling mergeReport
[b] Populated after calling mergeReport

Table 4: Sample dataframe, post-WRTDS

ColumnName	Type	Description	Units
yHat	number	jack-knife estimate of the log of concentration	numeric
SE	number	jack-knife estimate of the standard error of yHat	numeric
ConcHat	number	jack-knife unbiased estimate of concentration	mg/L

2.3 INFO

The INFO dataframe is used to store information about the measurements, such as station name, parameter name, drainage area, etc. There can be many additional, optional columns, but the columns in Table 5 are required to initiate the EGRET analysis. After running the WRTDS calculations (as will be described in 5), additional columns (Table 6) are automatically inserted into the INFO dataframe (the meaning of the values will be discussed further sections):

Table 5: INFO dataframe

ColumnName	Type	Description
shortName	string	Name of site, suitable for use in graphical headings
staAbbrev	string	Abbreviation for station name, used in saveResults
paramShortName	string	Name of constituent, suitable for use in graphical headings
constitAbbrev	string	Abbreviation for constituent name, used in saveResults
drainSqKm	numeric	Drainage area in km^2
paStart [a]	integer (1-12)	Starting month of period of analysis
paLong [b]	integer (1-12)	Length of period of analysis in months

[a] Inserted with the setPA function
[b] Inserted with the setPA function

Table 6: INFO dataframe, post-WRTDS

ColumnName	Description	Units
bottomLogQ	Lowest discharge in prediction surfaces	numeric
stepLogQ	Step size in discharge in prediction surfaces	numeric
nVectorLogQ	Number of steps in discharge, prediction surfaces	numeric
bottomYear	Starting year in prediction surfaces	numeric
stepYear	Step size in years in prediction surfaces	numeric
nVectorYear	Number of steps in years in prediction surfaces	numeric
windowY	Half-window width in the time dimension	years
windowQ	Half-window width in the log discharge dimension	numeric
windowS	Half-window width in the seasonal dimension	years
minNumObs	Minimum number of observations for regression	integer
minNumUncen	Minimum number of uncensored observations	integer

2.4 Units

EGRET uses entirely SI units to store the data, but for purposes of output, it can report results in a wide variety of units. The default is that concentration is measured in mg/L, discharge is cubic meters per second (cms), flux is kg/day, and drainage area is km^2. When discharge values are imported from USGS web services (using the dataRetrieval package), they are automatically converted from cubic feet per second (cfs) to cms unless the argument convet is set to FALSE. This can cause confusion if not careful.

Although the data is stored in the dataframes in SI, it is possible to report the results in a variety of units. For all functions that provide output, there are two arguments that can be defined to set the output units: qUnit and FluxUnit. qUnit and FluxUnit can be defined by a numeric code or name. There are two functions that can be called to see the options for qUnit and FluxUnit: printqUnitCheatSheet and printFluxUnitCheatSheet.

```
> printqUnitCheatSheet()
```

```
The following codes apply to the qUnit list:
1 =  cfs  ( Cubic Feet per Second )
2 =  cms  ( Cubic Meters per Second )
3 =  thousandCfs  ( Thousand Cubic Feet per Second )
4 =  thousandCms  ( Thousand Cubic Meters per Second )
5 =  mmDay  ( mm per day )
6 =  mmYear  ( mm per year )
```

When a function has an input argument qUnit, you can define the flow units with the index (1-6) as shown above. The choice should be based on the units that are customary for the audience, but also so that the discharge values don't have too many digits to the right or left of the decimal point.

```
> printFluxUnitCheatSheet()
```

```
The following codes apply to the fluxUnit list:
1 =  poundsDay  ( pounds/day )
2 =  tonsDay  ( tons/day )
3 =  kgDay  ( kg/day )
4 =  thousandKgDay  ( thousands of kg/day )
5 =  tonsYear  ( tons/year )
6 =  thousandTonsYear  ( thousands of tons/year )
7 =  millionTonsYear  ( millions of tons/year )
8 =  thousandKgYear  ( thousands of kg/year )
9 =  millionKgYear  ( millions of kg/year )
10 =  billionKgYear  ( billions of kg/year )
11 =  thousandTonsDay  ( thousands of tons/day )
12 =  millionKgDay  ( millions of kg/day )
```

When a function has an input argument FluxUnit, you can define the flux units with the index (1-12) as shown above. The choice should be based on the units that are customary for the audience, but also so that the flux values don't have too many digits to the right or left of the decimal point.

3 Flow History

This section describes functions included in the EGRET package that provide a variety of table and graphical outputs looking only at flow statistics based on time-series smoothing. These functions were designed for studies of long-term streamflow change and work best for daily streamflow data sets of 50 years or longer. This type of analysis might be useful for studying

At this point it is assumed that you can load the daily discharge record into R, create the Daily dataframe, and enter the required meta-data into the INFO dataframe. If not, see the dataRetrieval vignette:

```
> vignette("dataRetrieval")
```

We will walk through an example from the Rio Grande gaging station in Embodo, NM. This is the first stream gage station in the USGS, established by John Wesley Powell in 1888.

```
> #Rio Grande at Embudo, NM
> siteID <- "08279500"
> startDate <- ""
> endDate <- ""
> Daily <- getDVData(siteID,"00060",startDate,endDate,interactive=FALSE)
> INFO <- getMetaData(siteID,"",interactive=FALSE)
> INFO$shortName <- "Rio Grande at Embudo, NM"
```

The first choice you need to make is what period of analysis to use (pa). What is the period of analysis? If we want to examine our data set as a time series of water years, then the period of analysis is October through September. If we want to examine the data set as calendar years then the period of analysis should be January through December. We might want to examine the winter season, which we could define as December through February, then those 3 months become the period of analysis. The only constraints on the definition of a period of analysis are these: It must be defined in terms of whole months. It must be a set of contiguous months (like March-April-May). And it must have a length that is no less than 1 month and no more than 12 months. It can be uniquely defined by two arguments: paLong and paStart. paLong is the length of the period of analysis, and paStart is the first month of the period of analysis. Table 7 summarizes paLong and paStart.

Table 7: Period of Analysis Information

PeriodOfAnalysis	paStart	PaLong
Calendar Year	1	12
Water Year	10	12
Winter	12	3
September	9	1

To set a period running from December through February:

```
> INFO <- setPA(paStart=12,paLong=3)
```

To set the default value (water year):

```
> INFO <- setPA()
```

The next step is to create the annual series of flow statistics. These will be stored in a matrix called annualSeries that contain the statistics described in table 8.

Table 8: Index of Statistics Information

istat	Name
1	1-day minimum flow
2	7-day minimum flow
3	30-day minimum flow
4	median flow
5	mean flow
6	30-day maximum flow
7	7-day maximum flow
8	1-day maximum flow

To create the annualSeries matrix, using the function makeAnnualSeries:

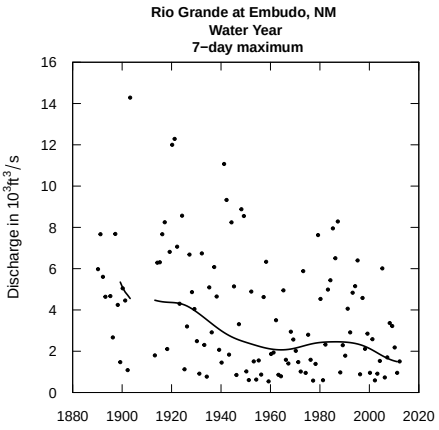
```
> annualSeries <- makeAnnualSeries()
```

Once the annualSeries matrix is created, the plots of any of the stored statistics can be generated with the plotFlowSingle function.

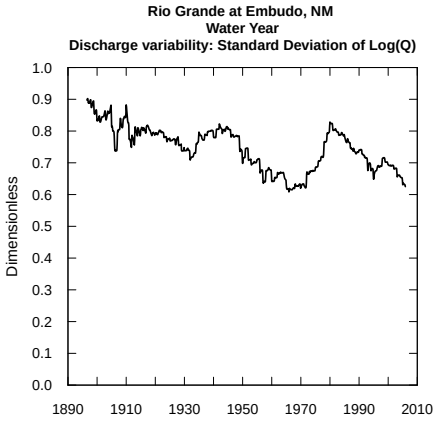
3.1 Plotting Options

This section will give examples of the available plots appropriate for studying flow history once the annualSeries has been created. The plots here will use the default variable input options. For any function, you can get a complete list of input variables (as described in the previous section) in a help file by typing a ? before the function name in the R console. See Appendix B.1 for information on the available input variables for these plotting functions. Also, the complete EGRET manual has more detailed information for each plot type (link to download).

```
> Daily <- getDVData(siteID,"00060",startDate,endDate,interactive=FALSE)
> INFO <- getMetaData(siteID,"",interactive=FALSE)
> INFO$shortName <- "Rio Grande at Embudo, NM"
> INFO <- setPA()
> annualSeries <- makeAnnualSeries()
> plotFlowSingle(istat=7,qUnit="thousandCfs")
> plotSDLogQ()
> plotQTimeDaily(1990,2010,qLower=1,qUnit=3)
> plotFour(qUnit=3)
> plotFourStats(qUnit=3)
```

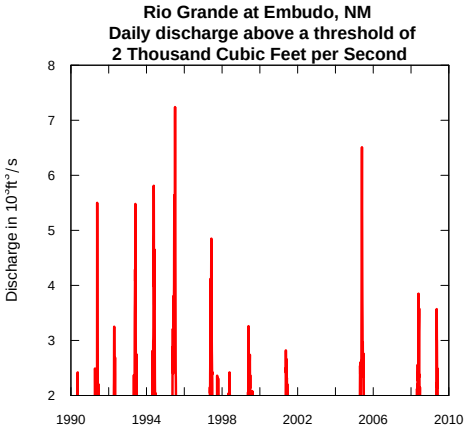


(a) Default plotFlowSingle



(b) Default plotSDLogQ

Figure 1



(a) Default plotQTimeDaily

Figure 2

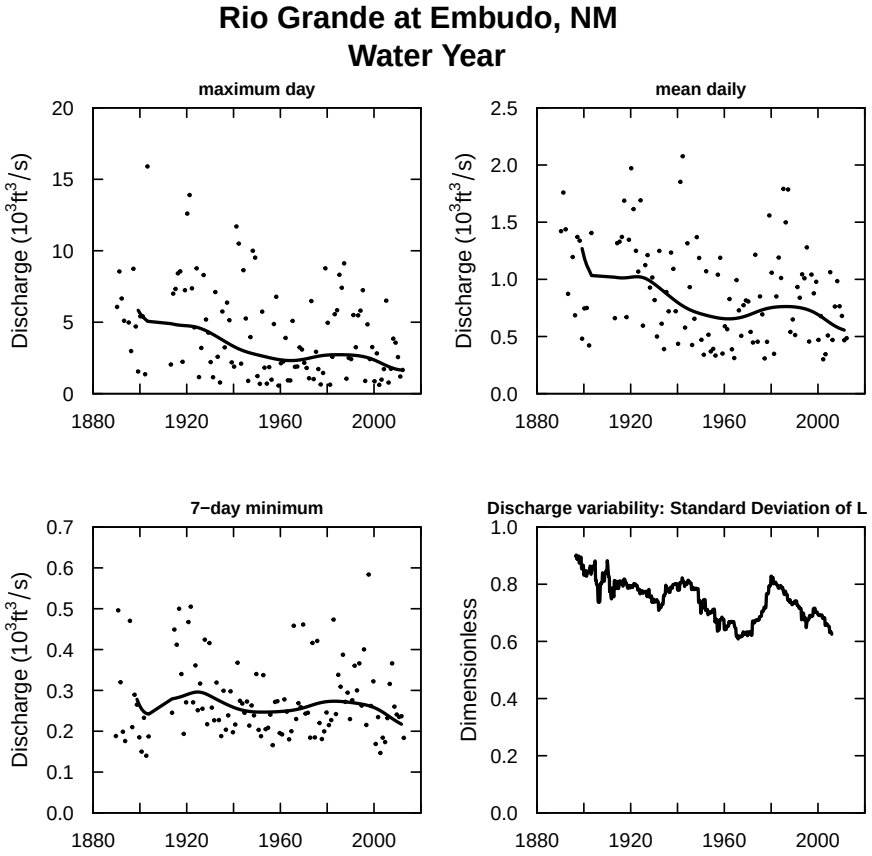


Figure 3: Default plotFour

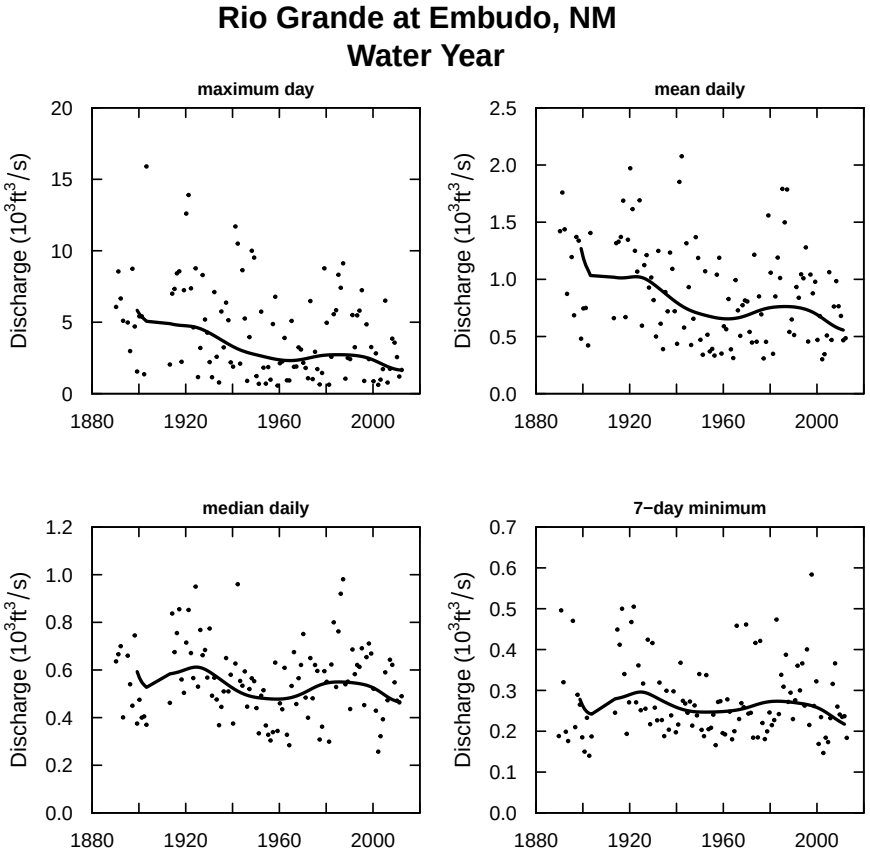


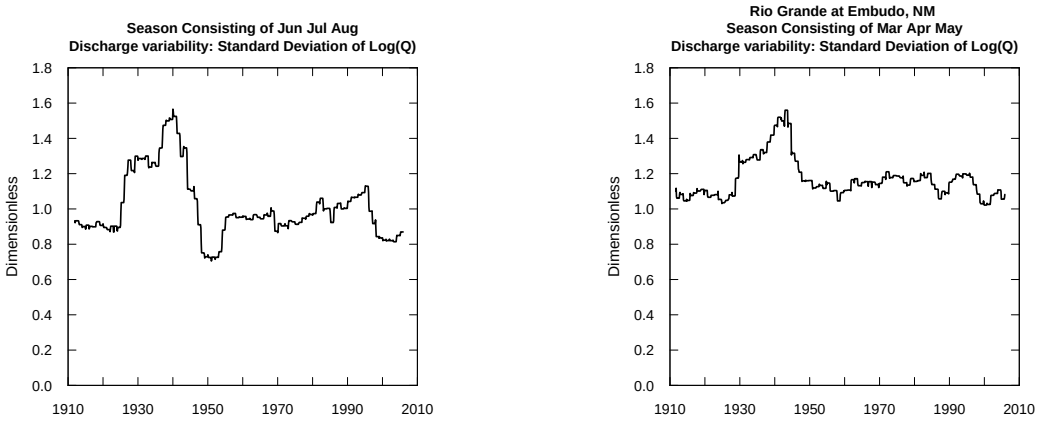
Figure 4: Default plotFourStats

The simplest way to look at these time series is with the function plotFlowSingle. The statistic index (istat) must be defined, but other input arguements can defined. To see a list of these optional arguments and other information about the function, type ?plotFlowSingle in the R console.

plotSDLogQ produces a graphic of the running standard deviation of the log of daily discharge over time. The idea is to get some idea of how variability of daily discharge is changing over time. By using the standard deviation of the log discharge the statistic becomes dimensionless. It also means that it is a way of looking at variability quite aside from average values, so, in the case of a system where discharge might be increasing over a period of years, this provides a way of looking at the variability relative to that changing mean value. It is much like a coefficient of variation, but it has sample properties that make it a smoother measure of variability. There are often comments about how things like urbanization or enhanced greenhouse gases in the atmosphere are bringing about an increase in variability, this is one way to explore that idea.

The Rio Grande does not contain much seasonal variablity, but as quick side example, you can see that there is a difference between variability in spring and summer on the Red River of the North (Figure 5).

```
> siteID <- "05082500"
> DailyRed <- getDVData(siteID,"00060","","",interactive=FALSE)
> INFORed <- getMetaData(siteID,"",interactive=FALSE)
> INFORed$shortName <- "Red River, ND"
> INFORed <- setPA(paStart=6, paLong=3,localINFO = INFORed)
> plotSDLogQ(localDaily = DailyRed, localINFO = INFORed, printStaName = FALSE)
```



(a) Summer (June-August)

(b) Spring (March-May)

Figure 5: Discharge variability on the Red River of the North, ND

plotQTimeDaily is simply a time series plot of discharge. But, it is most suited for showing events above some discharge threshold. In the simplest case, it can plot the entire record, but given the line weight and use of an arithmetic scale it will primarily provide a visual focus on the higher values. plotQTimeDaily requires startYear and endYear, along with some other optional arguements (see ?plotQTimeDaily for more details).

plotFour, plotFourStats, and plot15 are all designed to plot several graphs from the other functions all in a single figure.

3.2 Table Options

Sometimes easier to consider the results in table formats rather than graphically. Similar to the function plotFlowSingle, the printSeries will print the requested flow statistics (Table 8). A small sample of the output is printed below.

```
> printSeries(istat=3, qUnit=3)

Rio Grande at Embudo, NM
 Water Year
    30-day minimum
    Thousand Cubic Feet per Second
   year   annual   smoothed
           value    value
   1899    0.280    0.296
   1900    0.208    0.285
   1901    0.169    0.277
   1902    0.320    0.272
...
   2011    0.252    0.248
   2012    0.257       NA
```

Another way to look at the results is to consider how much the smoothed values change between various pairs of years. These changes can be represented in four different ways.

- As a change between the first and last year of the pair, expressed in the flow units selected.
- As a change between the first and last year of the pair, expressed as a percentage of the value in the first year
- As a slope between the first and last year of the pair, expressed in terms of the flow units per year.
- As a slope between the first and last year of the pair, expressed as a percentage change per year (a percentage based on the value in the first year).

There is another argument that can be very useful in this function: yearPoints. In the default case, the set of years that are compared are at 5 year intervals along the whole data set. If the data set was quite long this can be a daunting number of comparisons. For example, in an 80 year record, there would be 136 such pairs. Instead, we could look at changes for every 20 years starting in 1930:

```
> annualSeries <- makeAnnualSeries()
> tableFlowChange(istat=3, qUnit=3,yearPoints=c(1930,1950,1970,1990,2010))
```

```
Rio Grande at Embudo, NM
Water Year
 30-day minimum

           Streamflow Trends
time span          change        slope       change        slope
                   10^3 cfs 10^3 cfs /yr         %         %/yr
1930  to  1950       -0.054      -0.0027          -17        -0.84
1930  to  1970       -0.046      -0.0011          -14        -0.36
1930  to  1990       -0.029     -0.00048         -8.9        -0.15
1930  to  2010       -0.078     -0.00098          -24         -0.3
1950  to  1970        0.008        4e-04            3         0.15
1950  to  1990        0.025      0.00063          9.4         0.24
1950  to  2010       -0.024     -0.00041         -9.1        -0.15
1970  to  1990        0.017      0.00086          6.2         0.31
1970  to  2010       -0.032     -0.00081          -12        -0.29
1990  to  2010        -0.05      -0.0025          -17        -0.84
```

See Appendix C for instructions on converting an R dataframe to a table in Microsoft.

4 Water Quality Analysis (pre-WRTDS)

Before running the WRTDS model, it is very helpful to take a look at the measured data in a graphical way to understand its behavior and to identify things that might be errors in the data set or learn about the temporal distribution of the data (identify gaps) prior to running the model. It is always best to clear up these issues before moving forward.

We will now use the Choptank River at Greensboro, MD as our example case. The Choptank River is a major tributary of the Chesapeake Bay. Inorganic nitrogen (nitrate and nitrite) has been measured from 1979 onward. First, we need to get the streamflow and nitrate data into R, then use the mergeReport function to associate flow with the discrete measured water quality data.

```
> siteID <- "01491000" #Choptank River at Greensboro, MD
> startDate <- "1979-10-01"
> endDate <- "2011-09-30"
> param<-"00631"
> Daily <- getDVData(siteID,"00060",startDate,endDate)

There are  11688 data points, and  11688 days.

> INFO<- getMetaData(siteID,param,interactive=FALSE)
> INFO$shortName <- "Choptank River"
> INFO <- setPA()
> annualSeries <- makeAnnualSeries()
> Sample <- getSampleData(siteID,param,startDate,endDate)
> Sample <- mergeReport()

 Discharge Record is 11688 days long, which is 32 years
 First day of the discharge record is 1979-10-01 and last day is 2011-09-30
 The water quality record has 606 samples
 The first sample is from 1979-10-24 and the last sample is from 2011-09-29
 Discharge: Minimum, mean and maximum 0.00991 4.09 246
 Concentration: Minimum, mean and maximum 0.05 1.1 2.4
 Percentage of the sample values that are censored is 0.17 %
```

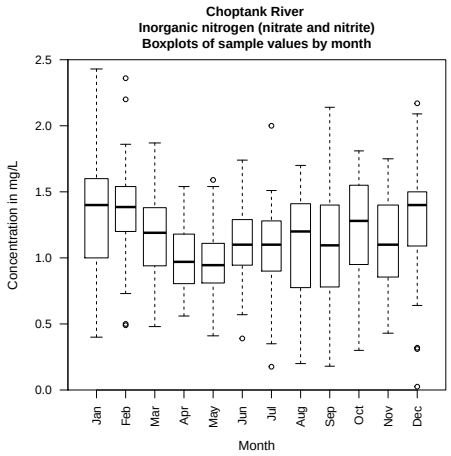
4.1 Plotting Options

This section will give examples of the available plots appropriate for analyzing the data prior to performing a WRTDS analysis. The plots here will use the default variable input options. For any function, you can get a complete list of input variables (as described in the previous section) in a help file by typing a ? before the function name in the R console. See Appendix B.2 for information on the available input variables for these plotting functions. Also, the complete EGRET manual has more detailed information for each plot type (link to download).

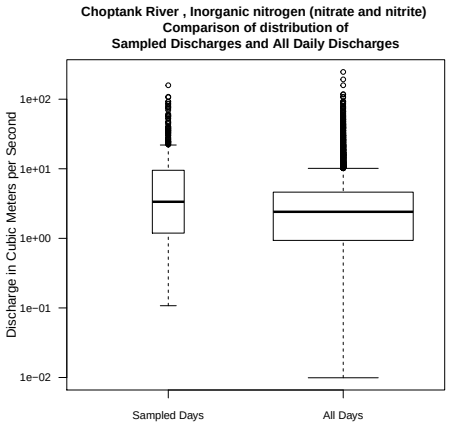
One note about any of the plotting functions that show the sample data: If a value in the data set is a non-detect. Then it is displayed on a graph as a vertical line. The top of the line is the reporting limit and the bottom is either zero, or if the graph is plotting log concentration values, the minimum value on the y-axis. This line is an 'honest' representation of what we know about that observation and doesn't involve us using a statistical model to fill in what we don't know.

```
> boxConcMonth()
> boxQTwice()
> plotLogConcTime()
> plotConcTime()
```

```
> plotConcQ()
> plotLogConcQ()
> plotLogFluxQ()
> multiPlotDataOverview()
```

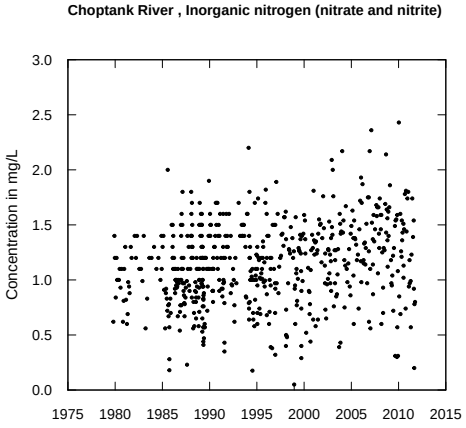


(a) Default boxConcMonth

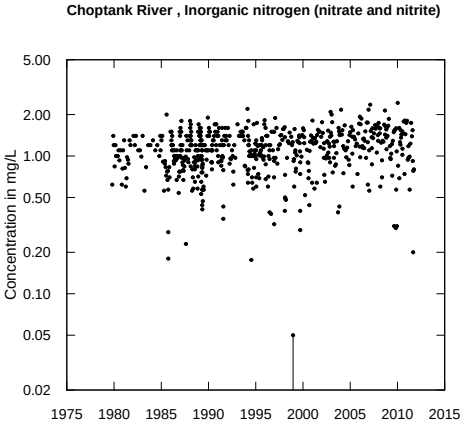


(b) Default boxQTwice

Figure 6: Monthly sample distributions (left) and flow distributions (right)

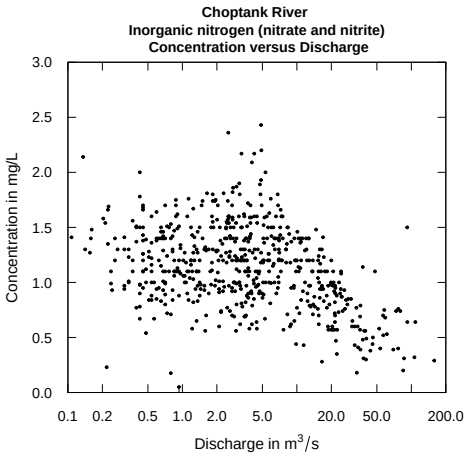


(a) Default plotConcTime

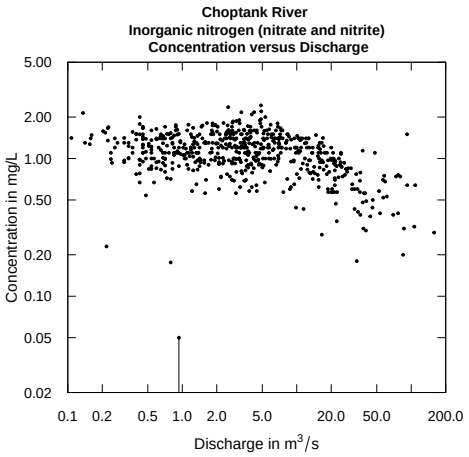


(b) Default plotLogConcTime

Figure 7: Concentration vs. Time (Linear=left, Log=right)

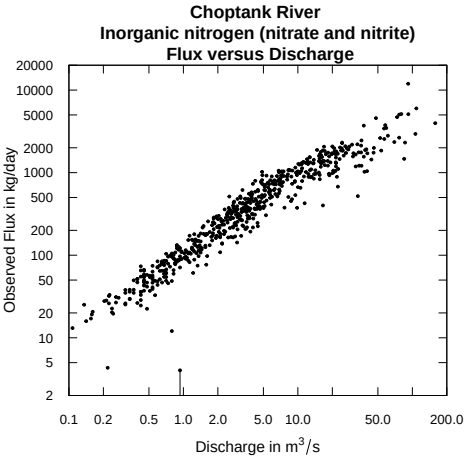


(a) Default plotConcQ



(b) Default plotLogConcQ

Figure 8: Concentration vs. Discharge (Linear=left, Log=right)



(a) Default plotLogFluxQ

Figure 9: Log flux vs. discharge

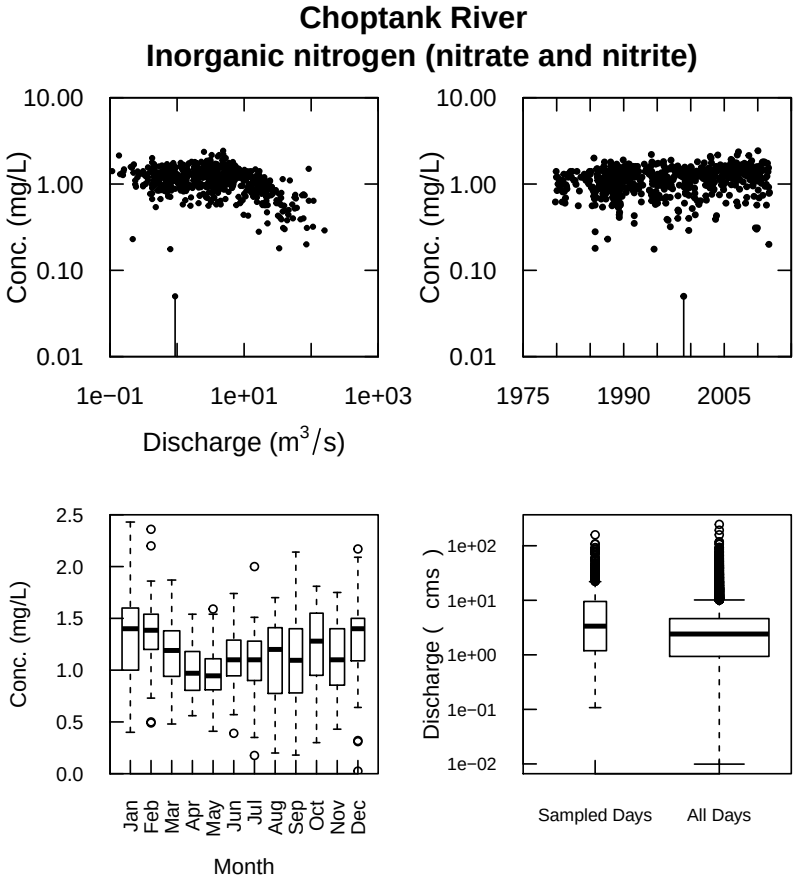


Figure 10: Default multiPlotDataOverview

4.2 Extending Plots Past Defaults

The basic plotting options were shown in the previous section. This section demonstrates some ideas on how to extend the capabilities of the EGRET plots.

```
>   bp <- boxConcMonth()
>   mtext(paste(bp$n, sep = ""), at = seq_along(bp$n),
        line = -1, side = 3, cex=0.75)
>   mtext("n = ",side=3,line=-1,adj=.01, cex=0.75)
```

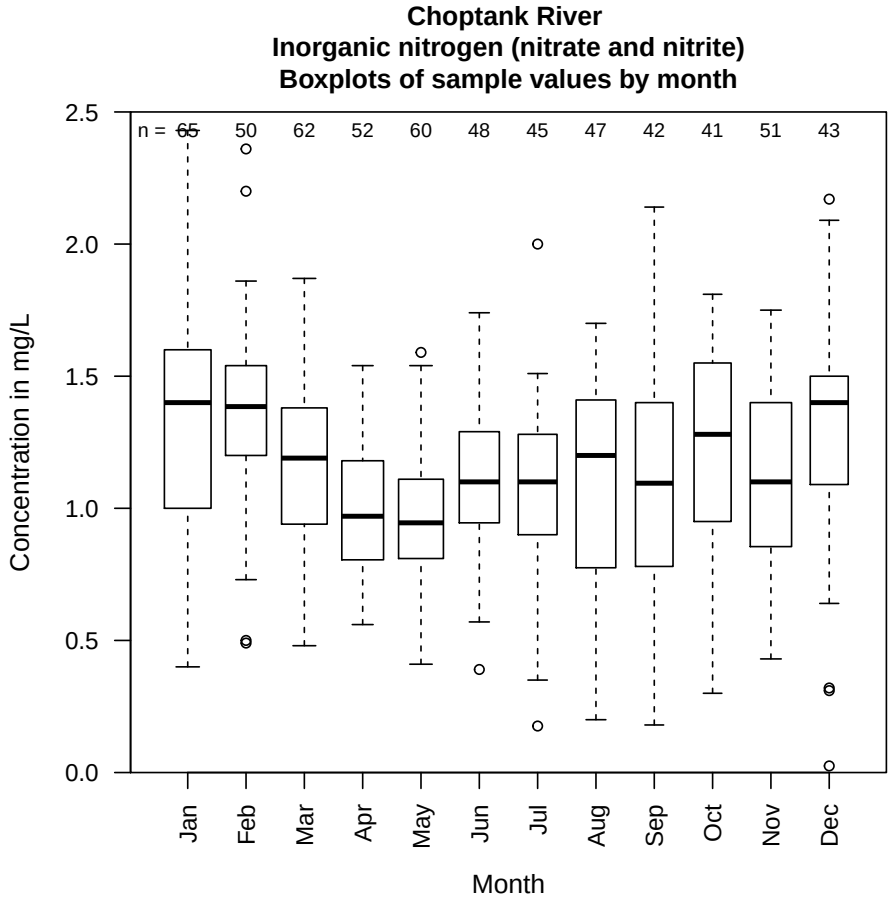


Figure 11: Adding number of samples to boxConcMonth

4.3 Table Options

Another useful tool for checking the data before running the WRTDS estimations is flowDuration. This is a utility function that can help define the flow ranges that we want to explore. It prints out key points on the flow duration curve. They are defined for a particular part of the year, although they can be done for the entire year.

```
> flowDuration()

Flow Duration for Choptank River

Flow duration is based on full year

Discharge units are Cubic Meters per Second
```

```
      min         5%        10%        25%        50%        75%        90%        95%        max
9.91e-03  3.40e-01  4.53e-01  9.34e-01  2.41e+00  4.62e+00  8.21e+00  1.31e+01  2.46e+02
```

5 WRTDS Analysis

Weighted Regressions on Time, Discharge and Season (WRTDS) creates a model of long-term trends in river-water quality, seasonal components, and discharge-related components of the behavior of measured water-quality parameters. In this section, we will step though the process require for a WRTDS analysis. The next section (6) will detail the available methods to view and evaluate the model results.

Once you have looked at your data using the tools described in section 4, and have determined there is sufficient representative data, it is time to run the WRTDS model. There are a few inputs that can be defined before running the model (see Appendix B.3).

Assuming you are using the defaults, with dataframes called Daily, Sample, and INFO, the modelEstimation function will run the WRTDS modeling algorithm:

```
> modelEstimation()
```

This function is slow, and shows the progress in percent complete. See the references and manual for more information. It's important to understand that this is the one function that will globally change your Daily, Sample, and INFO dataframes. It is unusual R programming behavior (and generally considered poor practice), but was chosen to make it easy for the user.

The next step is for the user to select the period of analysis (see section 3) to use in looking at the summary results from WRTDS. In the simplest case, where you would like to do the analysis by water years, the call would be:

```
> AnnualResults<-setupYears()
```

Finally, it is a good idea to save your results because of the computational time that has been invested in producing these results. Assuming that you have already created the object savePath, the command is just

```
> savePath <- "C:/Users/ldecicco/WRTDS_Output"
> saveResults(savePath)
```

This will now save all of the objects in your workspace.

6 WRTDS Results

At this point (after having run modelEstimation and setupYears) we can start considering how to view the annual averages for the variables that have been calculated. See Appendix B.4 for common input variables for these functions. Additionally, check the help files (in the R console, type ? followed by the function name).

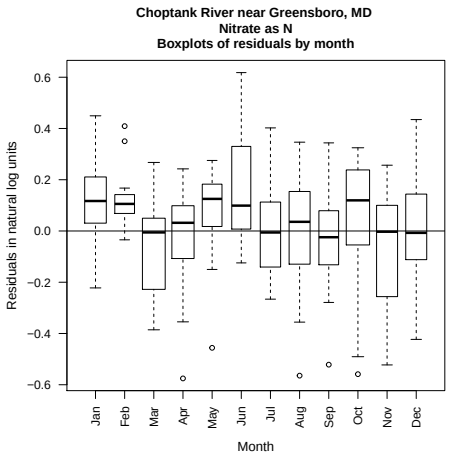
6.1 Plotting Options

Check the help files or manual for more details on the following functions. Defaults are shown here, unless an input is required:

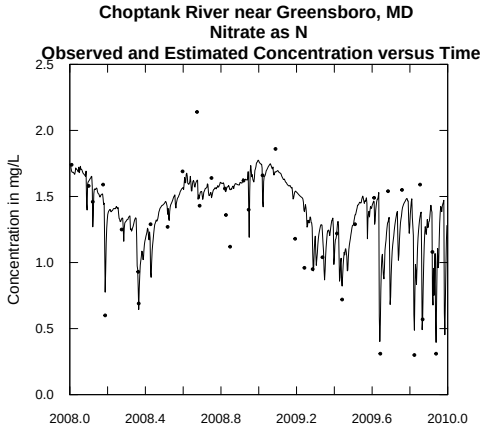
```
> #All plotting functions post-modelEstimation:
>
> yearStart <- 2008
> yearEnd <- 2010
> #Require Sample + INFO:
> boxResidMonth()
> plotConcTimeDaily(yearStart, yearEnd)
> plotFluxTimeDaily(yearStart, yearEnd)
> plotConcPred()
> plotFluxPred()
> plotLogConcPred()
> plotLogFluxPred(tinyPlot=FALSE)
> plotResidPred()
> plotResidQ()
> plotResidTime()
> #Require annualResults + INFO:
> AnnualResults <- setupYears()
> plotConcHist()
> plotFluxHist()
> # Multi-line plots:
> date1 <- "2000-09-01"
> date2 <- "2005-09-01"
> date3 <- "2009-09-01"
> plotLogConcQSmooth(date1, date2, date3, qBottom, qTop, concMax=2, concMin=0.1)
> plotConcQSmooth(date1, date2, date3, qBottom, qTop, concMax=2)
> q1 <- 10
> q2 <- 25
> q3 <- 75
> centerDate <- "07-01"
> plotConcTimeSmooth(q1, q2, q3, centerDate, 2000, yearEnd)
> # Multi-plot:
```

```
> fluxBiasMulti()
> #Contour plots:
> qBottom<-0.1
> qTop<-100
> clevel<-seq(0,2,0.5)
> maxDiff<-0.8
> plotContours(yearStart,yearEnd,qBottom,qTop, contourLevels = clevel)
> plotDiffContours(year0=2000,yearEnd,qBottom,qTop,maxDiff)
```

The following plots are generated:

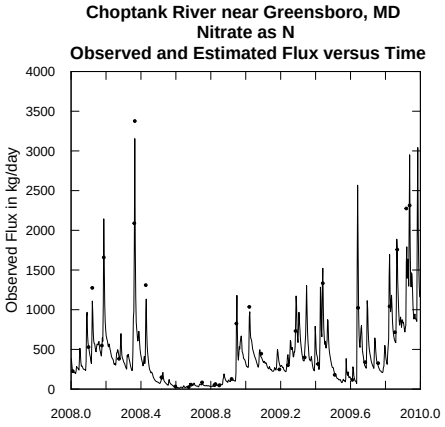


(a) Default boxResidMonth

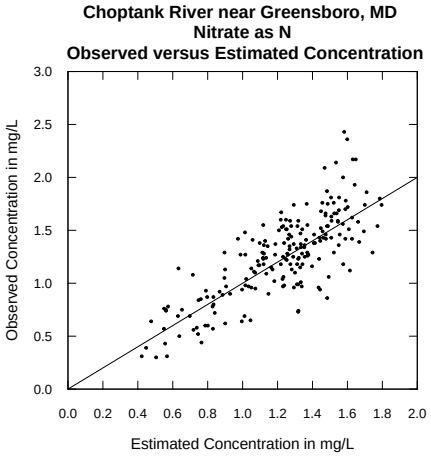


(b) Default plotConcTimeDaily

Figure 12

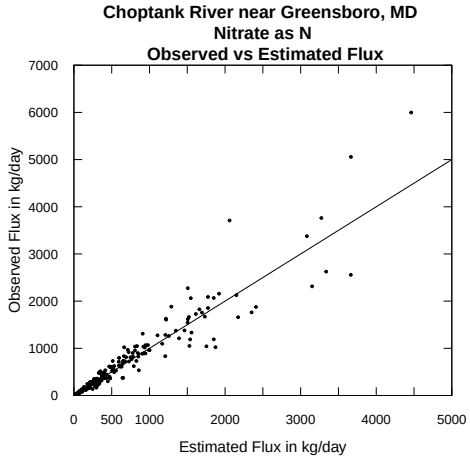


(a) Default plotFluxTimeDaily

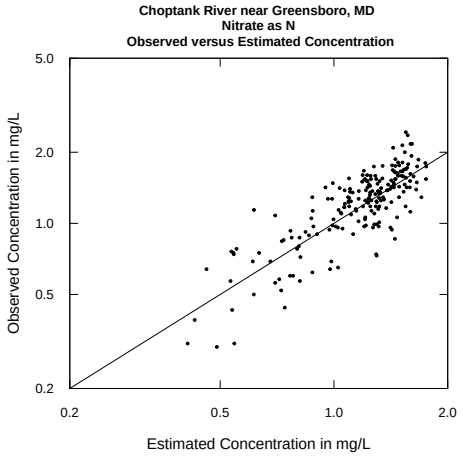


(b) Default plotConcPred

Figure 13

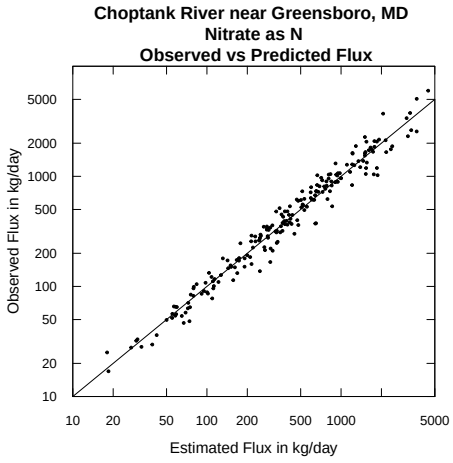


(a) Default plotFluxPred

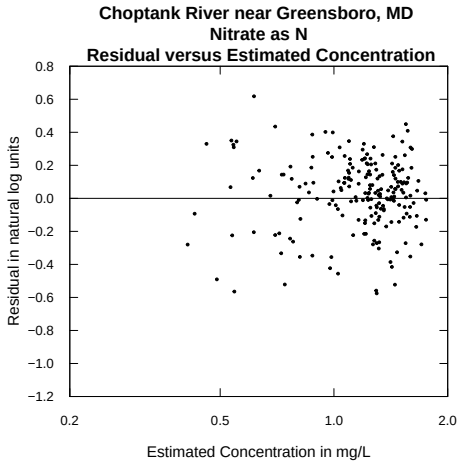


(b) Default plotLogConcPred

Figure 14

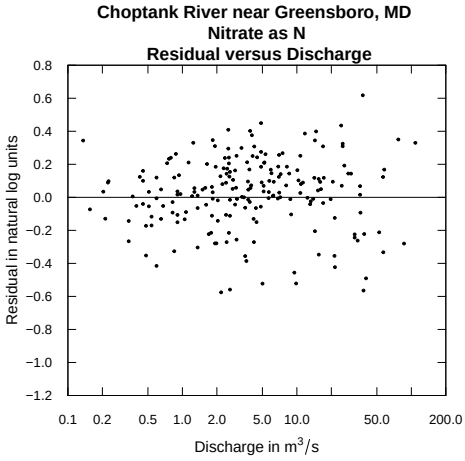


(a) Default plotLogFluxPred

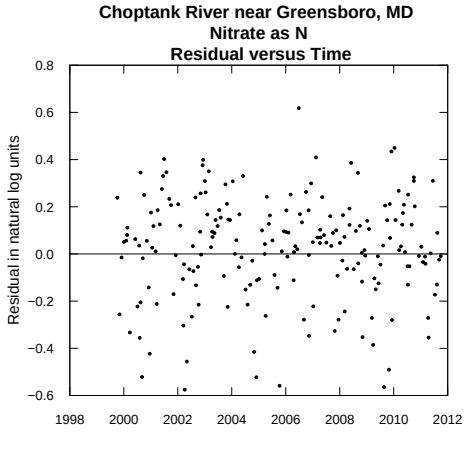


(b) Default plotResidPred

Figure 15

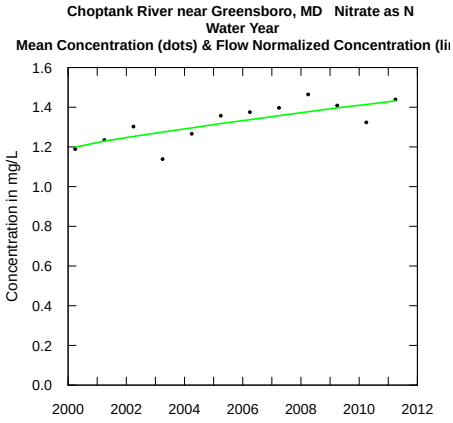


(a) Default plotResidQ

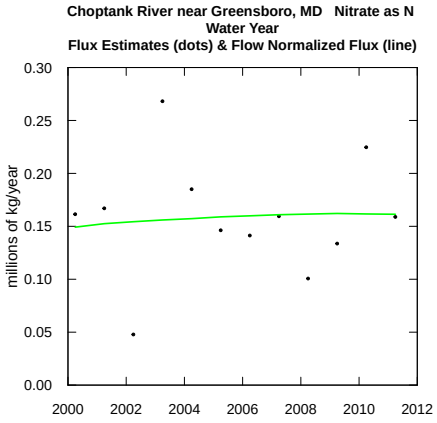


(b) Default plotResidTime

Figure 16

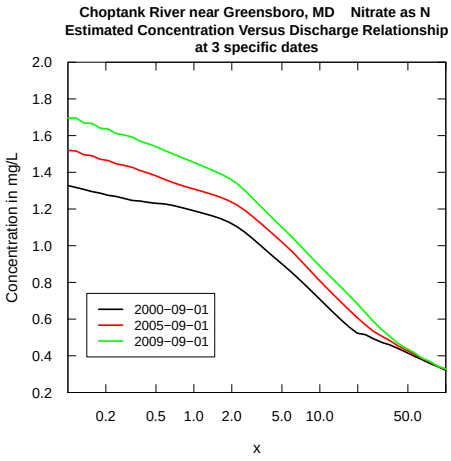


(a) Default plotConcHist

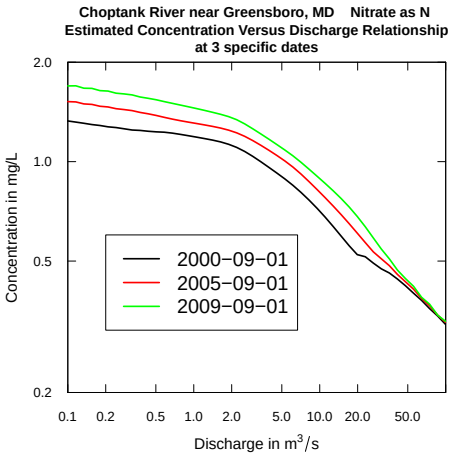


(b) Default plotFluxHist

Figure 17

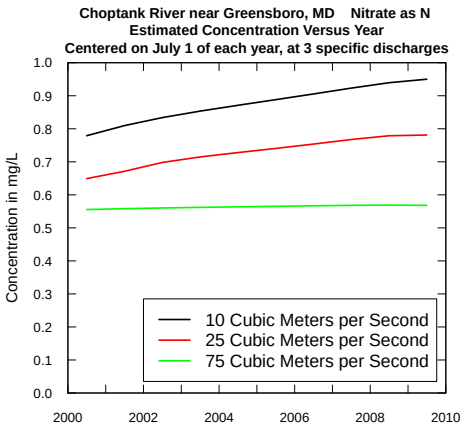


(a) Default plotConcQSmooth



(b) Default plotLogConcQSmooth

Figure 18



(a) Default plotConcTimeSmooth

Figure 19

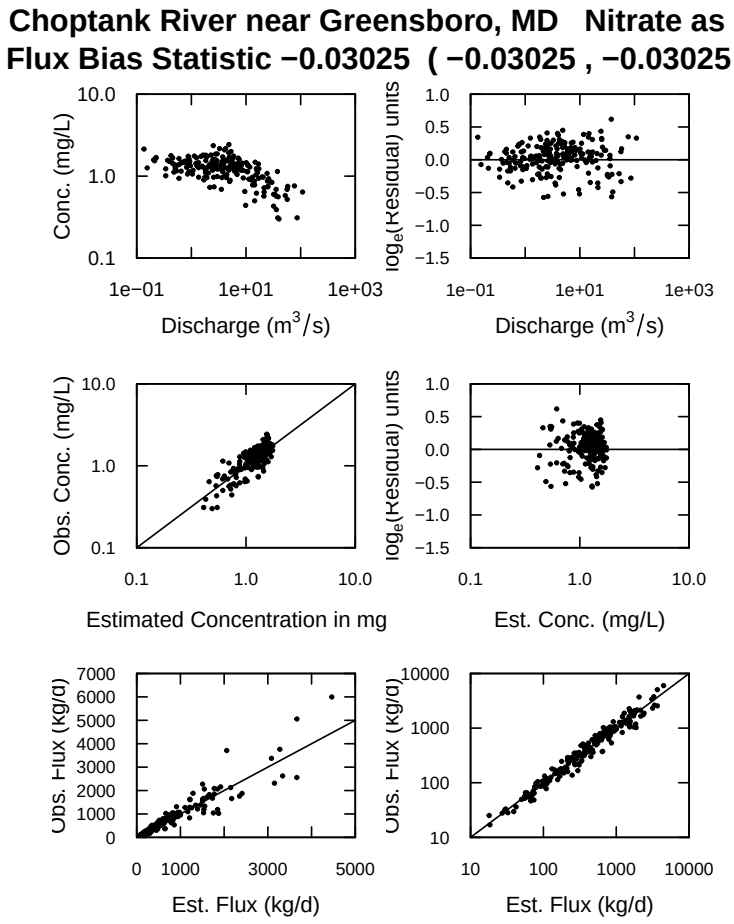


Figure 20: Default fluxBiasMulti

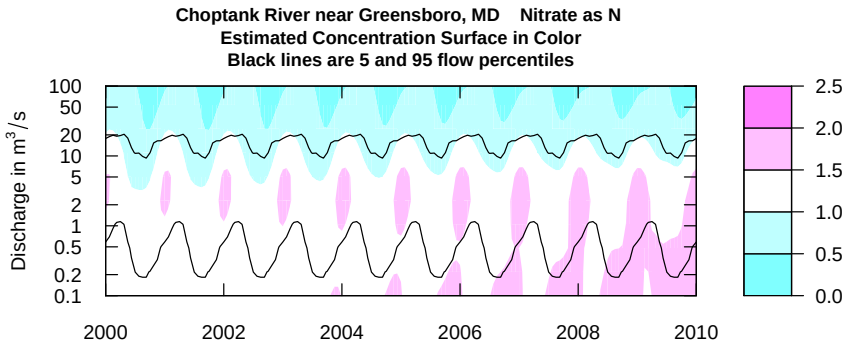


Figure 21: Default plotContours

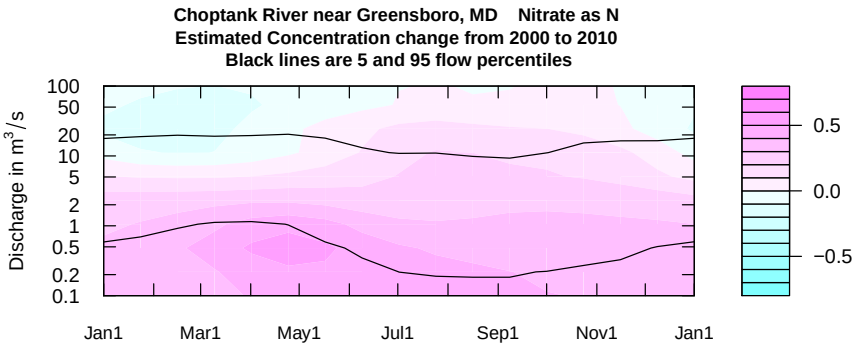


Figure 22: Default plotDiffContours

6.2 Table Options

Sometimes easier to consider the results in table formats rather than graphically. The function tableResults produces a simple text table that contains the annual values for the results. Each row of the output represents a year and it prints: year, average discharge, average concentration, flow normalized concentration, average flux, and flow normalized flux.

```
> tableResults()
```

```
   Choptank River near Greensboro, MD
   Nitrate as N
   Water Year

   Year   Discharge    Conc    FN_Conc     Flux    FN_Flux
             cms            mg/L             10^6 kg/yr

   2000      4.72      1.19      1.20    0.1614     0.149
   2001      4.88      1.23      1.23    0.1670     0.152
   2002      1.24      1.30      1.25    0.0478     0.154
   2003      8.64      1.14      1.28    0.2682     0.156
   2004      5.28      1.27      1.30    0.1851     0.157
   2005      3.81      1.36      1.32    0.1462     0.159
   2006      3.59      1.38      1.34    0.1413     0.160
   2007      4.28      1.40      1.36    0.1595     0.161
   2008      2.56      1.46      1.38    0.1007     0.162
   2009      3.68      1.41      1.40    0.1337     0.162
   2010      7.19      1.32      1.41    0.2247     0.162
   2011      5.24      1.44      1.43    0.1589     0.161
NULL
```

Additionally, you can have the tableResults return the results as a dataframe. This can be useful for further R analysis, or creating more visually appealing tables. For instance, using the xtable package, one can create a LaTeXtable. See Appendix C for instructions on converting an R dataframe to a table in Microsoft.

```
> tableOut <- tableResults(returnDataFrame = TRUE)
> data.table <- xtable(tableOut,
     caption="Table created from tableResults function",
     label="table:tableResults")
> print(data.table, caption.placement="top",
       include.rownames=FALSE,table.placement="!ht")
```

Table 9: Table created from tableResults function

Year	Discharge [cms]	Conc [mg/L]	FN_Conc [mg/L]	Flux [10^6kg/yr]	FN_Flux [10^6kg/yr]
2000	4.72	1.19	1.20	0.1614	0.149
2001	4.88	1.23	1.23	0.1670	0.152
2002	1.24	1.30	1.25	0.0478	0.154
2003	8.64	1.14	1.28	0.2682	0.156
2004	5.28	1.27	1.30	0.1851	0.157
2005	3.81	1.36	1.32	0.1462	0.159
2006	3.59	1.38	1.34	0.1413	0.160
2007	4.28	1.40	1.36	0.1595	0.161
2008	2.56	1.46	1.38	0.1007	0.162
2009	3.68	1.41	1.40	0.1337	0.162
2010	7.19	1.32	1.41	0.2247	0.162
2011	5.24	1.44	1.43	0.1589	0.161

The other table option is tableChange. This is a function that provides for the computation of changes or slopes between any selected pairs of time points. These computations are made only on the flow-normalized results.

```
> tableChange()
```

```
  Choptank River near Greensboro, MD
  Nitrate as N
  Water Year

           Concentration trends
  time span       change     slope    change     slope
                    mg/L   mg/L/yr         %       %/yr

2001  to  2006      0.11     0.022       8.8        1.8
2001  to  2011       0.2      0.02        16        1.6
2006  to  2011     0.095     0.019       7.1        1.4


                 Flux Trends
  time span           change        slope       change        slope
                  10^6 kg/yr    10^6 kg/yr /yr      %         %/yr
2001  to  2006       0.0074       0.0015          4.9         0.97
2001  to  2011       0.0089      0.00089          5.8         0.58
2006  to  2011       0.0015      0.00029         0.92         0.18
```

Finally, tableChangeSingle operates exactly the same as tableChange, but it provides either concentration results or flux results, but not both. This can be useful when producing many output tables for a report that is entirely focused on concentration or one that is entirely focused on flux. The arguments are identical to those for tableChange, except that the final two arguments. The first is a logical argument to indicate if a data frame of output should be returned (for later manipulation and printing through other programs such as Excel), this argument is returnDataFrame, and its default is FALSE. The final argument is flux, and the default is TRUE. When flux=TRUE the output is only for flux, and when flux=FALSE the output is only for concentration. Additionally, this function allows for the return of a dataframe. See Appendix C for instructions on converting an R dataframe to a table in Microsoft.

```
> changeOutput <- tableChangeSingle(fluxUnit=3, flux=TRUE,
         yearPoints=c(2002,2004,2006), returnDataFrame = TRUE)
> data.table <- xtable(changeOutput,
         caption="Table created from tableChangeSingle function",
         label="table:tableChangeSingle")
> print(data.table, caption.placement="top",
       include.rownames=FALSE,table.placement="!ht")
```

Table 10: Table created from tableChangeSingle function

Year1	Year2	change [kg/day]	slope [kg/day]	change[%]	slope [%/yr]
2002.00	2004.00	429.00	215.00	34265.00	17133.00
2002.00	2006.00	437.00	109.00	34848.00	8712.00
2004.00	2006.00	437.00	218.00	33687.00	16844.00

A Getting Started

This section describes the options for downloading and installing the dataRetrieval package.

A.1 New to R?

If you are new to R, you will need to first install the latest version of R, which can be found here: `http://www.r-project.org/`.

There are many options for running and editing R code, one nice environment to learn R is RStudio. RStudio can be downloaded here: `http://rstudio.org/`. Once R and RStudio are installed, the environment package needs to be installed as described in the next section.

At any time, you can get information about any function in R by typing a question mark before the functions name. This will open a file (in RStudio, in the Help window) that describes the function, the required arguments, and provides working examples.

```
> ?getJulian
```

To see the raw code for a particular code, type the name of the function:

```
> getJulian
```

```
function (date)
{
    dateTime <- as.Date(date)
    Julian <- as.numeric(julian(dateTime, origin = as.Date("1850-01-01")))
    return(Julian)
}
<environment: namespace:EGRET>
```

A.2 R User: Installing EGRET

To install the EGRET packages and it's dependencies:

```
> install.packages(c("zoo","survival","methods","fields","spam"))
> install.packages("dataRetrieval", repos="http://usgs-r.github.com/",
                   type="source")
> install.packages("EGRET", repos="http://usgs-r.github.com/",
                   type="source")
```

It is a good idea to re-start the R enviornment after installing the package if installing an updated version (that is, restart RStudio).

After installing the package, you need to open the library each time you re-start R. This is done with the simple command:

```
> library(dataRetrieval)
> library(EGRET)
```

Using RStudio, you could alternatively click on the checkbox for dataRetrieval and EGRET in the Packages window.

A.3 R Developers: Installing EGRET from gitHub

Alternatively, R-developers can install the most recent (not-necessarily stable) version of EGRET directly from gitHub using the devtools package. devtools is available on CRAN. Simply type the following commands into R to install the latest version of EGRET available on gitHub. Rtools (for Windows) and appropriate LaTeX tools are required. Be aware that the version installed using this method isn't necessarily the same as the version in the stable release branch.

```
> library(devtools)
> install_github("dataRetrieval", "USGS-R")
> install_github("EGRET", "USGS-R")
```

To then open the library, simply type:

```
> library(dataRetrieval)
> library(EGRET)
```

B Common Function Variables

B.1 flowHistory Plotting Input

Table 11: Variables used in flow history plots

Argument	Definition
istat	Which flow statistic to plot: 1-8. Must be specified, see Table 8.
yearStart[3]	What is the decYear value where you want the graph to start?
yearEnd[3]	What is the decYear value where you want the graph to end?
qMax	User specified upper limit on y axis (can be used when we want several graphs to all share the same scale) will be specified in the discharge units that the user selects.
printTitle	can be TRUE or FALSE, you may want FALSE if it is going to be a figure with a caption or if it is a part of a multipanel plot.
tinyPlot	Can be TRUE or FALSE, the TRUE option assures that there will be a small number of tick marks, consistent with printing in a small space
runoff	Can be TRUE or FALSE. If true then discharge values are reported as runoff in mm/day. This can be very useful in multi-site analyses.
qUnit	An index indicating what discharge units to use. Options run from 1 to 13 (see 2.4). The choice should be based on the units that are customary for the audience but also, the choice should be made so that the discharge values don't have too many digits to the right or left of the decimal point.
printStaName[4]	Can be TRUE or FALSE, if TRUE the name of the streamgage is stated in the plot title.
printPA[4]	Can be TRUE or FALSE, if TRUE the period of analysis is stated in the plot title.
printIstat[4]	Can be TRUE or FALSE, if TRUE the name of the statistic (e.g. 7-day minimum flow) is stated in the plot title.

[3]Setting yearStart and yearEnd will determine where the graphs start and end, but they don't determine where the smoothing analysis starts and ends. There are situations, typically where many sites are be analyzed together, where you may want to run the smoothing on a consistent period of record across all sites. Doing this requires modifying the Daily data frame before running makeAnnualSeries

[4]If the printTitle argument is set to FALSE, then it really makes no difference what you do with printSta, printPA, or printIstat. They can all be left as their default values and thus there is no need to include them in the call for the function.

B.2 Water Quality Plotting Input

Table 12: Variables used in water quality analysis plots

Argument	Definition
qUnit	Determines what units will be used for discharge, see 2.4
printTitle	If TRUE the plot has a title. If FALSE no title (useful for publications where there will be a caption)
qLower	The lower bound on the discharge on the day of sampling that will be used in forming a subset of the sample data set that will be displayed in the graph. It is expressed in the units specified in qUnit. If qLower = NA, then the lower bound is set to zero.
qUpper	The upper bound on the discharge on the day of sampling that will be used in forming a subset of the sample data set that will be displayed in the graph. It is expressed in the units specified in qUnit. If qUpper = NA, then the lower bound is set to infinity.
paLong	The length of the time period that will be used in forming a subset of the sample data set that will be displayed in the graph, expressed in months.
paStart	The starting month for the time period that will be used in forming a subset of the sample data set that will be displayed in the graph. It is expressed in months (calendar months).
concMax	The upper limit on the vertical axis of graphs showing concentration values in mg/L (NA sets value to just above maximum).
concMin	The lower limit on the vertical axis of graphs showing concentration values in mg/L (NA sets value to just below minimum for log scales, zero for linear).
fluxUnit	Determines what units will be used for flux (see Section 2.4).
fluxMax	The upper limit on the vertical axis of graphs showing flux values.

B.3 WRTDS Estimation Input

Table 13: Variables in WRTDS

Argument	Definition	Default
windowY	The half window width for the time weighting, measured in years. Values much shorter than 10 usually result in a good deal of oscillations in the system that are likely not very realistic	10
windowQ	The half window width for the weighting in terms of ln(Q). For very large rivers (average discharge values in the range of many tens of thousands of cfs) a smaller value than 2 may be appropriate, but probably not less than 1	2
windowS	The half window width for the seasonal weighting, measured in years. Any value >0.5 will make data from all seasons have some weight. Values should probably not be lower than 0.3 and there is no need to go higher than 0.5	0.5
minNumObs	This is the minimum number of observations with non-zero weight that the individual regressions will require before they will be used. If there too few observations the program will iterate, making the windows wider until the number increases above this minimum. The only reason to lower this is in cases where the data set is rather small. It should always be set to a number at least slightly smaller than the sample size. Any value lower than about 60 is probably in the 'dangerous' range, in terms of the reliability of the regression	100
minNumUncen	This is the minimum number of uncensored observations with non-zero weight that the individual regressions will require before they will be used. If there are too few uncensored observations the program will iterate, making the windows wider until the number increases above this minimum. The only reason to lower this is in cases where the number of uncensored values is rather small. The method has never been tested in situations where there are very few uncensored values	50

B.4 WRTDS Plotting Input

Table 14: Variables used in WRTDS analysis plots

Argument	**Definition**
qUnit	Determines what units will be used for discharge, see 2.4
fluxUnit	An index indicating what flux units will be used , see 2.4
stdResid	This is an option. If FALSE, it prints the regular residuals (they are in ln concentration units). If TRUE, it is the standardized residuals. These are the residuals divided by their estimated standard error (each residual has its own unique standard error). In theory, the standardized residuals should have mean zero and standard deviation of 1
printTitle	If TRUE the plot has a title. If FALSE no title (useful for publications where there will be a caption)
startYear	The starting date for the graph, expressed as decimal years, for example, 1989.0
endYear	The ending date for the graph, expressed as decimal years, for example, 1996.0
moreTitle	A character variable that adds additional information to the graphic title. Typically used to indicate what the estimation method was (e.g. WRTDS or LOADEST). Default is ' ' which indicates that nothing is added to title
fluxMax	The upper limit on the vertical axis of graphs showing flux values.
plotFlowNorm	If TRUE the graph shows the annual flux values as circles and the flow-normalized values as a green curve. If false, it only shows the annual flux values.

Table 15: Variables used in WRTDS contour plots

Argument	Definition
qUnit	Determines what units will be used for discharge, see 2.4
qBottom	The discharge value that should form the bottom of the graph
qTop	The discharge value that should form the top of the graph
printTitle	If TRUE the plot has a title. If FALSE no title (useful for publications where there will be a caption)
yearStart	The starting date for the graph, expressed as decimal years, for example, 1989.0
yearEnd	The ending date for the graph, expressed as decimal years, for example, 1996.0
whatSurface	default = 3. This should generally be at its default value. At whatSurface = 3, the plotted surface shows the expected value of concentration. For whatSurface = 1, it shows the yHat surface (natural log of concentration). For whatSurface = 2, it shows the SE surface (the standard error in log concentration).
contourLevels	Default value is NA. With the default value the contour intervals are set automatically. These will generally NOT be a very good choice, but they may provide a starting point.
span	Default value = 60. Specifies the smoothness of the flow duration information that goes on this graph. A larger value will make it smoother. The default should work well in most cases.
pval	Default value = 0.05. The probability value for the flow frequency information shown on the plot. The plot has two black curves on it. In the default value case these are at the 5 and 95 percent levels on the seasonal flow duration curve. pval = 0.01 would place these at the 1 and 99 percent points. pval = 0.1 would place them at 10 and 90.
vert1	Default = NA. This simply plots a vertical black line on the graph at a particular time (defined in decimal years). It is used to illustrate the idea of a 'vertical slice' through the contour plot, which might then be shown in a subsequent use of plotConcQSmooth.
vert2	Default = NA. This gives the location of a second vertical black line on the graph at a particular time (defined in decimal years).
horiz	Default = NA. This simply plots a horizontal black line on the graph at a particular discharge value (defined in the units specified by qUnit). It is used to illustrate the idea of the seasonal cycle in concentrations for a given discharge and the long-term change in this cycle.
flowDuration	Default = TRUE. If TRUE it draws the flow duration lines at the specified probabilities. If FALSE, the flow duration lines are left off.

Table 16: Variables used in WRTDS multi-line plots

Argument	Definition
date1	This is the date for the first curve to be shown on the plot. It must be in the form 'yyyy-mm-dd' (it must be in quotes)
date2	This is the date for the second curve to be shown on the plot ('yyyy-mm-dd'), If you don't want a second curve then the argument must be date2=NA
date3	This is the date for the third curve to be shown on the plot ('yyyy-mm-dd'), If you don't want a third curve then the argument must be date3=NA
qUnit	Determines what units will be used for discharge, see printqUnitCheatSheet()
qLow	The discharge value that should form the left edge of the graphic.
qHigh	The discharge value that should form the right edge of the graphic.
legendLeft	This determines the placement of the legend on the graph. It establishes the left edge of the legend and is expressed in the flow units being used. The default (which is NA) will let it be placed automatically. The legend can end up conflicting with one or more of the curves. Once the location of the curves is established then this can be set in a way that avoids conflict.
legendTop	This determines the placement of the legend on the graph. It establishes the top edge of the legend and is expressed according to the concentration values on the y-axis. The default (which is NA) will let it be placed automatically. The legend can end up conflicting with one or more of the curves. Once the location of the curves is established then this can be set in a way that avoids conflict.
concMax	Maximum value for the vertical axis of the graph. The default is NA. The reason to set concMax is if you want to make several plots that have the same vertical axis.
concMin	[This one is only used in plotLogConcQSmooth]. Minimum value for the vertical axis of the graph. The default is NA. The reason to set concMin is if you want to make several plots that have the same vertical axis.
bw	Default is FALSE, which means we want a color plot. If bw=TRUE that means it should be black and white.
printTitle	If TRUE the plot has a title. If FALSE no title (useful for publications where there will be a caption).
printValues	If TRUE the estimated values that make up the plotted lines are printed on the console. If FALSE they are not printed. Default is FALSE. This could be useful if you wanted to compute various comparisons across time periods.
windowY	This is the half-window width for time in WRTDS. It has units of years. The default value is 10.
windowQ	This is the half-window width for discharge in WRTDS. It has units of ln(discharge). The default value is 2.
windowS	This is the half-window width for seasons in WRTDS. It has units of years. The default value is 0.5.

C Creating tables in Microsoft from R

There are a few steps that are required in order to create a table in a Microsoft product (Excel, Word, Powerpoint, etc.) from an R dataframe. There are a variety of good methods, one of which is detailed here. The example we will step through here will be to create a table in Microsoft Word based on the dataframe tableData:

```
> siteNumber <- '01491000'
> ChoptankAvailableData <- getDataAvailability(siteNumber)
> ChoptankDailyData <- ChoptankAvailableData["dv" == ChoptankAvailableData$service,]
> ChoptankDailyData <- ChoptankDailyData["00003" == ChoptankDailyData$statCd,]
> pCodeINFO <- getMultipleParameterNames(ChoptankDailyData$parameter_cd, interactive=FALSE,
> ChoptankDailyData <- merge(ChoptankDailyData,pCodeINFO,by="parameter_cd")
> tableData <- with(ChoptankDailyData,
       data.frame(
       shortName=srsname,
       Start=startDate,
       End=endDate,
       Count=count,
       Units=parameter_units)
       )
```

First, save the dataframe as a tab delimited file (you don't want to use comma delimited because there are commas in some of the data elements):

```
> write.table(tableData, file="tableData.tsv",sep="\t",
             row.names = FALSE,quote=FALSE)
```

This will save a file in your working directory called tableData.tsv. You can see your working directory by typing getwd() in the R console. Opening the file in a general-purpose text editor, you should see the following:

```
shortName  Start  End Count Units
Temperature, water 2010-10-01 2012-06-24 575 deg C
Stream flow, mean. daily 1948-01-01 2013-03-13 23814 cfs
Specific conductance 2010-10-01 2012-06-24 551 uS/cm @25C
Suspended sediment concentration (SSC) 1980-10-01 1991-09-30 3651 mg/l
Suspended sediment discharge 1980-10-01 1991-09-30 3652 tons/day
```

To open this file in Excel:

1. Open Excel

2. Click on the File tab
3. Click on the Open option
4. Browse to the working directory (as shown in the results of getwd())
5. Next to the File name text box, change the dropdown type to All Files (*.*)
6. Double click tableData.tsv
7. A text import wizard will open up, in the first window, choose the Delimited radio button if it is not automatically picked, then click on Next.
8. In the second window, click on the Tab delimiter if it is not automatically checked, then click Finished.
9. Use the many formatting tools within Excel to customize the table

From Excel, it is simple to copy and paste the tables in other Microsoft products. An example using one of the default Excel table formats is here.

shortName	Start	End	Count	Units
Temperature, water	10/1/2010	6/24/2012	575	deg C
Stream flow, mean. daily	1/1/1948	3/13/2013	23814	cfs
Specific conductance	10/1/2010	6/24/2012	551	uS/cm @25C
Suspended sediment concentration (SSC)	10/1/1980	9/30/1991	3651	mg/l
Suspended sediment discharge	10/1/1980	9/30/1991	3652	tons/day

Figure 23: A simple table produced in Microsoft Excel

References

[1] Helsel, D.R. and R. M. Hirsch, 2002. Statistical Methods in Water Resources Techniques of Water Resources Investigations, Book 4, chapter A3. U.S. Geological Survey. 522 pages. http://pubs.usgs.gov/twri/twri4a3/

[2] Hirsch, R. M., Moyer, D. L. and Archfield, S. A. (2010), Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs. JAWRA Journal of the American Water Resources Association, 46: 857-880. doi: 10.1111/j.1752-1688.2010.00482.x http://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.2010.00482.x/full

[3] Sprague, L. A., Hirsch, R. M., and Aulenbach, B. T. (2011), Nitrate in the Mississippi River and Its Tributaries, 1980 to 2008: Are We Making Progress? Environmental Science & Technology, 45 (17): 7209-7216. doi: 10.1021/es201221s http://pubs.acs.org/doi/abs/10.1021/es201221s