



Overview of dataRetrieval and EGRET

USGS R Packages: by Robert Hirsch and Laura De Cicco

Robert M. Hirsch

USGS

2015-02-11

These are two R-packages

Free, quality assured, open-source,
platform independent, documented

`dataRetrieval` retrieves many types of
water data from USGS-NWIS and from
the Portal

`EGRET` is for data exploration: it depends
on `dataRetrieval` and it implements
WRTDS (Weighted Regressions on
Time, Discharge, and Season)

Outline of the presentation

Motivations for the packages

The WRTDS concept

Overview of dataRetrieval

How EGRET works

Motivation for EGRET: Quote From Ralph Keeling

The only way to figure out what is
happening to our planet is to
measure it,

and this means tracking changes
decade after decade

and poring over the records.

Keeling, 2008, Recording Earth's vital signs, Science, p1771-1772

Motivation for dataRetrieval:

Make it easy to import and organize
all types of water data into the R
environment

So you can get to work and **pore**
over your data.

EGRET (Exploration and Graphics for RivEr Trends):

- 1) Obtain and organize: Sample data, daily discharge data, and meta-data**
- 2) Use the WRTDS method to explore evolving water quality conditions**
- 3) Produce graphs and tables**

Guiding ideas for WRTDS

- Describe the evolving behavior of the watershed. No mathematical straight-jacket!!
- Estimate both concentration & flux (averages as well as trends).
- Estimate the actual history but also a flow-normalized history.
- Resolve a serious bias in flux estimates.
- Be quantitative but also exploratory.

Data requirements

- Low intra-day variability (not flashy)
- Requires a complete daily discharge record
- Intended for >200 samples, but has been used for some purposes with as few as 60 samples
- Water quality samples cover most of the discharge range
- For trend studies: 20+ years, but can do less
- For average flux computations: 5 – 10 years.



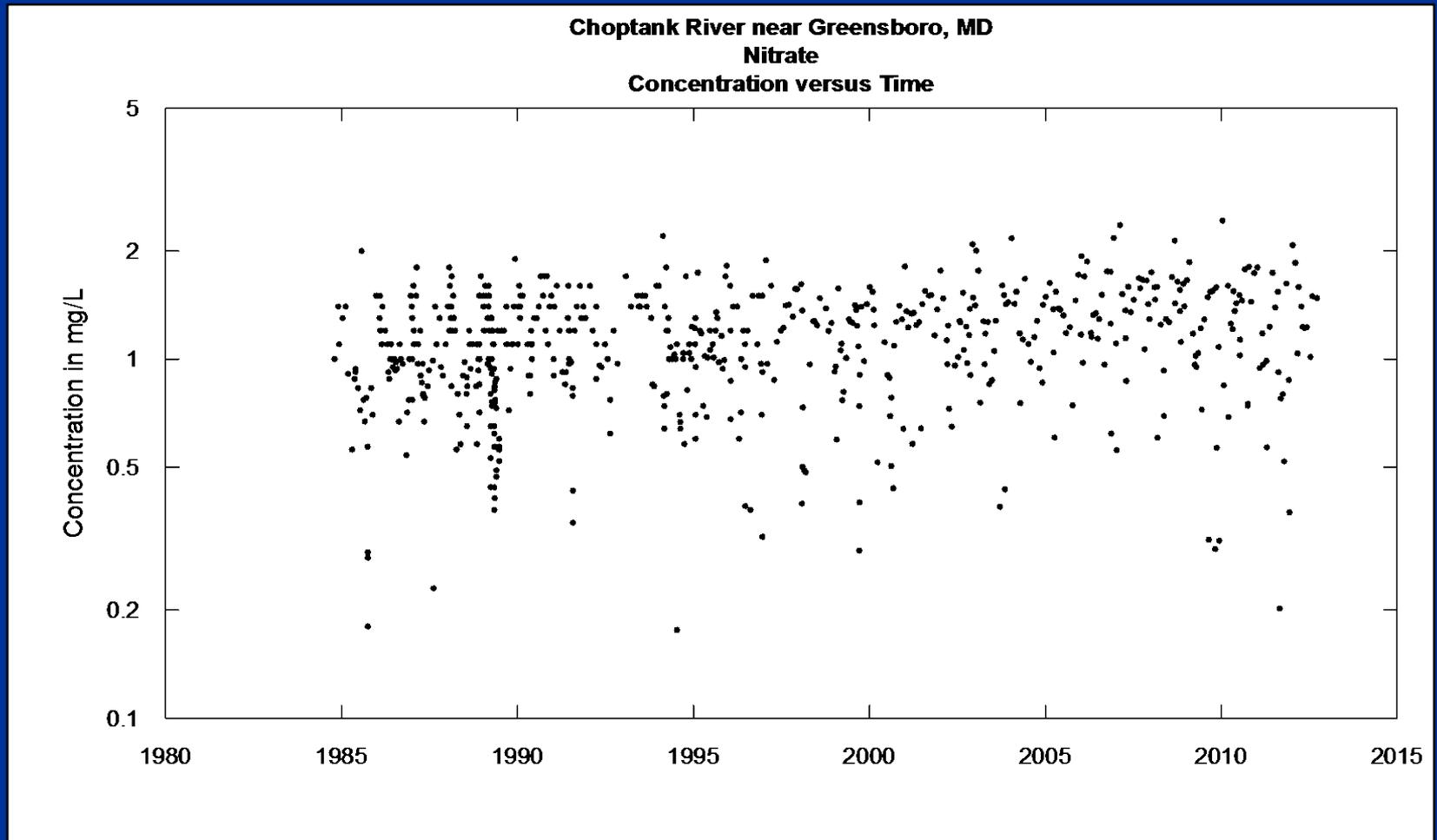
WRTDS Example



**Choptank
River,
293 km² watershed**

“Data without models are chaos, but models without data are fantasy”

Nesbit, Dlugokencky and Bousquet, Science, 31 January 2014, pp. 493-495



Use the data and a simple, highly-flexible smoothing model to decompose the data into 4 components.

1) Discharge related component

2) Seasonal component

3) Time trend

4) Random component

**Weighted Regressions on Time,
Discharge and Season (WRTDS)**

Locally Weighted Regression

For any location in time - discharge space (t and Q) we assume that concentration (c) follows this model

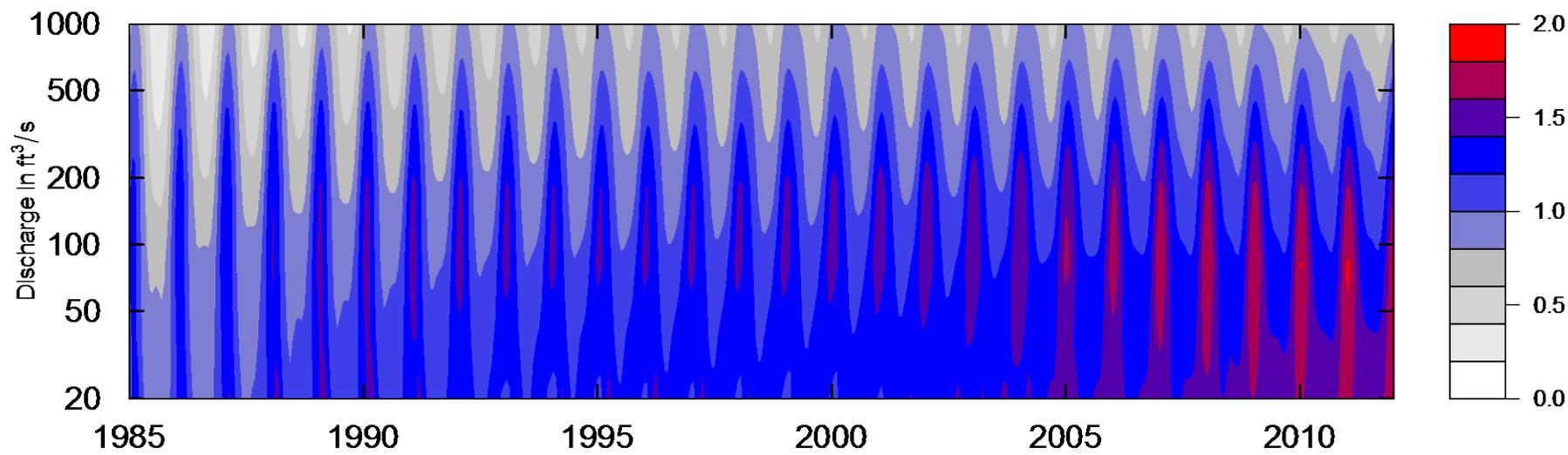
$$\ln(c) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot \ln(Q) + \beta_3 \cdot \sin(2\pi t) + \beta_4 \cos(2\pi t) + \varepsilon$$

But the coefficients should be smoothly changing as we move through the space

Use weighted regression at many points in that space. The weight on each sample is determined by its “relevance” to that particular point in the space.

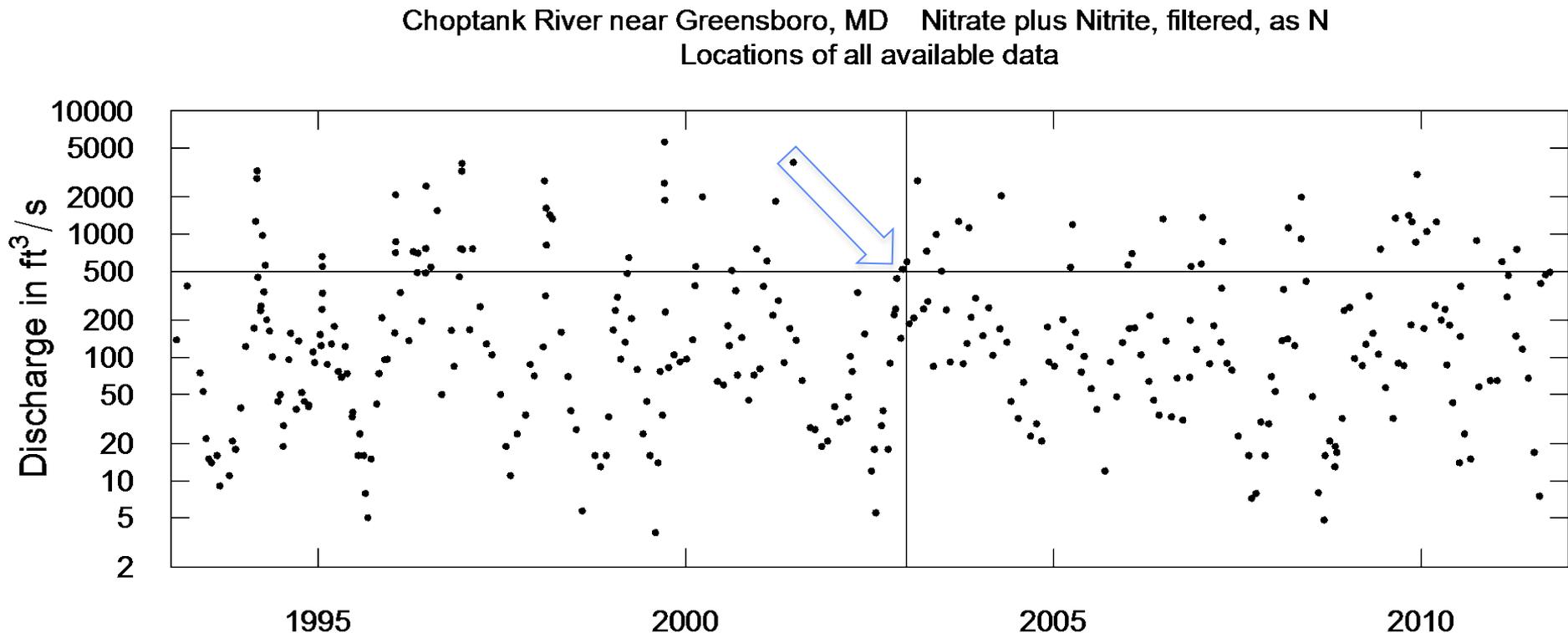
WRTDS view of the evolving behavior of nitrate

Choptank River near Greensboro, MD Nitrate plus Nitrite, Filtered, as N
Estimated Concentration Surface in Color



How is this surface created?

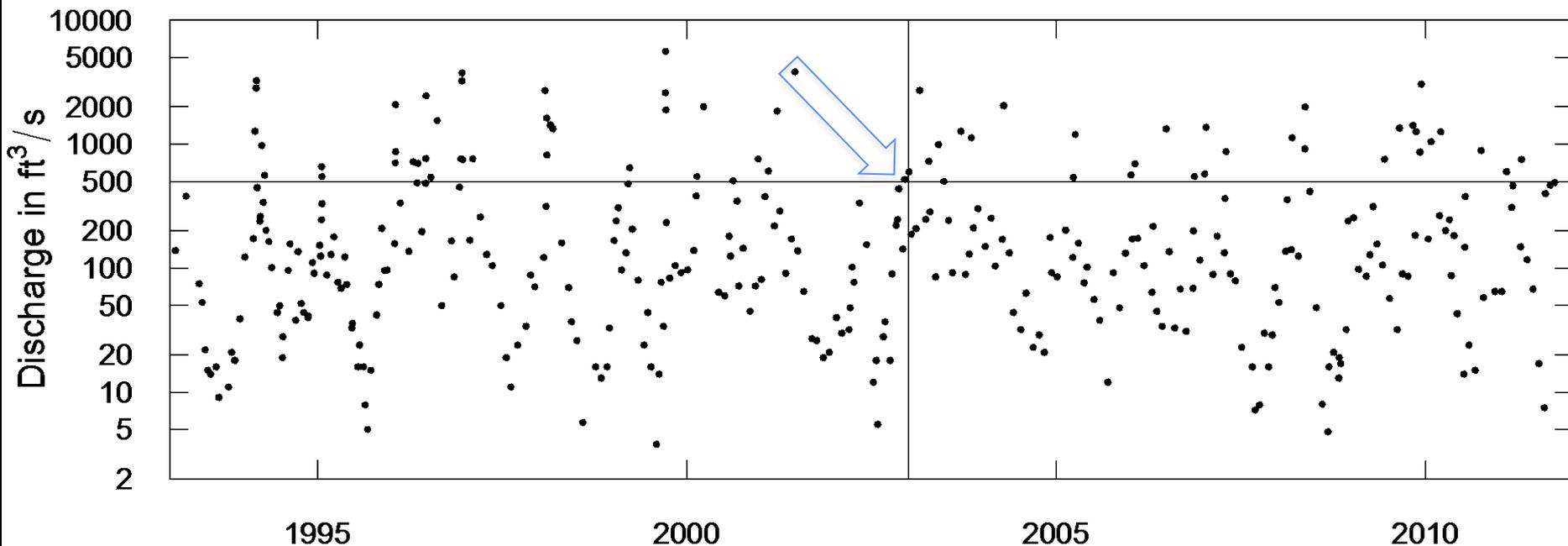
Every dot is a data point from 1993 to 2012
Let's say we want to use the data to estimate the expected value of concentration for January 1, 2003 at Q=500 cfs



The principle is this:

Do a weighted regression at this point. The weights on each observation are related to their “distance”

Choptank River near Greensboro, MD Nitrate plus Nitrite, filtered, as N
Locations of all available data

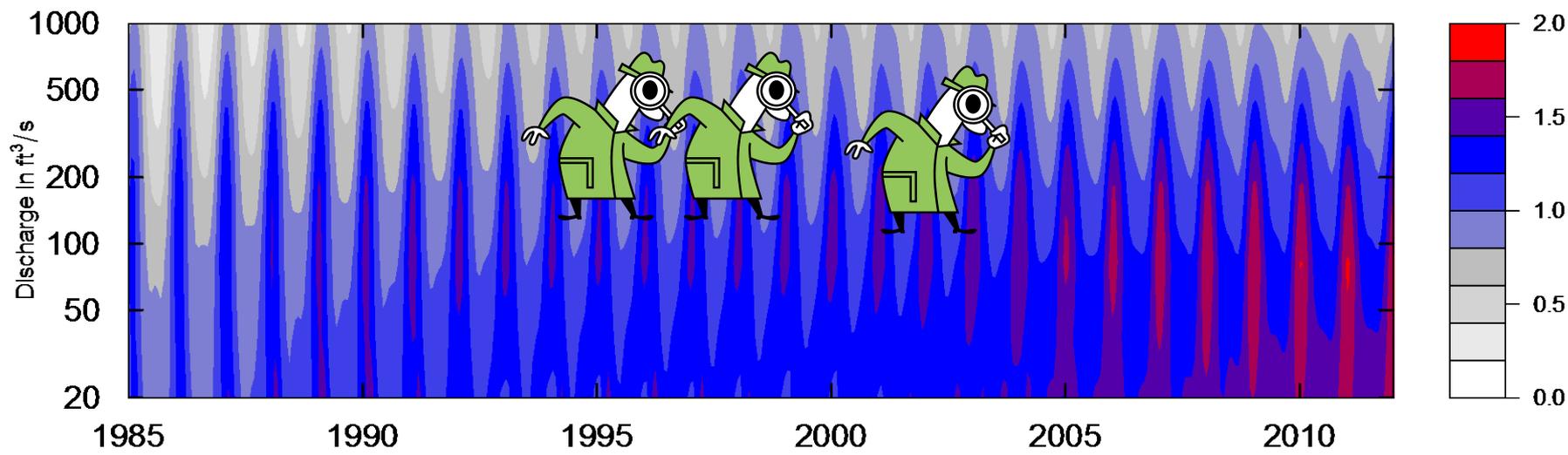


Distance in time, in $\log(Q)$, and season.

Now move to the next point and do it all over again.

This kind of weighted regression gets done about 6000 times to form this whole surface!!

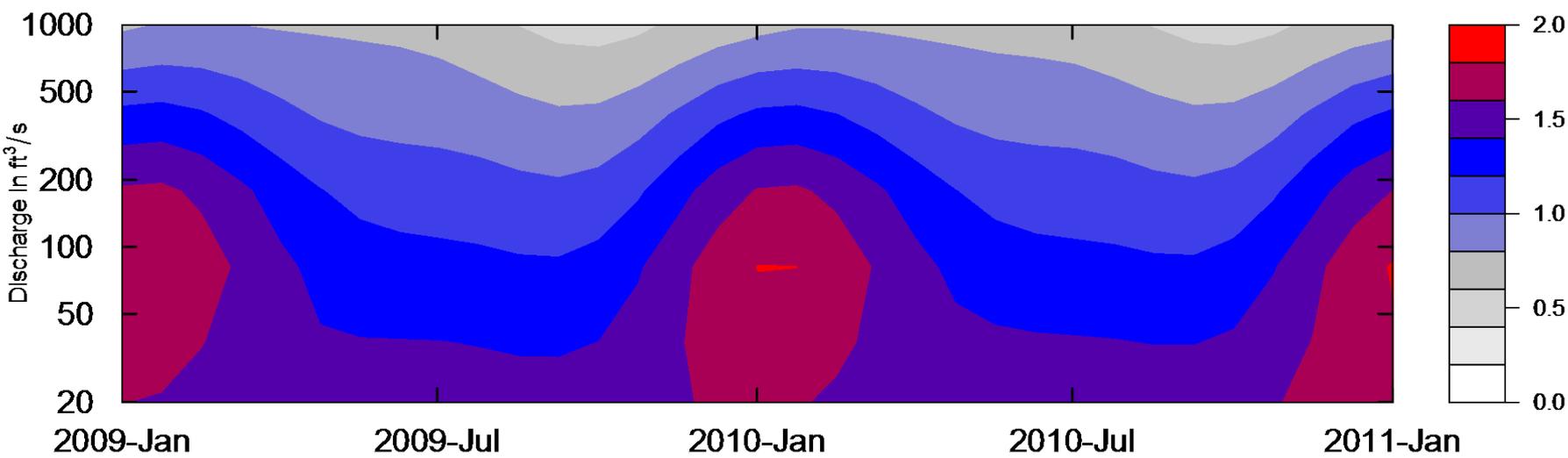
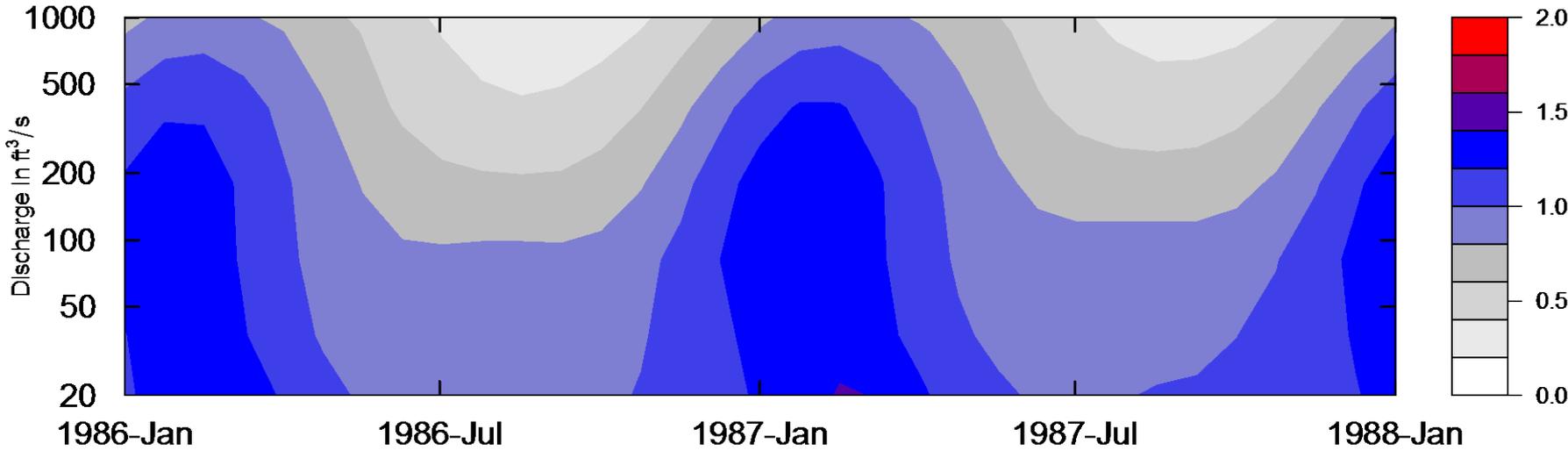
**Choptank River near Greensboro, MD Nitrate plus Nitrite, Filtered, as N
Estimated Concentration Surface in Color**



**You must be kidding. This is a ton of computations!!
That's right! But it's what we need to make order out of chaos.**

Here are two, more detailed looks at this surface

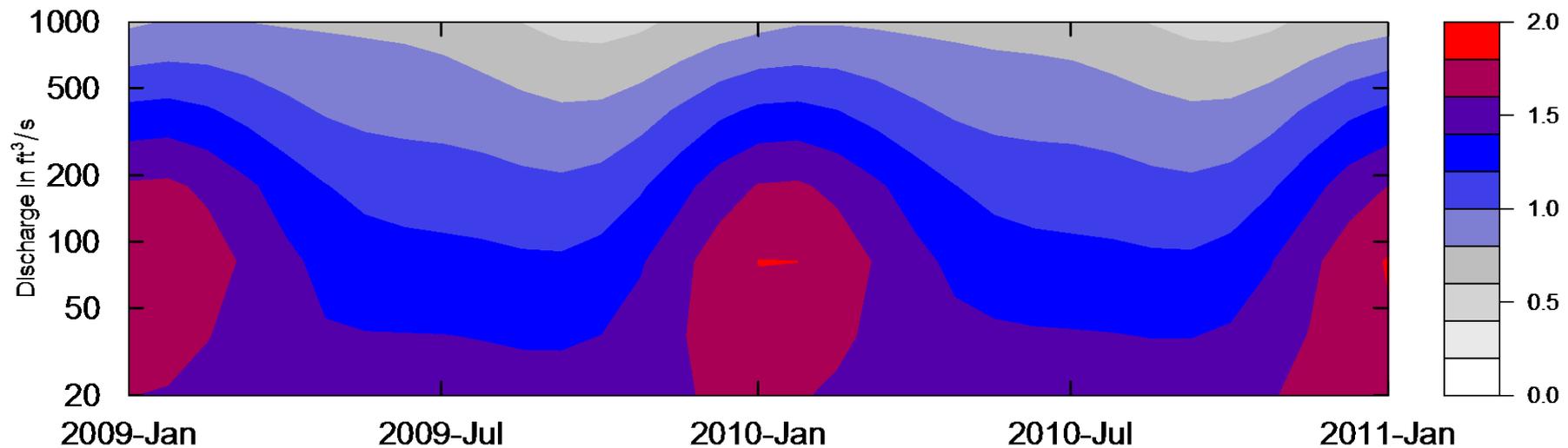
Choptank River near Greensboro, MD Nitrate plus Nitrite, Filtered, as N
Estimated Concentration Surface in Color



Now, for every one of 10,227 days in the record from 1985 through 2012:

We can use the date and the observed discharge to compute the expected value of concentration.

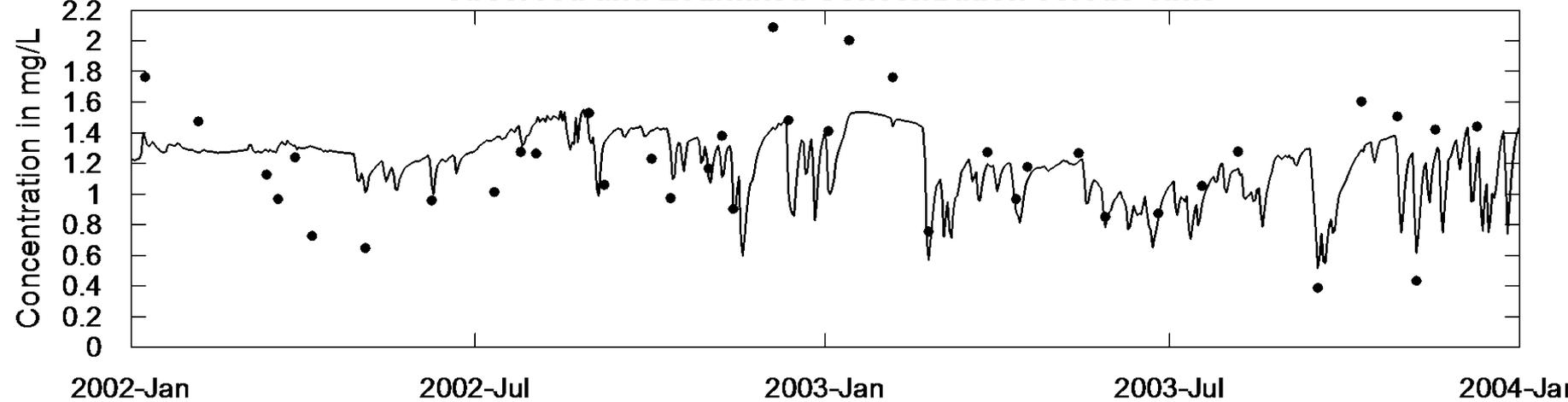
From that value we can compute the expected value of flux.



Then we can sum these estimates by year to compute estimates of annual mean concentration & annual mean flux

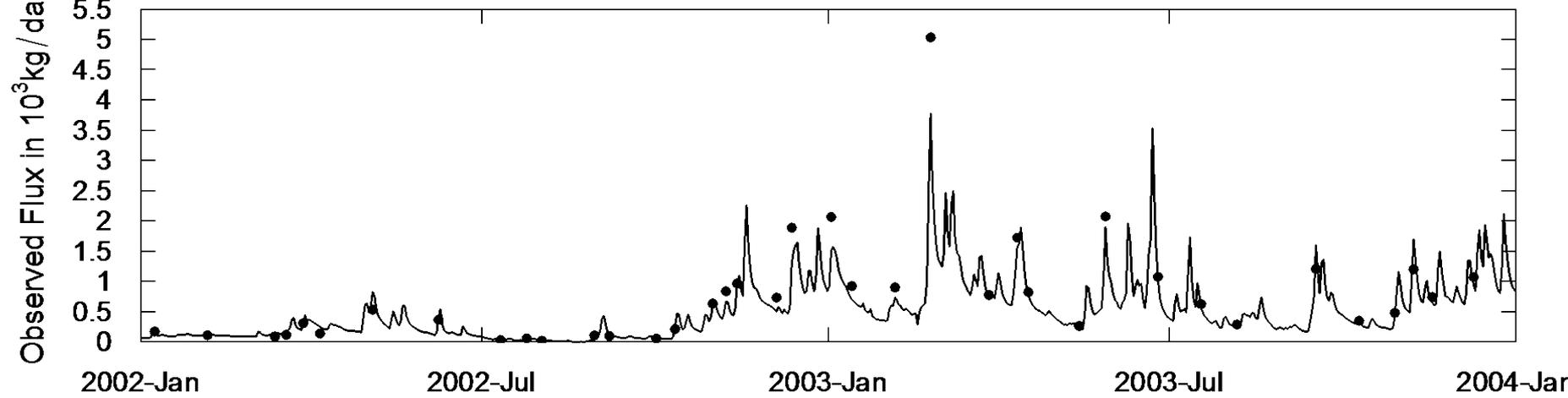
**Choptank River near Greensboro, MD
Nitrate**

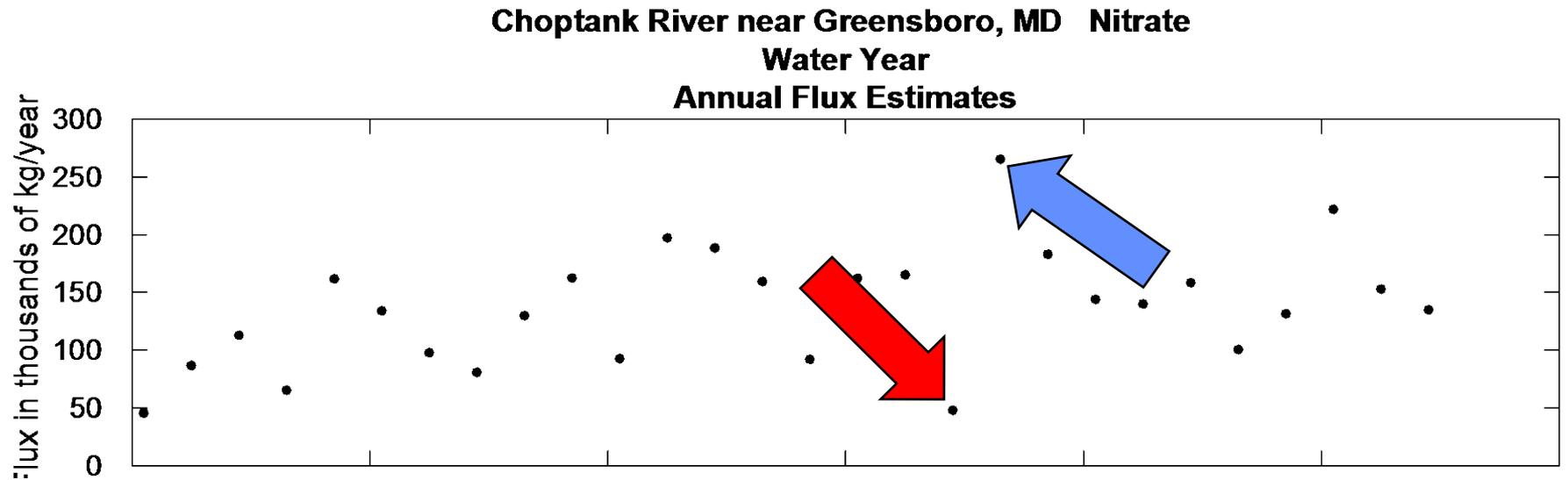
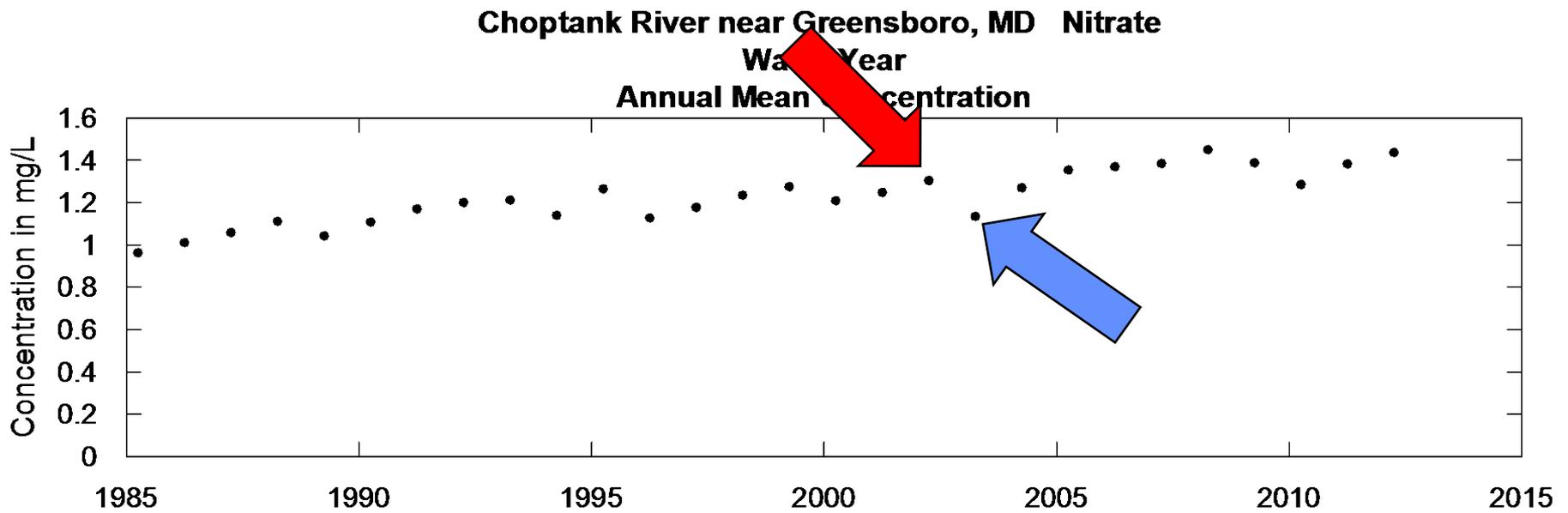
Observed and Estimated Concentration versus Time



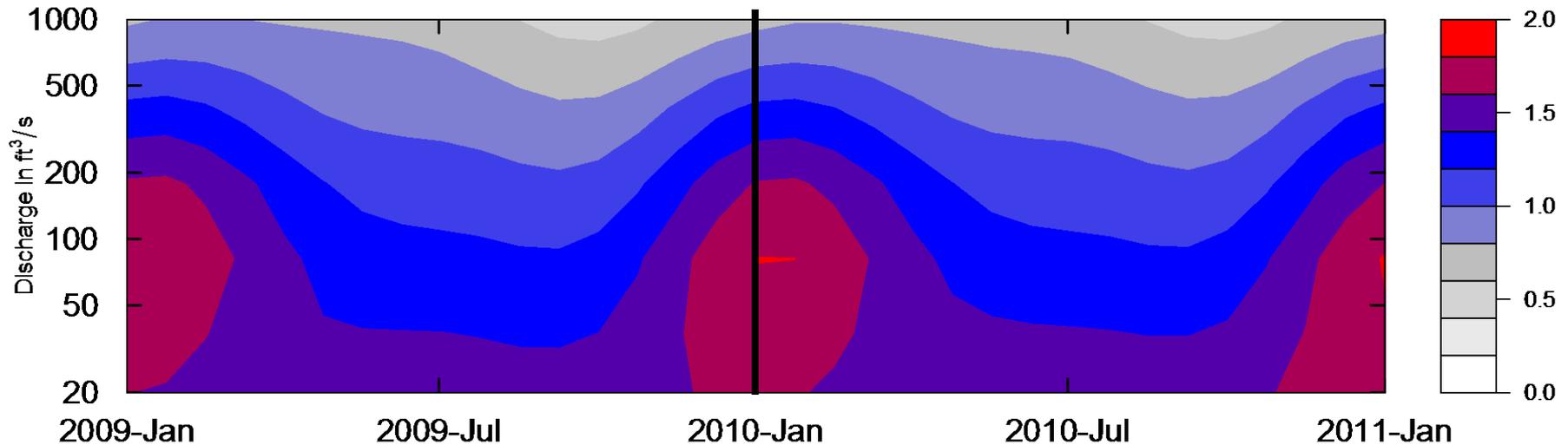
**Choptank River near Greensboro, MD
Nitrate**

Observed and Estimated Flux versus Time





Can we filter out this flow-driven variation to see the underlying change?



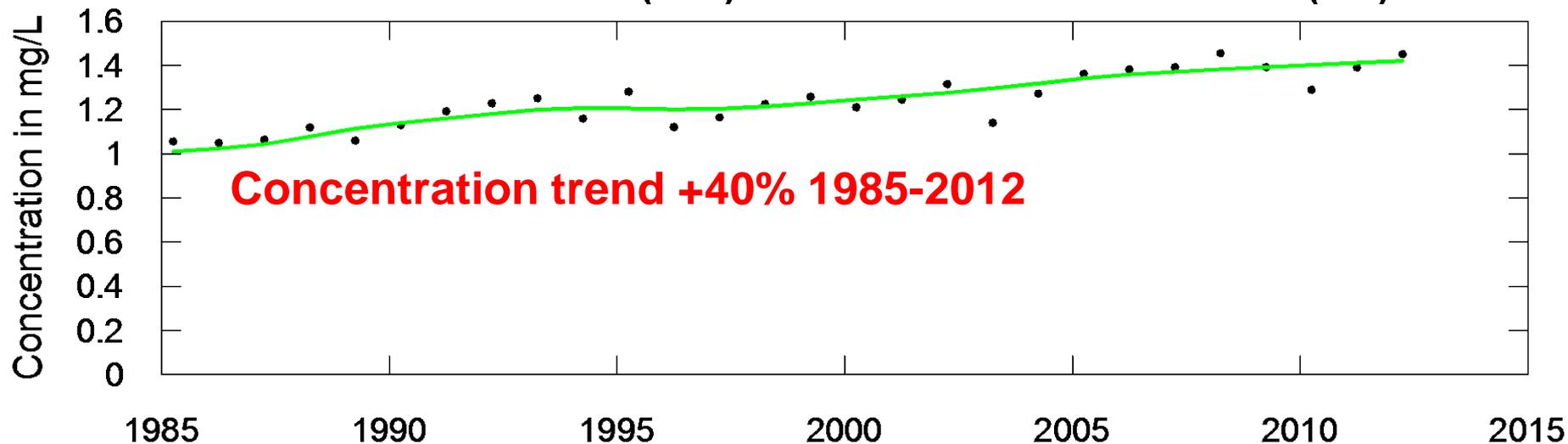
**The “flow normalized concentration” on any given day is:
 $c=f(Q,T)$ integrated over the probability distribution of Q
 for that day of the year.**

**Flow normalized flux is just $c \times Q$ integrated over
 discharge.**

**Sum those over the year to get annual flow-normalized
 mean concentration and flux.**

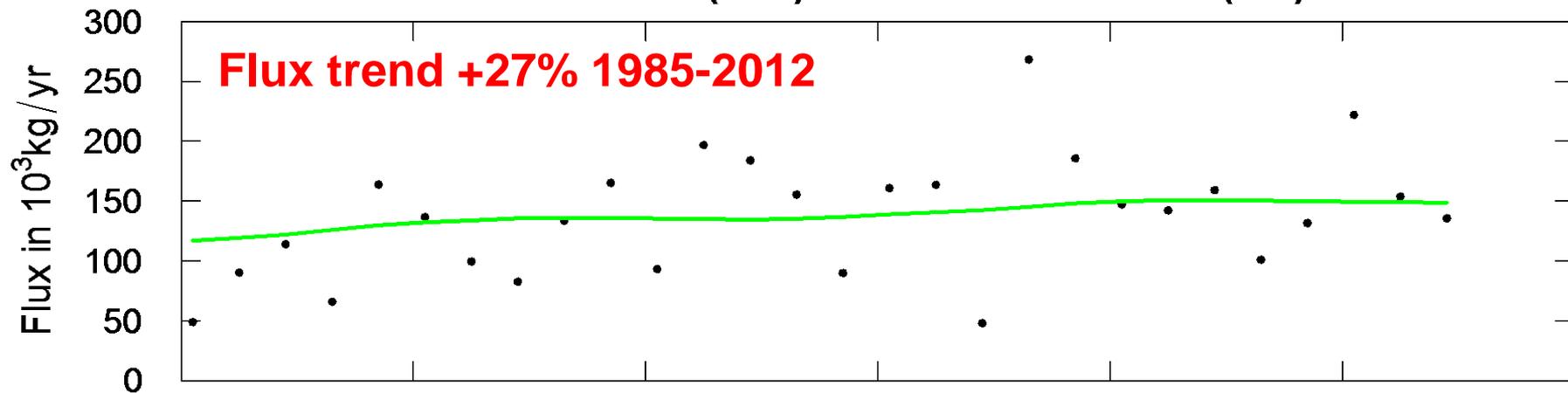
**Choptank River near Greensboro, MD Nitrate
Water Year**

Mean Concentration (dots) & Flow Normalized Concentration (line)



**Choptank River near Greensboro, MD Nitrate
Water Year**

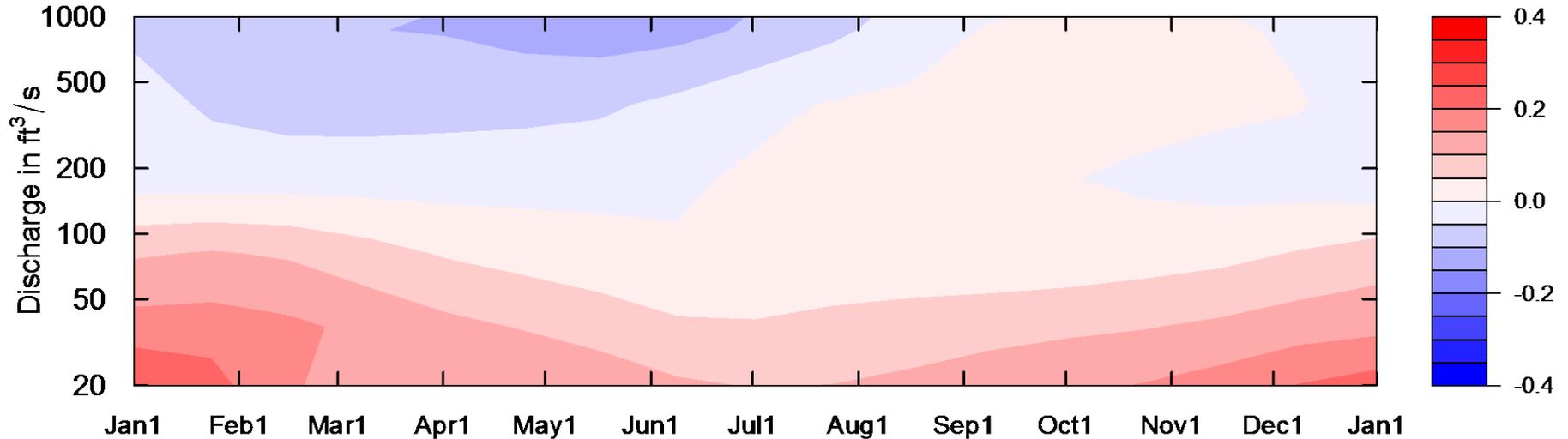
Flux Estimates (dots) & Flow Normalized Flux (line)



Look at changes in just the last few years.

This is a graphic of differences 2007 to 2012

Choptank River near Greensboro, MD Nitrate plus Nitrite, Filtered, as N
Estimated Concentration change from 2007 to 2012



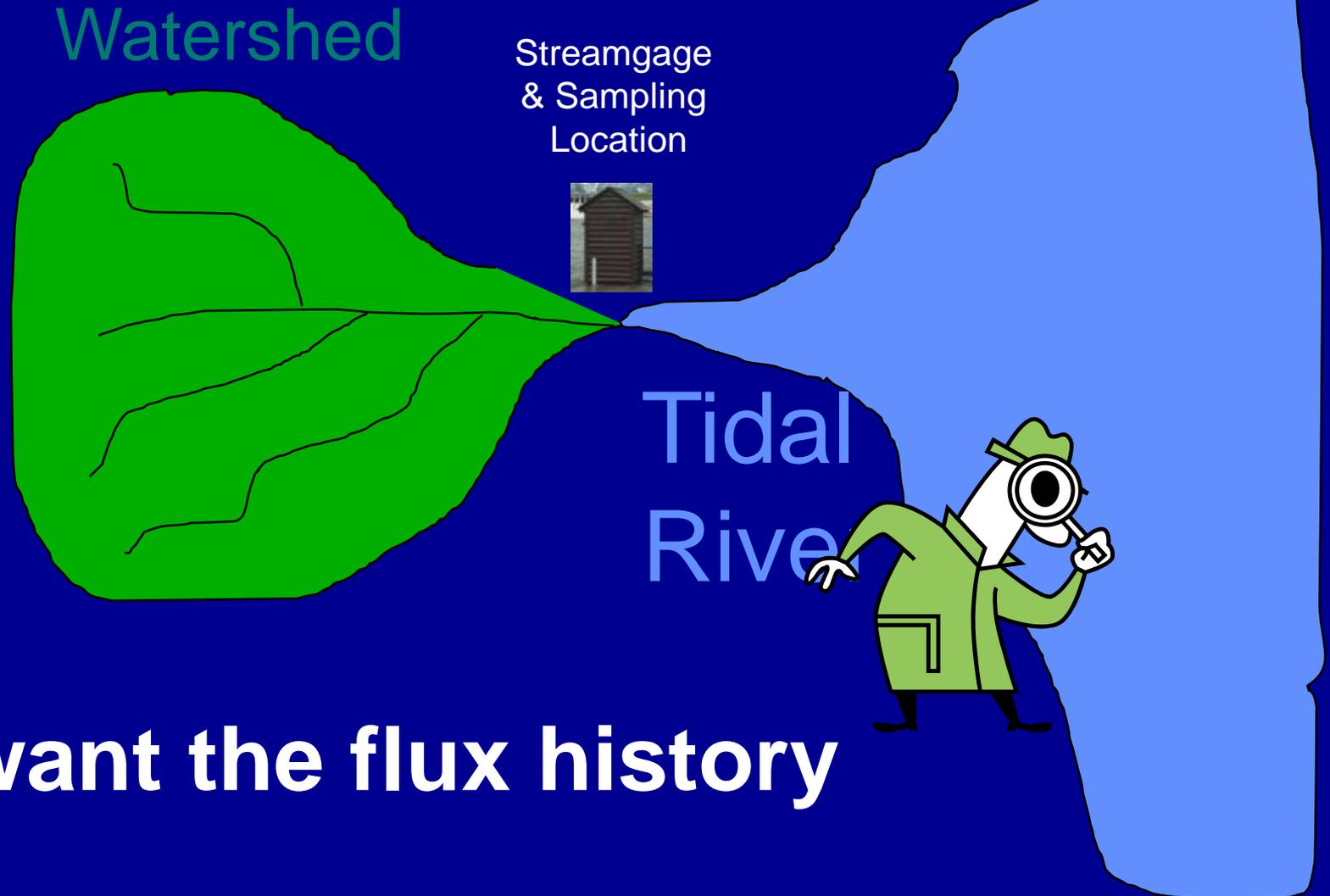
Hypothesis, cover crops are helping at higher flows particularly in the winter. Low flows are still responding to legacy of nitrate enriched groundwater.

Why all this complexity?

Different products for different purposes

- Concentration versus flux
- Actual history versus flow-normalized history

For understanding impact on the estuary ecosystem



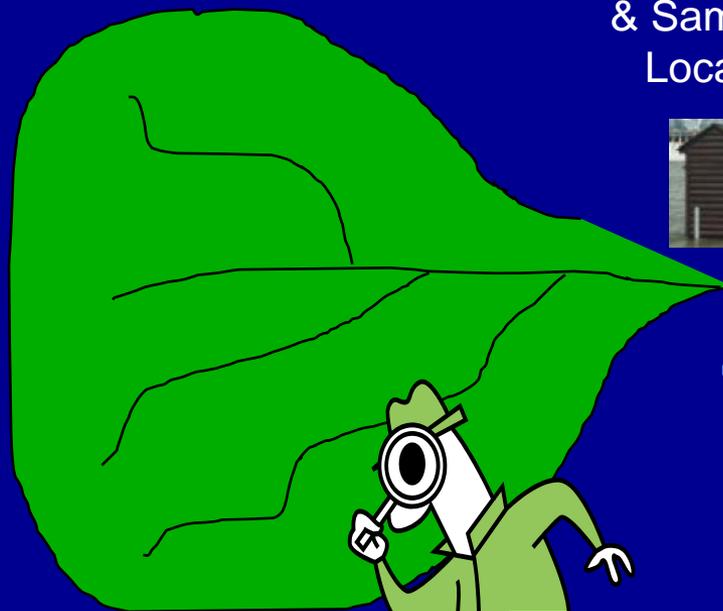
We want the flux history

For understanding
progress in the watershed

Estuary

Watershed

Streamgage
& Sampling
Location



Tidal
River

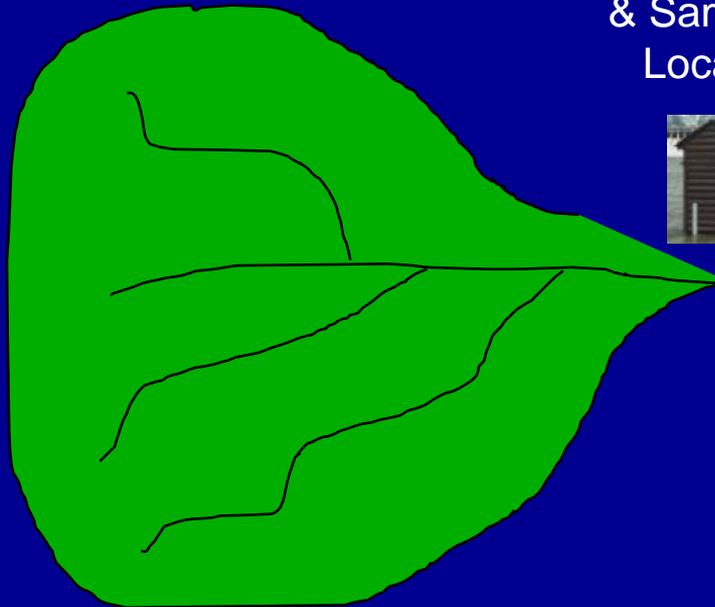
We want the flow-normalized flux history

For understanding the changes in the rivers

Estuary

Watershed

Streamgage
& Sampling
Location

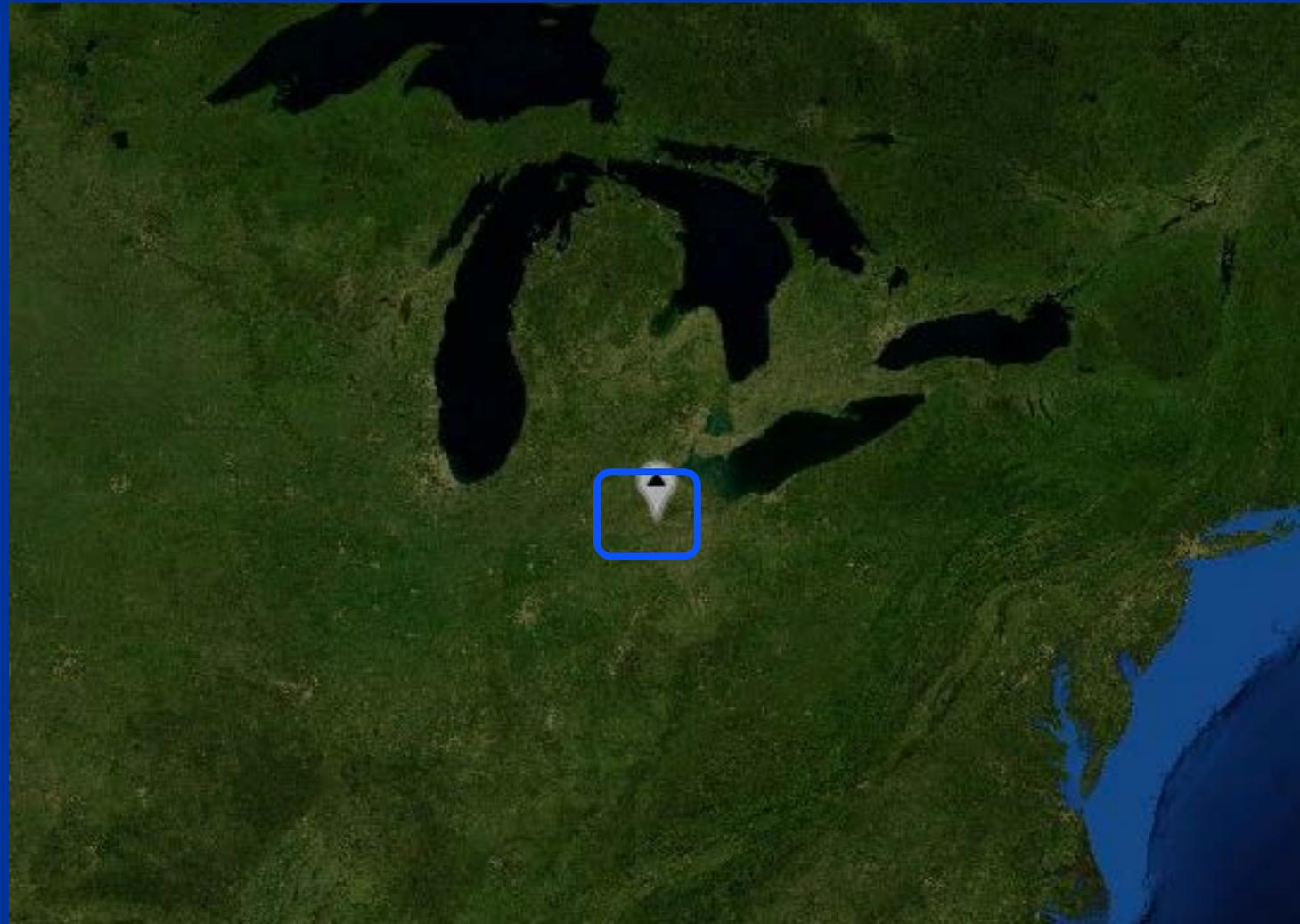


Tidal
River



We want the concentration history

Maumee River – 16,000 km² Tributary to Lake Erie

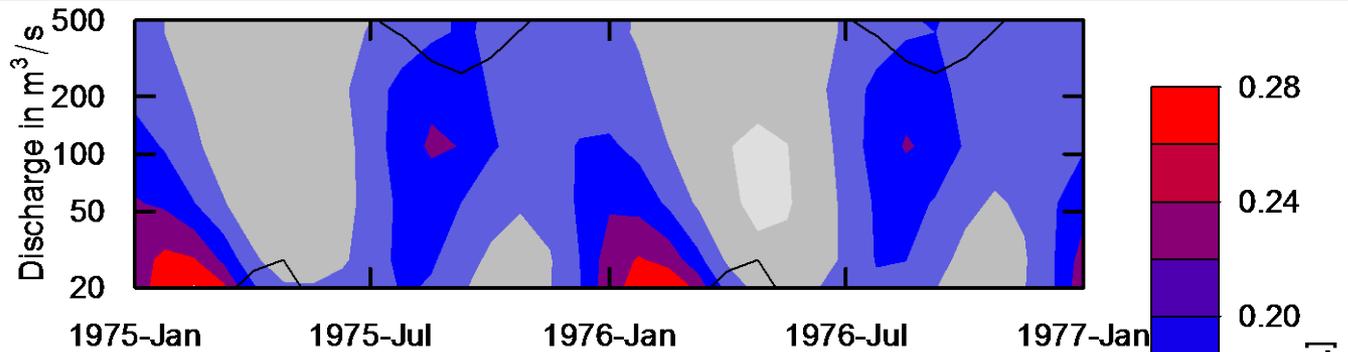


Cyanobacter – Lake Erie

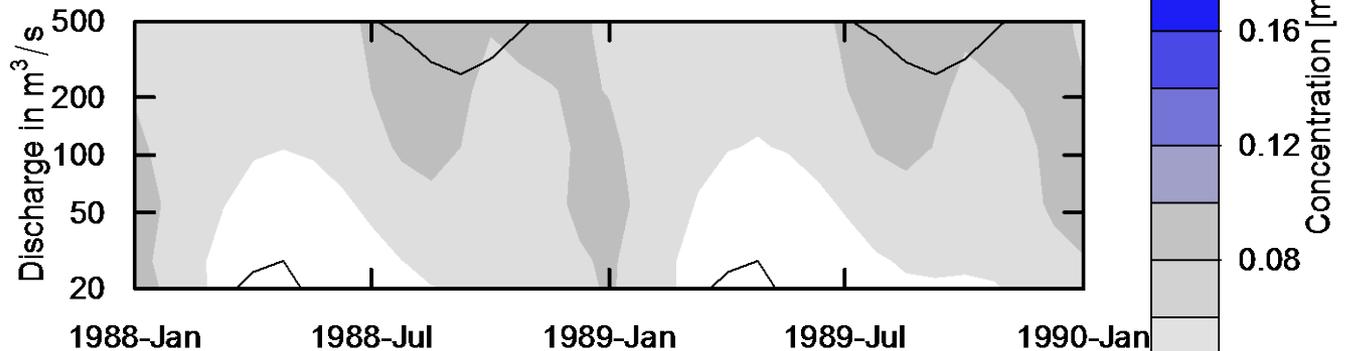


Dissolved Reactive Phosphorus, Maumee River, at Waterville, OH

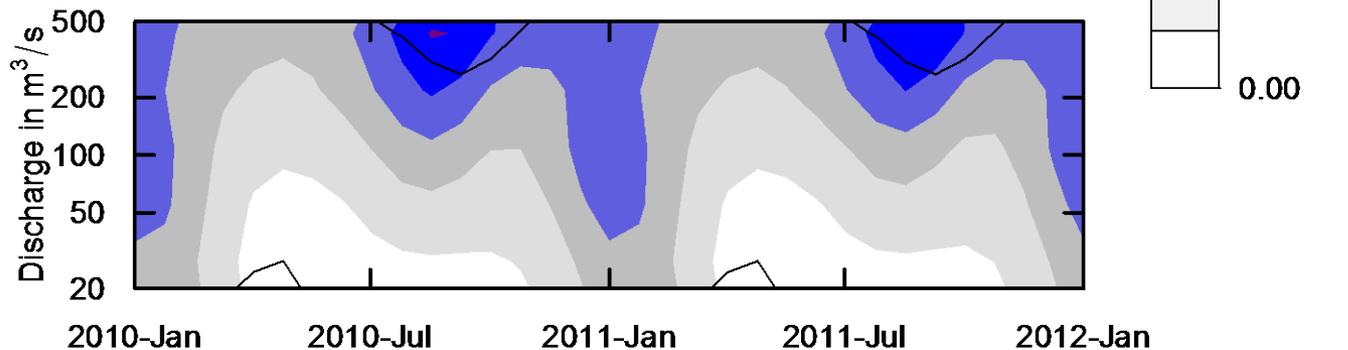
Mid 1970's



Late 1980's



Early 2010's

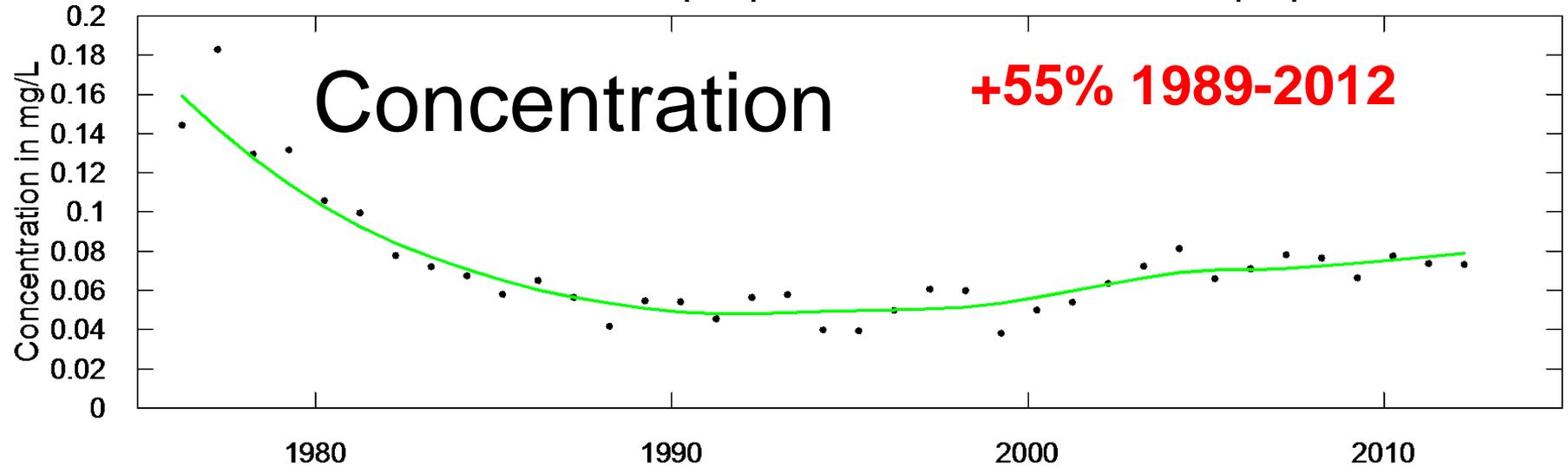


Dissolved Reactive Phosphorus, Maumee River, at Waterville, OH

Maumee River at Waterville OH Dissolved Reactive Phosphorus

Water Year

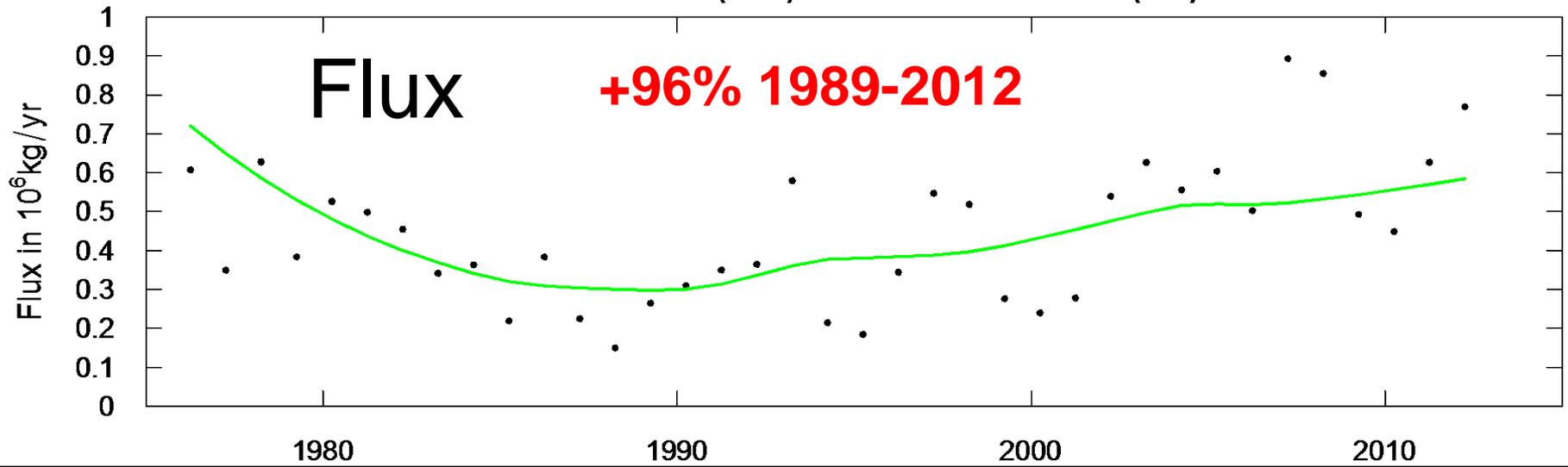
Mean Concentration (dots) & Flow Normalized Concentration (line)



Maumee River at Waterville OH Dissolved Reactive Phosphorus

Water Year

Flux Estimates (dots) & Flow Normalized Flux (line)



The software: how do I get it?

- Need to install R (freely downloaded from <http://cran.us.r-project.org/>) on your computer

- Once you start R, you can load the software:
`install.packages("dataRetrieval", "EGRET")`

`library(dataRetrieval)`

`library(EGRET)`

You are ready to go

check out our new developments at:

<https://github.com/USGS-R/EGRET/wiki>

dataRetrieval

- Brings data from various sources into R
- Organizes it into “data frames”
- From there, users can run summaries, build graphics, and build and evaluate statistical models.
- All the functionality of R is at your fingertips.
- You can save your work and you can share it with others. (by sending your “workspace”)

dataRetrieval functions:

By information source and purpose

Information Source	Site Query	Meta Data	Data
NWIS			
Water Quality Portal			

dataRetrieval functions: By information source and purpose

Information Source	Site Query	Meta Data	Data
NWIS	whatNWISsites whatNWISdata	readNWISsite readNWISpCode	readNWISdata readNWISdv readNWISqw readNWISuv readNWISrating readNWISmeas readNWISpeak readNWISgwl
Water Quality Portal	whatWQPsites		readWQPqw readWQPdata

**A little example of one
of these functions**



Unit values retrieval (not used by EGRET)

- Raccoon River at Van Metre, IA
- Nitrate sensor data
- March – Sept 2013

Bring in all the data

```
Unit<-readNWISuv("05484500",parameterCd=c("99133","00060"),"2013-03-01","2013-09-30")
```

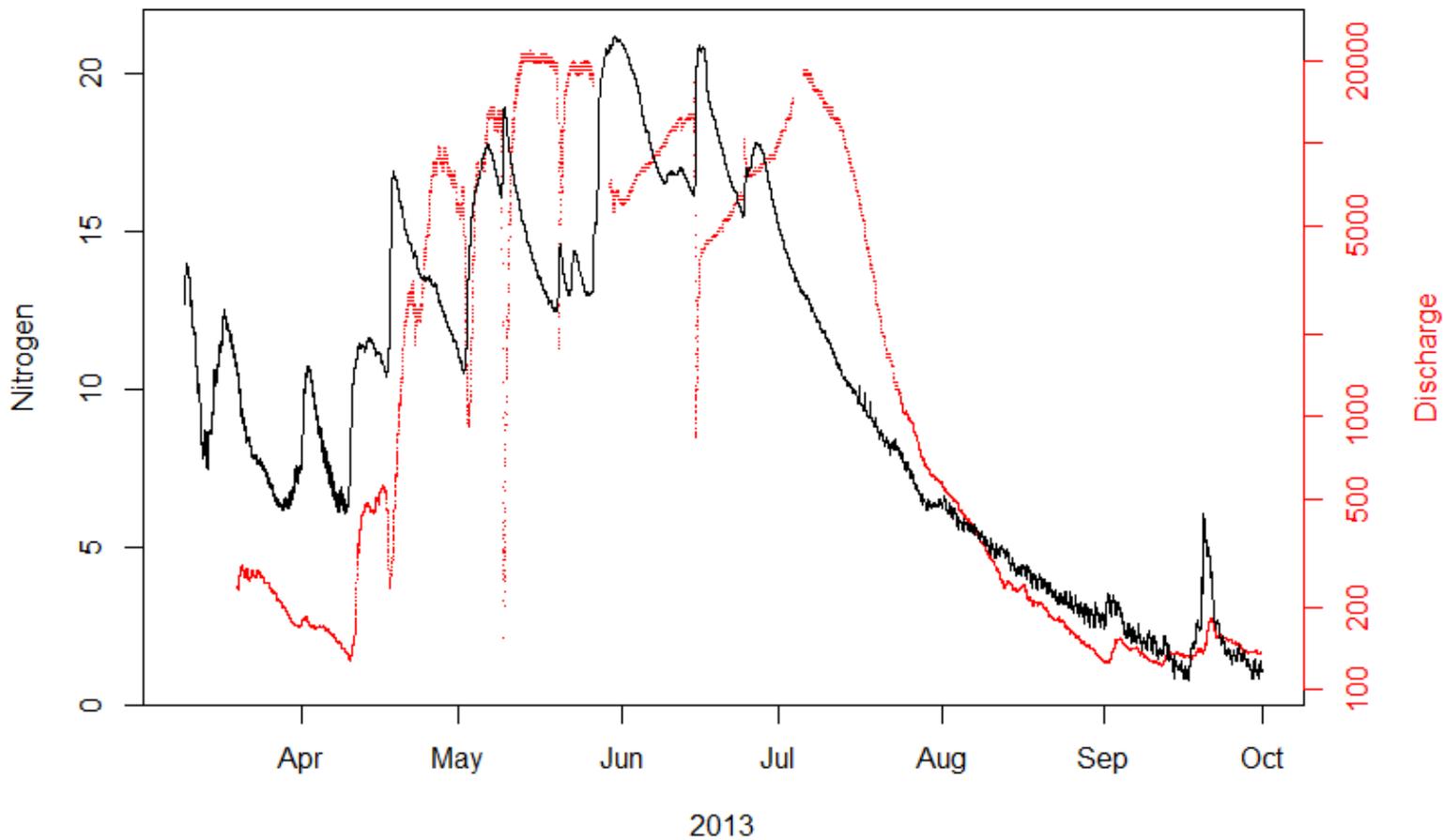
**Creates a data frame of 19,568 time steps,
discharge missing for 86 of them,
nitrate missing for 1,848 of them.**

One more command

```
Unit<-na.omit(Unit)
```

**Now we have a data from of 17,634 rows
all of which have discharge and nitrate**

Many options for graphics and modeling of concentration and discharge together



Using EGRET

- For each session the code needs to be loaded:
library(dataRetrieval)
library(EGRET)
- Once this is done you will have access to **help** and to the **package vignettes**.
- To get help with a function (such as the function readNWISSample) just type ?readNWISSample

How can we enter data?

- **For the water quality sample data**
 - From USGS web services
 - From Water Quality Portal
 - From a user supplied file
- **For the daily discharge data**
 - From USGS web services
 - From a user supplied file
- **For the meta-data**
 - From USGS or Water Quality Portal
 - From user entries

```
> library(dataRetrieval)
> library(EGRET)
> site <- "01491000"
> parameterCd <- "00631"
> startDate <- "1979-10-01"
> endDate <- "2014-09-28"
> Sample <- readNWISSample(site,parameterCd,startDate,endDate)
> summary(Sample)
```

The result: we have created a data frame of 708 rows (one per sample) with columns for:

Date, Concentration, Days since January 1, 1850, Month of the year, Day of the year, Decimal year, sine and cosine of time of year, and censoring information.

```
Daily <- readNWISDaily(site,"00060",startDate,endDate)
```

The result: we have created a data frame of 12,782 rows (one per day) with columns for:

Date, Discharge, Days since January 1, 1850,

Month of the year, Day of the year, Decimal year,
mean flow for past 7 days, mean flow for past 30
days

Storing the metadata

- For NWIS data

```
INFO<-readNWISInfo(site,parameterCd)
```

- Similar function for the Water Quality Portal
- The contents of INFO are used to label tables and figures as well as document the site and constituent information
- Creates a system of abbreviations to keep track of **workspace** files

Two more commands before we can start our analysis of the data

```
> eList <- mergeReport(INFO,Daily,Sample)
```

```
> eList <- mergeReport(INFO,Daily,Sample)
```

Discharge Record is 12782 days long, which is 35 years

First day of the discharge record is 1979-10-01 and last day is 2014-09-28

The water quality record has 708 samples

The first sample is from 1979-10-24 and the last sample is from 2014-08-13

Discharge: Minimum, mean and maximum 0.00991 4.17 246

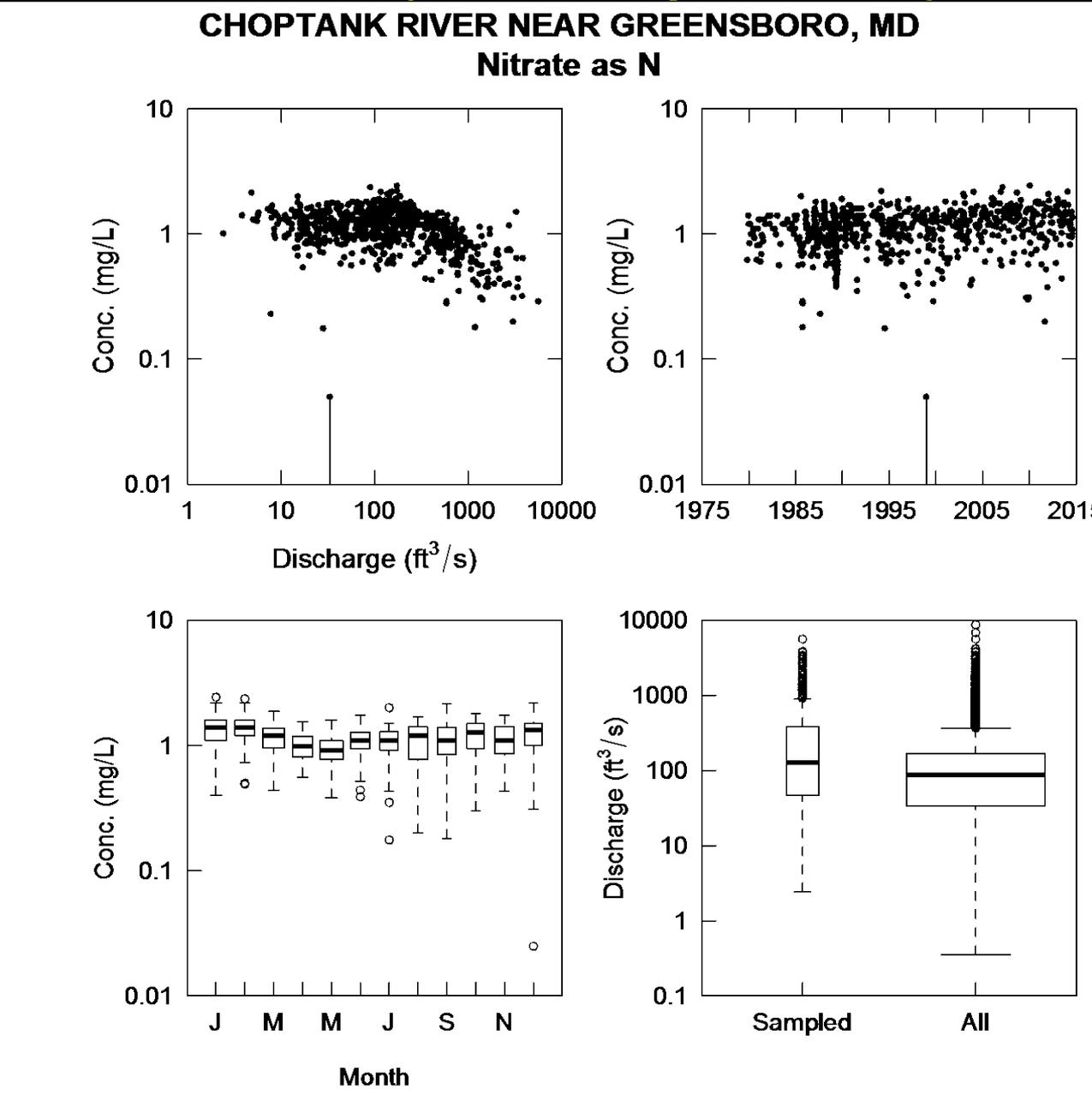
Concentration: Minimum, mean and maximum 0.05 1.1 2.4

Percentage of the sample values that are censored is 0.14 %

Now, look at your data.

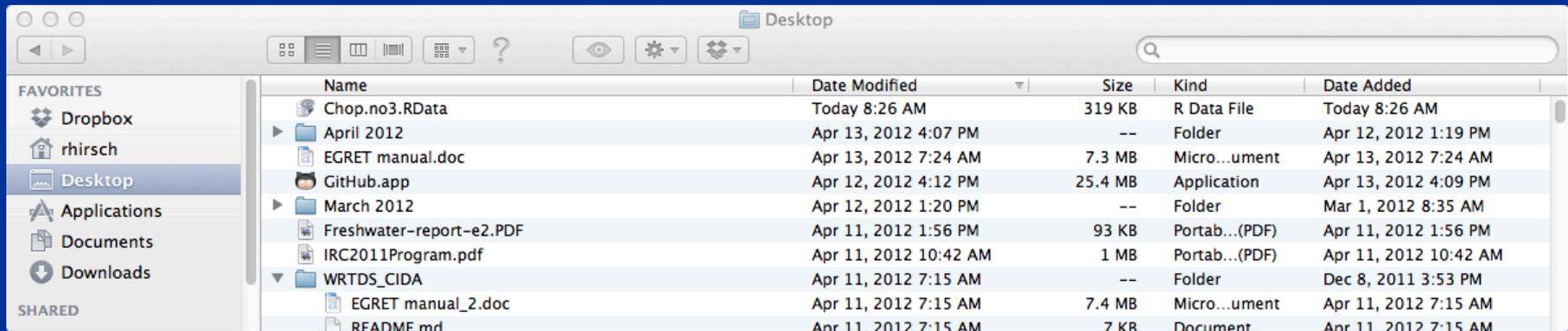
No excuses!!

> multiPlotDataOverview(eList, qUnit=1)



We've gone to all this effort, let's save our work

```
> savePath<-"~/Users/rhirsch/Desktop/"  
> saveResults(savePath,eList)
```



Chop.no3.RData

Save it over and over as
you proceed and add
results



We now have 3 data frames, bound together in eList

- Sample (708 rows, 14 columns)**
- Daily (12,782 rows, 12 columns)**
- INFO (1 row, 53 columns)**

> **modelEstimation(eList)**

- Runs the model in cross-validation mode
- Estimates the “surface” for concentration as a function of time and discharge
- Uses the surface to compute daily values of
 - Concentration
 - Flux
 - Flow-normalized concentration
 - Flow-normalized flux
- Adds those to the Daily data frame

User has choices about some of the parameters of the WRTDS model

Now what is in Daily?

It now has dimensions (12782, 19)

It has added columns for **daily estimates of:**

the log of concentration,

the standard error of the log of concentration,

the concentration,

the flux,

the flow-normalized concentration,

flow-normalized flux

“Period of Analysis” concept in EGRET.

- Could be water year
- Could be calendar year
- Could be April-May-June
- Could be Dec-Jan-Feb-Mar
- Could be only May...

paStart = calendar month that starts Period

paLong = length of Period, in months

Period of analysis set up

Say we want calendar year

```
eList <- setPA(eList,paStart = 1, paLong=12)
```

Say we want April, May, June

```
eList <- setPA(eList,paStart = 4, paLong = 3)
```

Default is water year

Units in EGRET

Everything stored as:

m^3/s , kg/day , or mg/L

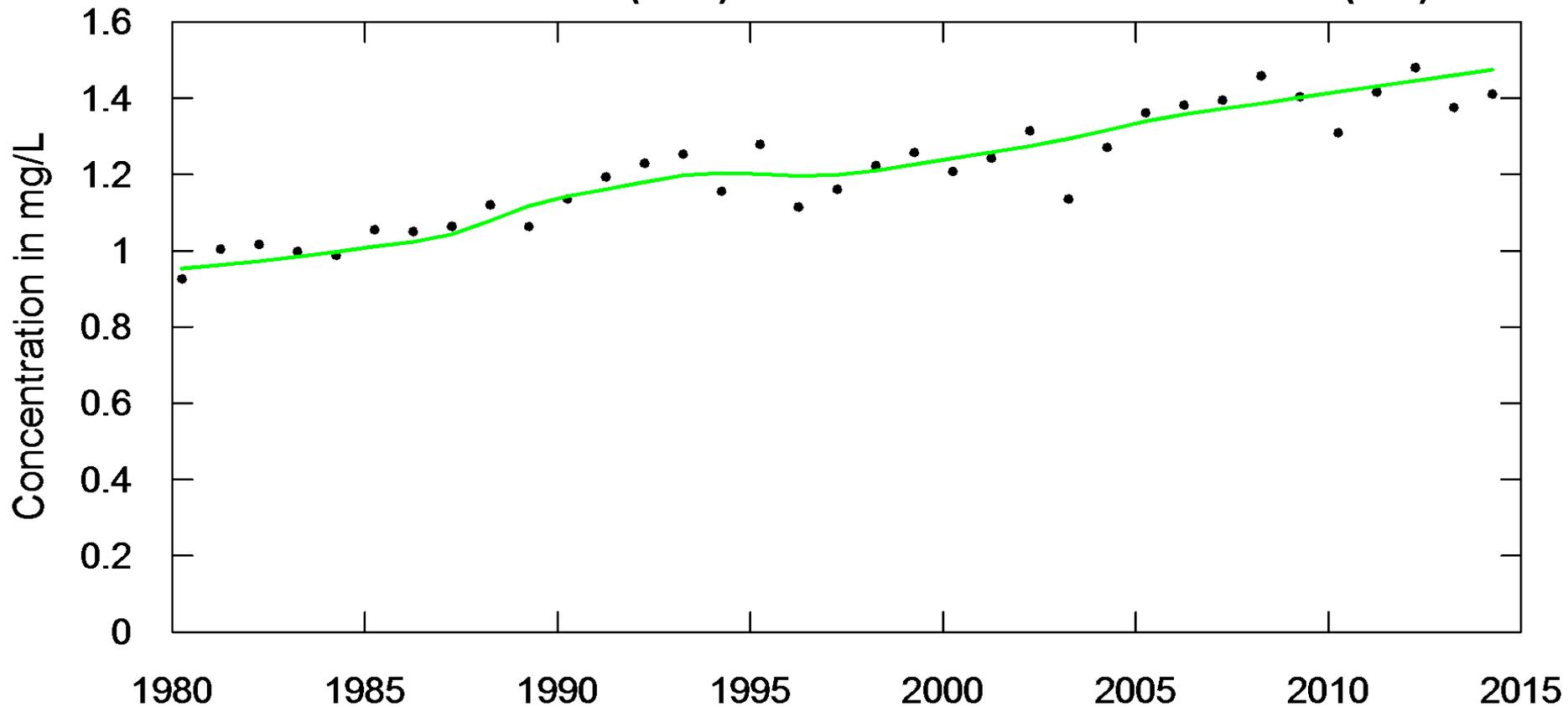
But each graphic or table has a wide choice of units (English and SI) that the user can select

Now lets see some trend results

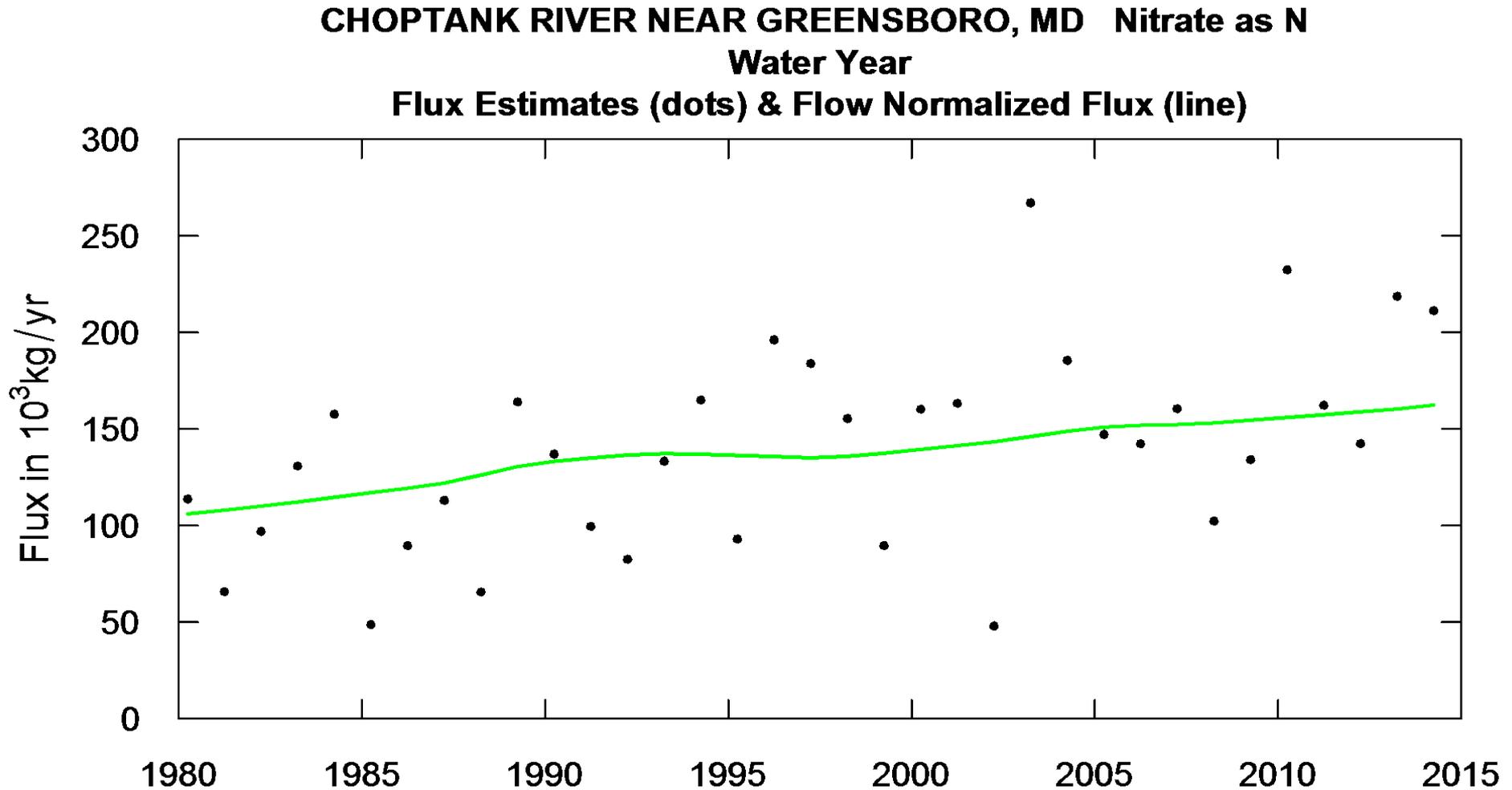
```
> plotConcHist(eList)
```

**CHOPTANK RIVER NEAR GREENSBORO, MD Nitrate as N
Water Year**

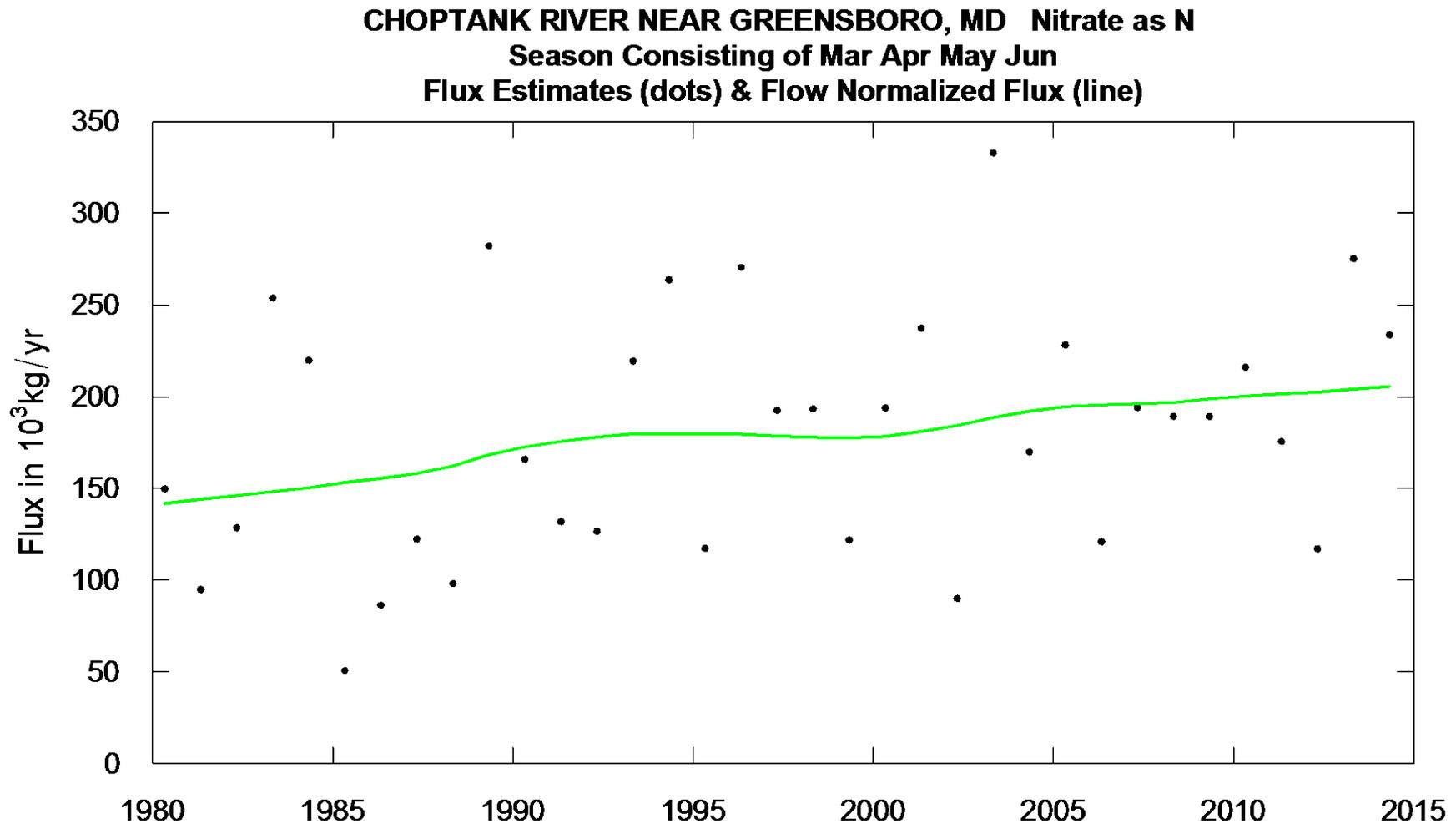
Mean Concentration (dots) & Flow Normalized Concentration (line)



```
> plotFluxHist(eList,fluxUnit=8)
```



```
> eList <- setPA(paStart=3,paLong=4)
> plotFluxHist(fluxUnit=8)
```



```
> tableResults(eList, qUnit = 1, fluxUnit = 5)
```

CHOPTANK RIVER NEAR GREENSBORO, MD

Nitrate as N

Water Year

Year	Discharge cfs	Conc	FN_Conc mg/L	Flux tons/yr	FN_Flux
1980	150.2	0.926	0.953	125.5	117
1981	78.3	1.004	0.963	72.6	119
1982	107.6	1.017	0.972	107.0	121
1983	176.1	0.998	0.984	144.4	124
1984	201.9	0.988	0.997	173.9	126
1985	53.6	1.055	1.011	53.8	129
1986	92.8	1.050	1.023	98.9	132
1987	119.1	1.064	1.043	124.7	135
1988	66.0	1.121	1.079	72.4	139
.					
.					
.					
2007	151.2	1.395	1.373	177.1	168
2008	90.5	1.459	1.386	112.8	169
2009	130.0	1.404	1.402	147.9	170
2010	254.0	1.310	1.417	256.4	172
2011	185.2	1.417	1.431	179.0	174
2012	122.6	1.480	1.445	157.1	175
2013	226.0	1.376	1.460	241.1	177
2014	191.8	1.411	1.475	233.0	179



```
> tableChange(eList, fluxUnit=5, yearPoints=c(1980,1995,2014))
```

CHOPTANK RIVER NEAR GREENSBORO, MD
Nitrate as N
Water Year

Concentration trends

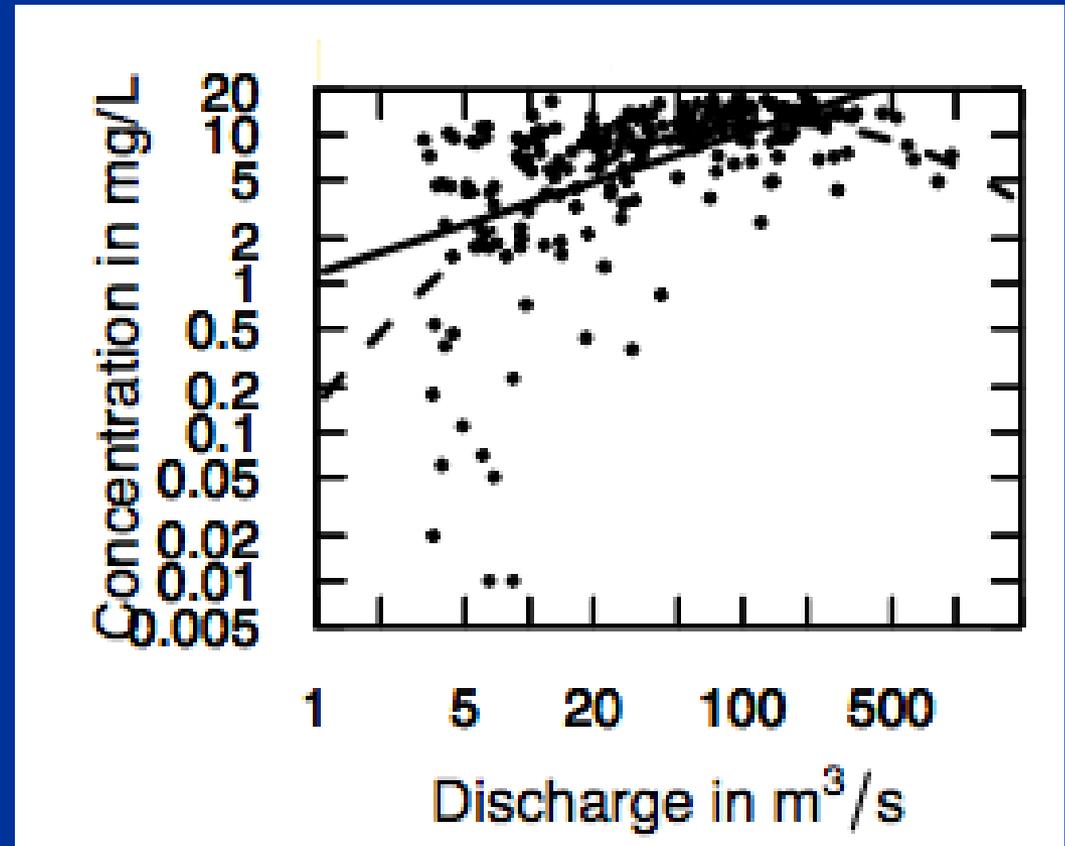
time span			change mg/L	slope mg/L/yr	change %	slope %/yr
1980	to	1995	0.25	0.017	26	1.7
1980	to	2014	0.52	0.015	55	1.6
1995	to	2014	0.27	0.014	23	1.2

Flux Trends

time span			change tons/yr	slope tons/yr /yr	change %	slope %/yr
1980	to	1995	33	2.2	29	1.9
1980	to	2014	62	1.8	53	1.6
1995	to	2014	29	1.5	19	1



I'm going to switch data sets to Nitrate for the Raccoon River at Des Moines Iowa

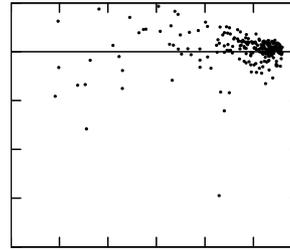


EGRET

produces a
diagnostic
plot to help
spot
serious
problems
with the
model

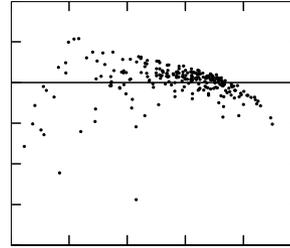
fluxBiasMulti(eList,
fluxUnit=4)

trate
Model is WRTDS Flux Bias Statistic -0.00237



This same
type of plot
can be
used to
look at
other
models,
here the
LOADEST7

n River at Des Moines, IA Nitrate
Model is L7 Flux Bias Statistic 0.319



Diagnostics and potential problems with estimating mean flux, see:

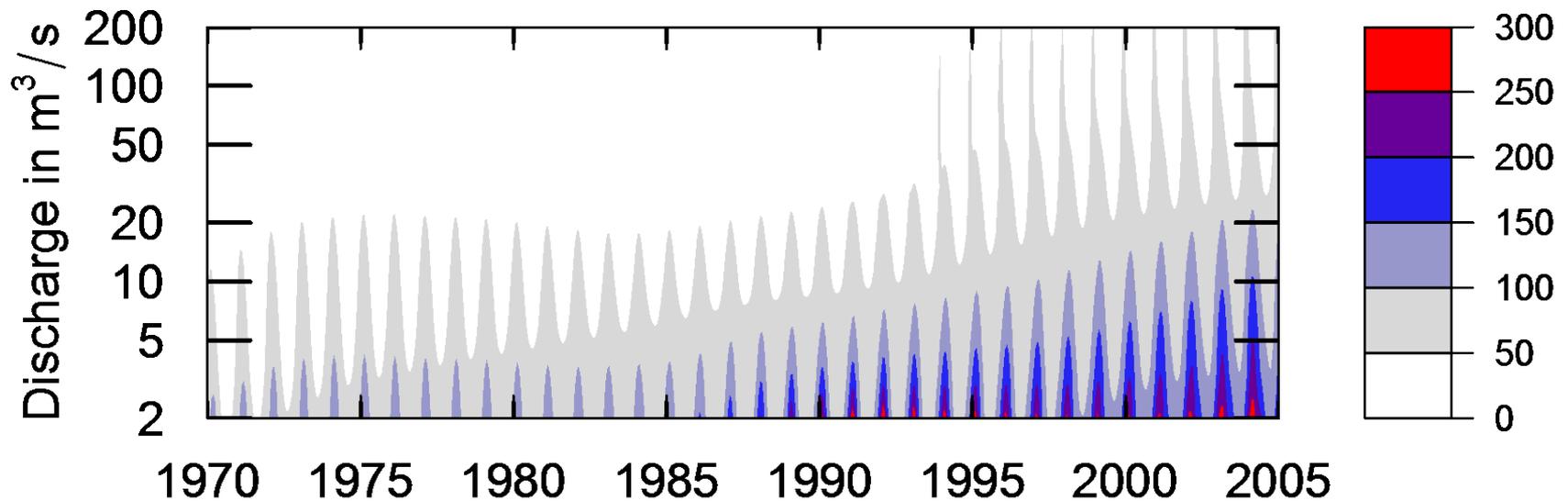
Hirsch, R.M., 2014, Large biases in regression-based constituent flux estimates: causes and diagnostics. Journal of the American Water Resources Association.

Bottom line, look at the fit before you use a statistical model!!!

How difficult is it to make those contour plots?

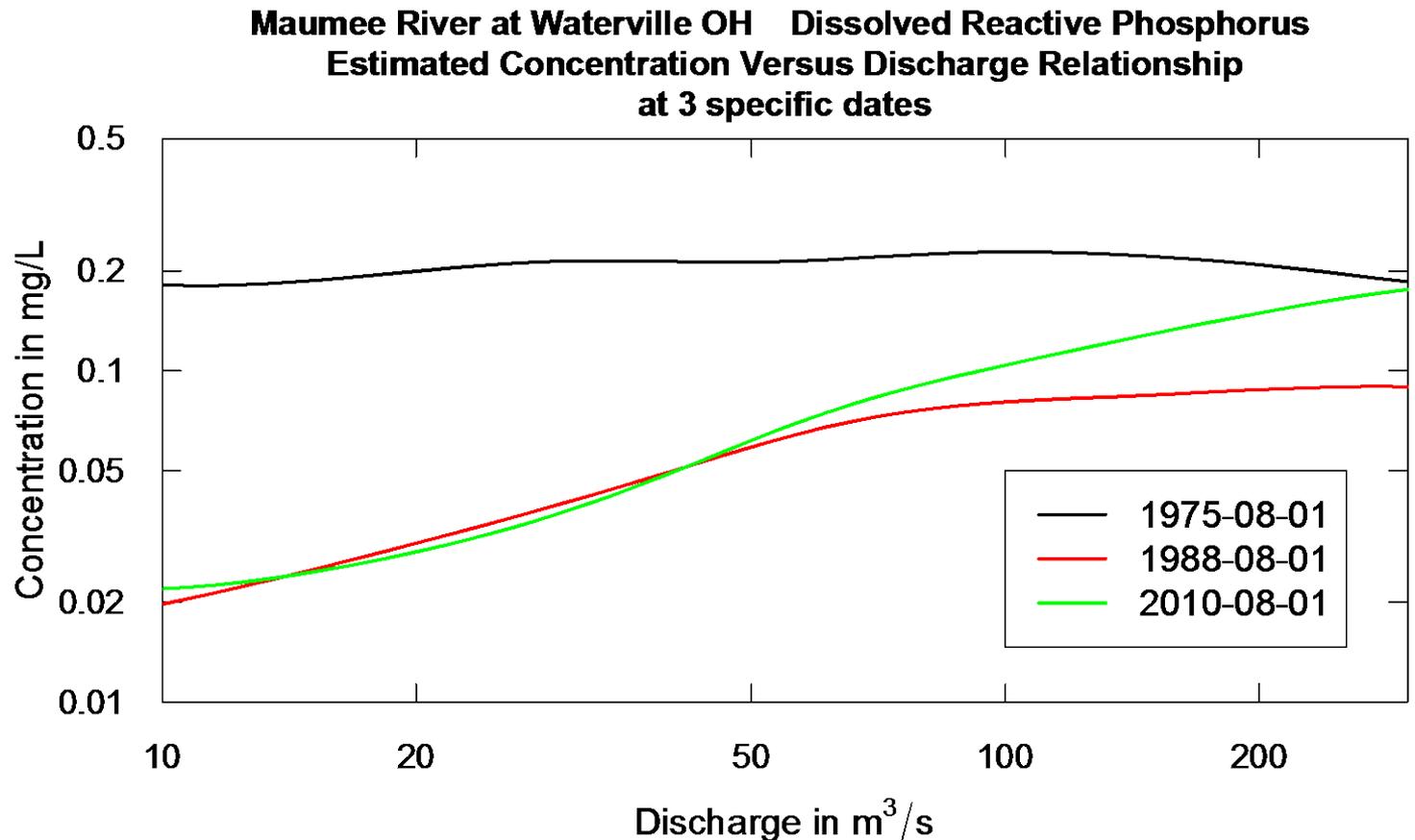
```
>plotContours(eList,yearStart=1970, yearEnd=2005,  
qBottom=2, qTop=200, qUnit=2,  
contourLevels=seq(0,300,50))
```

**Milwaukee River at Milwaukee, WI Chloride
Estimated Concentration Surface in Color**



There are many more graphics, for example

```
> plotConcQSmooth(eList,"1975-08-01", "1988-08-01", "2010-08-01",  
qLow=10, qHigh=300, qUnit=2, logScale=TRUE, legendLeft=100,  
legendTop=0.05)
```



Anticipated enhancements

- Significance levels and confidence intervals for trends (in review)
- Dealing with ephemeral streams
- Estimation of trends in frequency of exceedances of threshold values
- Dealing with nonstationarity in Q
- Improved estimates of yearly fluxes
- *Users ideas?*

dataDelivery and EGRET

• <https://github.com/USGS-R/EGRET/wiki>

“The only way to figure out what is happening to our planet is to measure it,

and this means tracking changes decade after decade,

and poring over the records.”

“Models without data are fantasy, but data without models are chaos”

