Introduction to the EGRETci package

By Robert M. Hirsch, Stacey A. Archfield, and Laura A. De Cicco

March 9, 2015

Contents

Figures

Tables

1 Introduction to EGRET Confidence Intervals

This package EGRETci is an early test version of the package that will encompass the various approaches to uncertainty estimation associated with trend analysis for the EGRET package. The various functions included here are all discussed in the draft manuscript '*A bootstrap method for estimating uncertainty of water quality trends*' by Hirsch, Archfield, and De Cicco (draft version of March 5, 2015). The present version of the EGRETci package is designed for interactive use only, but later versions will describe how batch workflows can be created. Also, the current version can only do analysis of trends for water years, but later versions will allow it to evaluate any group of months (i.e. any 'Period of Analysis' to use the term in the EGRET package). The EGRETci package is designed to carry out three specific types of tasks.

1) Evaluate a trend over a specific span of years and produce a variety of tabular results. This is done with a short workflow involving four functions: `trendSetUp`, `setPA`, `setForBoot`, and `wBT`. The results come both in the form of a) console output, which shows the bootstrap replicate process as it is underway and provides text output of results, and b) a list of outputs called eBoot. The contents of eBoot are described below.

2) Prepare histograms of values for the trend magnitudes, expressed in percent, for flow-normalized concentration and flow-normalized flux. This is done with the function `plotHistogramTrend`. It depends on outputs contained in eBoot.

3) Create confidence bands around the computed trends in flow-normalized concentration and flow-normalized flux. This is done using a single interactive function called xxxxx and then, using the output from that function running two functions that produce the confidence band graphics for concentration and flux respectively (`plotConcHistBoot`, and `plotFluxHistBoot`).

Table 1. This table defines the contents of the eBoot list

Data Frame	Column	Definition
bootOut	rejectC	Reject Ho, (no trend in concentration), TRUE or FALSE
	pValC	p-value for no trend in concentration
	estC	standard WRTDS estimate of change from starting year to ending year in mg/L
	lowC	Lower confidence limit (90%) on concentration trend
	upC	Upper confidence limit (90%) on concentration trend
	lowC50	Lower confidence limit (50%) on concentration trend
	upC50	Upper confidence limit (50%) on concentration trend
	lowC95	Lower confidence limit (95%) on concentration trend
	upC95	Upper confidence limit (95%) on concentration trend
	likeCUp	Likelihood that trend in concentration is upwards
	likeCDown	Likelihood that trend in concentration is downwards
	rejectF	Reject Ho, (no trend in flux), TRUE or FALSE
	pValF	p-value for no trend in flux
	estF	standard WRTDS estimate of change from starting year to ending year, in 10^6 kg/yr
	lowF	Lower confidence limit (90%) on flux trend
	upF	Upper confidence limit (90%) on flux trend
	lowF50	Lower confidence limit (50%) on flux trend
	upF50	Upper confidence limit (50%) on flux trend
	lowF95	Lower confidence limit (95%) on flux trend
	upF95	Upper confidence limit (95%) on flux trend
	likeFUp	Likelihood that trend in flux is upwards
	likeFDown	Likelihood that trend in flux is downwards
	baseConc	The estimated mean flow-normalized concentration in starting year, in mg/L
	baseFlux	The estimated mean flow-normalized flux in starting year, in 10^6 kg/yr
	iBoot	The actual number of bootstrap replicates used
wordsOut		a vector of four character variables (self explanatory)
xConc		a vector of length iBoot, of the change in flow normalized concentration computed by each bootstrap replicate (mg/L)
xFlux		a vector of length iBoot, of the change in flow normalized flux computed by each bootstrap replicate (10^6 kg/yr)

There is also a data frame called `caseSetUp`, which contains a number of important parameters that define the way that the test was implemented. They are presented here.

Table 2. This table defines the contents of the `caseSetUp` data frame

Column	Definition
year1	the water year that is the start of the trend period (an integer)
yearData1	the water year that is the start of the data set (an integer)
year2	the water year that is the end of the trend period (an integer)
yearData2	the water year that is the end of the data set (an integer)
numSamples	number of samples in eList$Sample
nBoot	maximum number of replicates (called Mmax in paper)
bootBreak	minimum number of replicates (called Mmin in paper)
blockLength	length of blocks for bootstrap (called B in the paper)
confStop	1 - alphap, the width of the confidence interval used in adaptive stopping rule (default alphap=0.3 so confStop=0.7)

2 EGRETci Workflow

```
library(EGRET)
library(EGRETci)
eList <- Choptank_eList

#Interactive function to set up trend analysis:
caseSetUp <- trendSetUp(eList)

#Currently, only water-year calculations are supported
eList <- setPA(eList)
eList <- setForBoot(eList)

eBoot <- wBT(eList,caseSetUp,
             saveOutput = TRUE, fileName = "outputText.txt")

#Interactive save output function:
saveEGRETci(eList, eBoot)
```

The output in output.txt is:

```
Choptank River    Inorganic nitrogen (nitrate and nitrite)

 Bootstrap process, for change from Water Year 1985 to Water Year 2005
                  data set runs from WaterYear 1980 to Water Year 2011
 Bootstrap block length in days 200
 bootBreak is 9  confStop is 0.7

WRTDS estimated concentration change is   0.328  mg/L
WRTDS estimated flux change is          0.03148  10^6 kg/yr
value is bootstrap replicate result (deltack or deltafk in paper)
nPos is cumulative number of positive trends
post_p is posterior mean estimate of probability of a positive trend
Lower and Upper are estimates of the 90% CI values for magnitude of trend

     rep            Concentration             |              Flux
         value     nPos post_p   Lower   Upper  |    value   nPos  post_p    Lower    Upper
      1   0.267     1     0.75   0.267   0.267  |  0.02585     1    0.75  0.02585  0.02585
      2   0.295     2    0.833   0.267   0.295  |  0.02635     2   0.833  0.02585  0.02635
      3   0.318     3    0.875   0.267   0.318  |  0.03256     3   0.875  0.02585  0.03256
      4   0.233     4      0.9   0.233   0.318  |  0.02718     4     0.9  0.02585  0.03256
      5   0.373     5    0.917   0.233   0.373  |  0.03897     5   0.917  0.02585  0.03897
      6   0.309     6    0.929   0.233   0.373  |  0.03404     6   0.929  0.02585  0.03897
      7   0.354     7    0.938   0.233   0.373  |  0.03824     7   0.938  0.02585  0.03897
      8   0.293     8    0.944   0.233   0.373  |  0.02632     8   0.944  0.02585  0.03897
      9   0.415     9     0.95   0.233   0.415  |  0.04439     9    0.95  0.02585  0.04439
     10   0.421    10    0.955   0.233   0.421  |  0.04115    10   0.955  0.02585  0.04439
     11   0.374    11    0.958   0.233   0.421  |  0.03796    11   0.958  0.02585  0.04439
     12   0.362    12    0.962   0.233   0.421  |  0.03708    12   0.962  0.02585  0.04439
     13   0.309    13    0.964   0.233   0.421  |  0.03073    13   0.964  0.02585  0.04439
     14   0.353    14    0.967   0.233   0.421  |  0.03364    14   0.967  0.02585  0.04439
     15    0.18    15    0.969    0.18   0.421  |  0.01796    15   0.969  0.01796  0.04439
     16   0.312    16    0.971    0.18   0.421  |  0.02388    16   0.971  0.01796  0.04439
     17   0.401    17    0.972    0.18   0.421  |  0.04159    17   0.972  0.01796  0.04439
     18    0.34    18    0.974    0.18   0.421  |  0.03359    18   0.974  0.01796  0.04439
```

```
      19   0.285    19   0.975    0.18   0.421   |    0.03073    19   0.975  0.01796  0.04439
      20   0.291    20   0.976   0.183   0.421   |    0.02888    20   0.976  0.01826  0.04425

Should we reject Ho that Flow Normalized Concentration Trend = 0 ? Reject Ho
 best estimate is   0.328 mg/L
  Lower and Upper 90% CIs    0.183    0.421
 also 95% CIs    0.180    0.421
 and 50% CIs    0.292    0.370
 approximate two-sided p-value for Conc     0.048
* Note p-value should be considered to be < stated value
 Likelihood that Flow Normalized Concentration is trending up = 0.976  is trending down = 0.0238

Should we reject Ho that Flow Normalized Flux Trend = 0 ? Reject Ho
 best estimate is  0.03148 10^6 kg/year
  Lower and Upper 90% CIs   0.0183   0.0442
 also 95% CIs   0.0180   0.0444
 and 50% CIs   0.0266   0.0382
 approximate two-sided p-value for Flux     0.048
* Note p-value should be considered to be < stated value
 Likelihood that Flow Normalized Flux is trending up = 0.976  is trending down = 0.0238

 Upward trend in concentration is highly likely
 Upward trend in flux is highly likely
 Downward trend in concentration is highly unlikely
 Downward trend in flux is highly unlikely
```

3 Histograms

These functions plot histograms of all of the trend slopes from the full set of replicates created by wBT. These slopes are stored in the eBoot list. They can be plotted as an individual histogram for trends in Flow Normalized Concentration and Flow Normalized Flux, or they can be combined using the script shown below in Figure 3:

```
library(EGRET)
library(EGRETci)

# Example data included in package:
eList <- Choptank_eList # Example data from EGRET package
eBoot <- Choptank_eBoot
caseSetUp <- Choptank_caseSetUp

#Concentration:
plotHistogramTrend(eBoot, caseSetUp, eList,
                   flux=FALSE, xSeq = seq(-20,60,5))
```

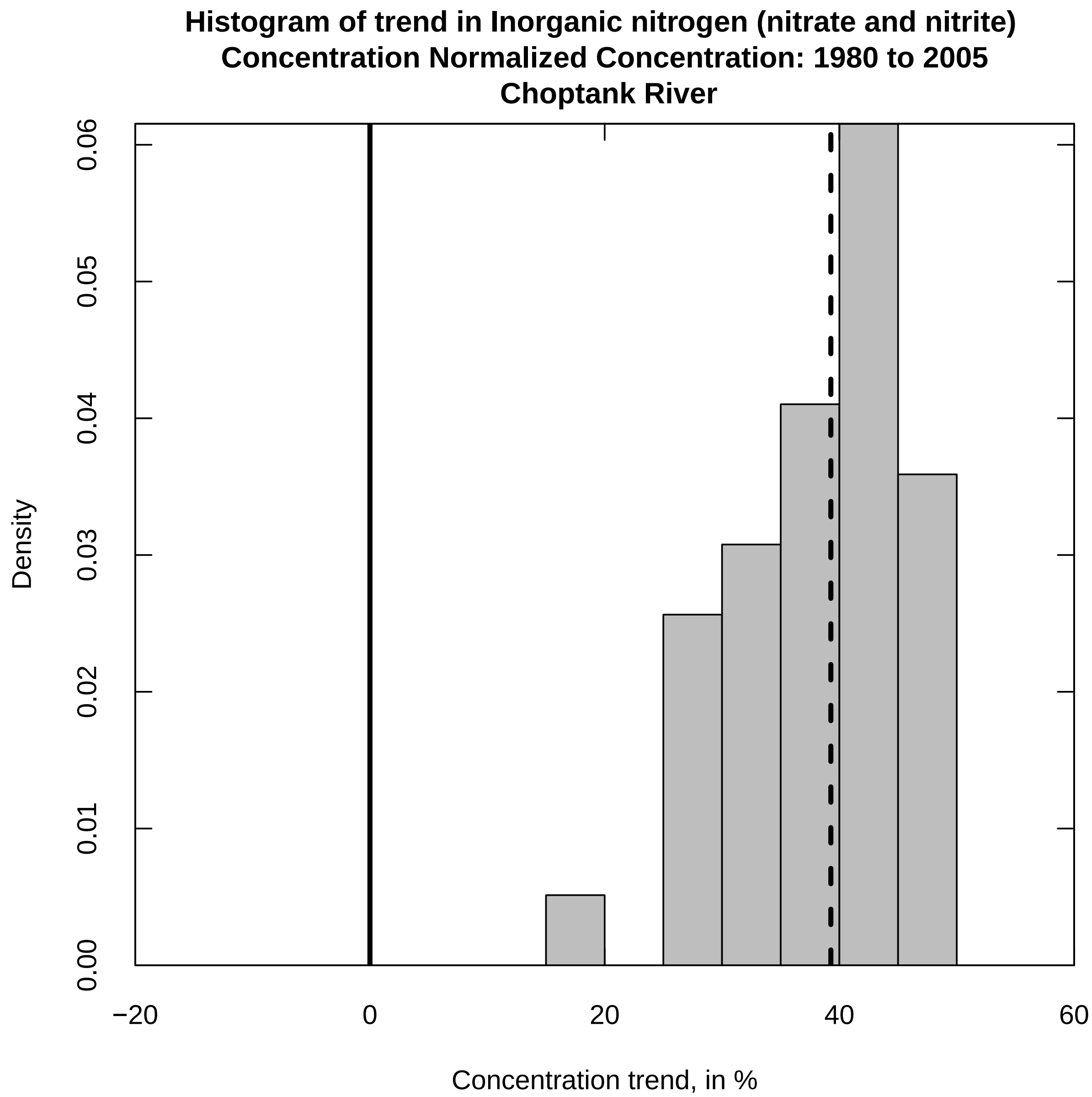

Figure 1. Concentration histogram

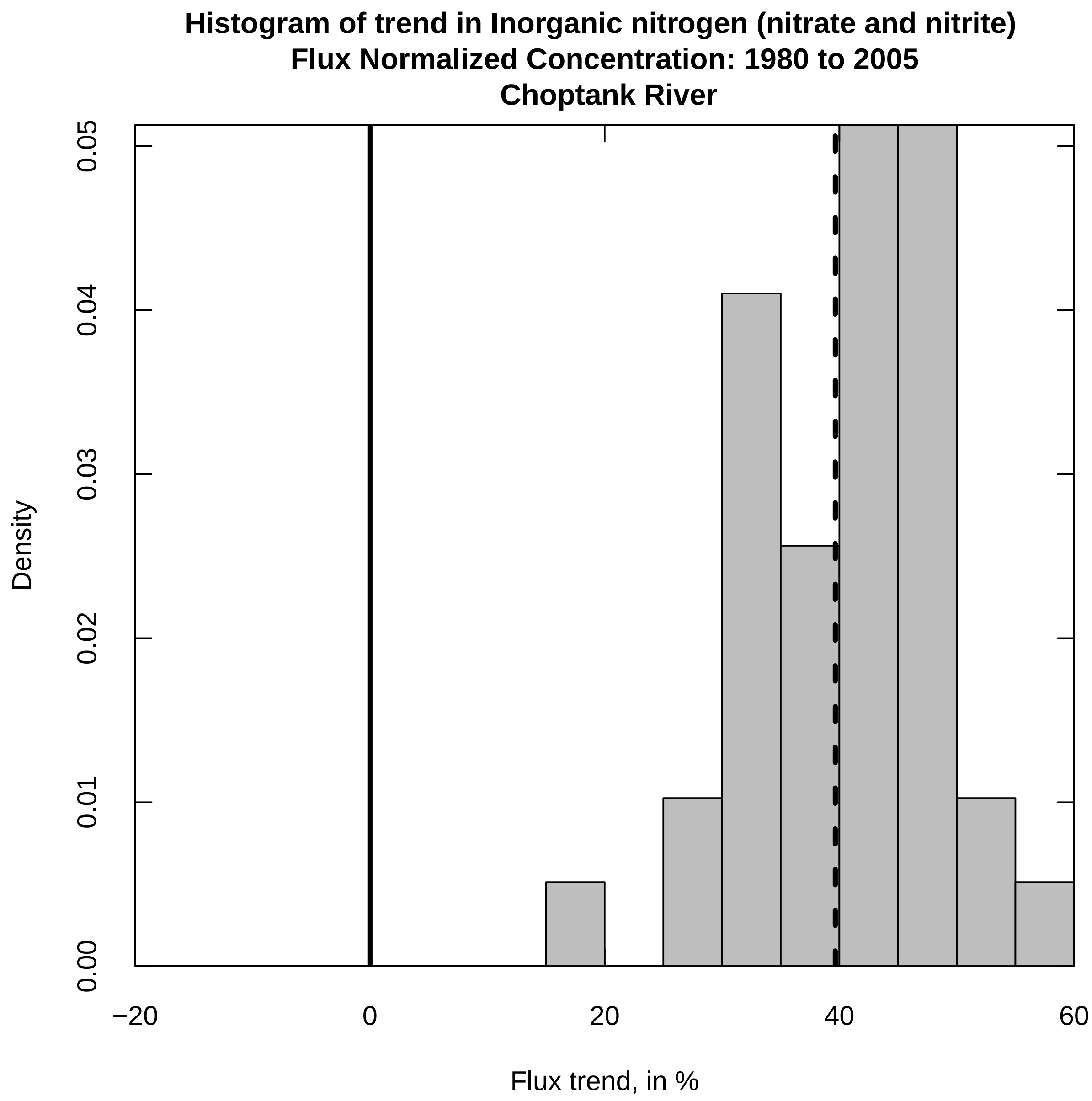

Figure 2. Flux histogram

```
#Flux
plotHistogramTrend(eBoot, caseSetUp, eList,
                   flux=TRUE, xSeq = seq(-20,60,5))
```

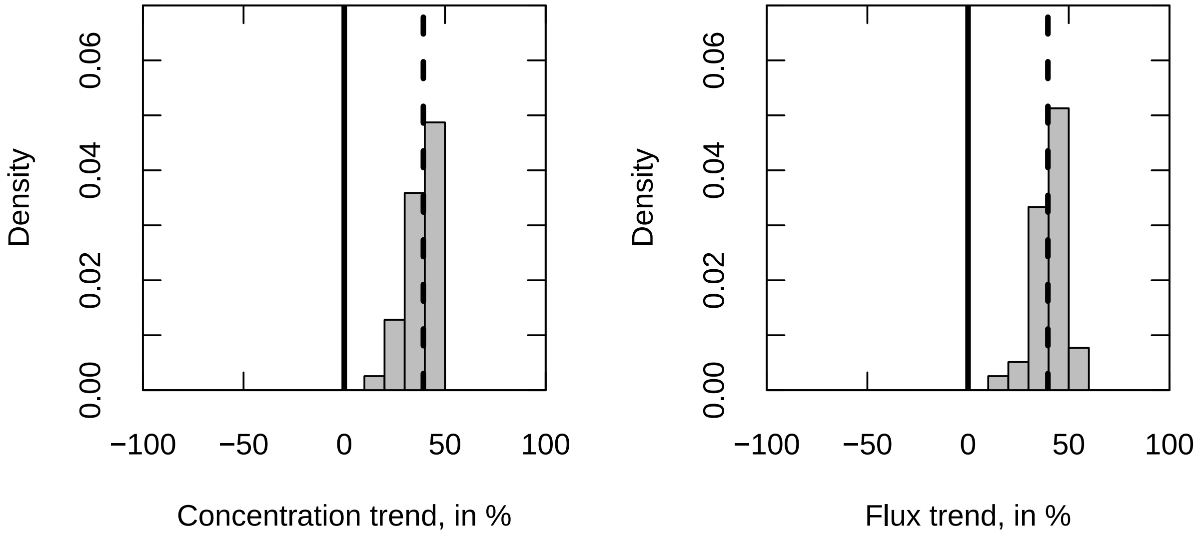

Figure 3. Combo example

Use the par function to set up both functions to plot side-by-side:

```
par(mfrow=c(1,2))
plotHistogramTrend(eBoot, caseSetUp, eList, flux=FALSE,
                   printTitle=FALSE, ylim=c(0,0.07))
plotHistogramTrend(eBoot, caseSetUp, eList, flux=TRUE,
                   printTitle=FALSE, ylim=c(0,0.07))
```

4 Confidence Bands

There are two versions of the workflow for confidence bands. The first is very simple, but results in the use of a very large amount of computer time (easily an hour). It is best done in the Terminal or some other window, rather than in the console, because it will make it impossible for other work in R to take place if it is done in the console. The second way it can be done is using parallel computing in R. A script for that purpose is provided. It does require three extra packages be installed (foreach, doParallel, and iterators). It is also best to run this in the terminal, because if it is run in the console, no other work can be done while it is running. Regardless of the way that the computations are done, they will result in the creation of a small data frame called CIAnnualResults. Once that data frame exists, the graphics can be produced using the functions plotConcHistBoot and plotFluxHistBoot. In either case they only require two arguments (eList and CIAnnualResults). However, the user can specify a number of other arguments. These other arguments are the same ones used in the plotConcHist and plotFluxHist functions in the base EGRET package.

Base-R workflow for confidence bands on model results:

```
library(EGRET)
library(EGRETci)

eList <- Choptank_eList

CIAnnualResults <- ciCalculations(eList)

save(CIAnnualResults, file="CIAnnualResults.RData")
```

Taking advantage of the foreach package to do parallel computing:

```
library(foreach)
library(doParallel)
library(iterators)
library(EGRET)
library(EGRETci)

eList <- Choptank_eList

nBoot <- 100
blockLength <- 200
coreOut <- 2 #Number of cores to leave out of processing tasks

widthCI <- 90
ciLower <- (50-(widthCI/2))/100
ciUpper <- (50+(widthCI/2))/100
probs <- c(ciLower,ciUpper)

nCores <- detectCores() - coreOut
cl <- makeCluster(nCores)
registerDoParallel(cl)
repAnnualResults <- foreach(n = 1:nBoot,.packages=c('EGRETci')) %dopar% {
   annualResults <- bootAnnual(eList, blockLength)
```

```
}
stopCluster(cl)

# save(repAnnualResults, file="repAnnualResults.RData")

CIAnnualResults <- ciBands(eList, repAnnualResults, probs)
save(CIAnnualResults, file="CIAnnualResults.RData")
```

The following concentration and flux plots can then be generated from the CIAnnualResults data.

```
#Load package data:
eList <- Choptank_eList
CIAnnualResults <- Choptank_CIAnnualResults

plotConcHistBoot(eList, CIAnnualResults)
```

```
plotFluxHistBoot(eList, CIAnnualResults)
```

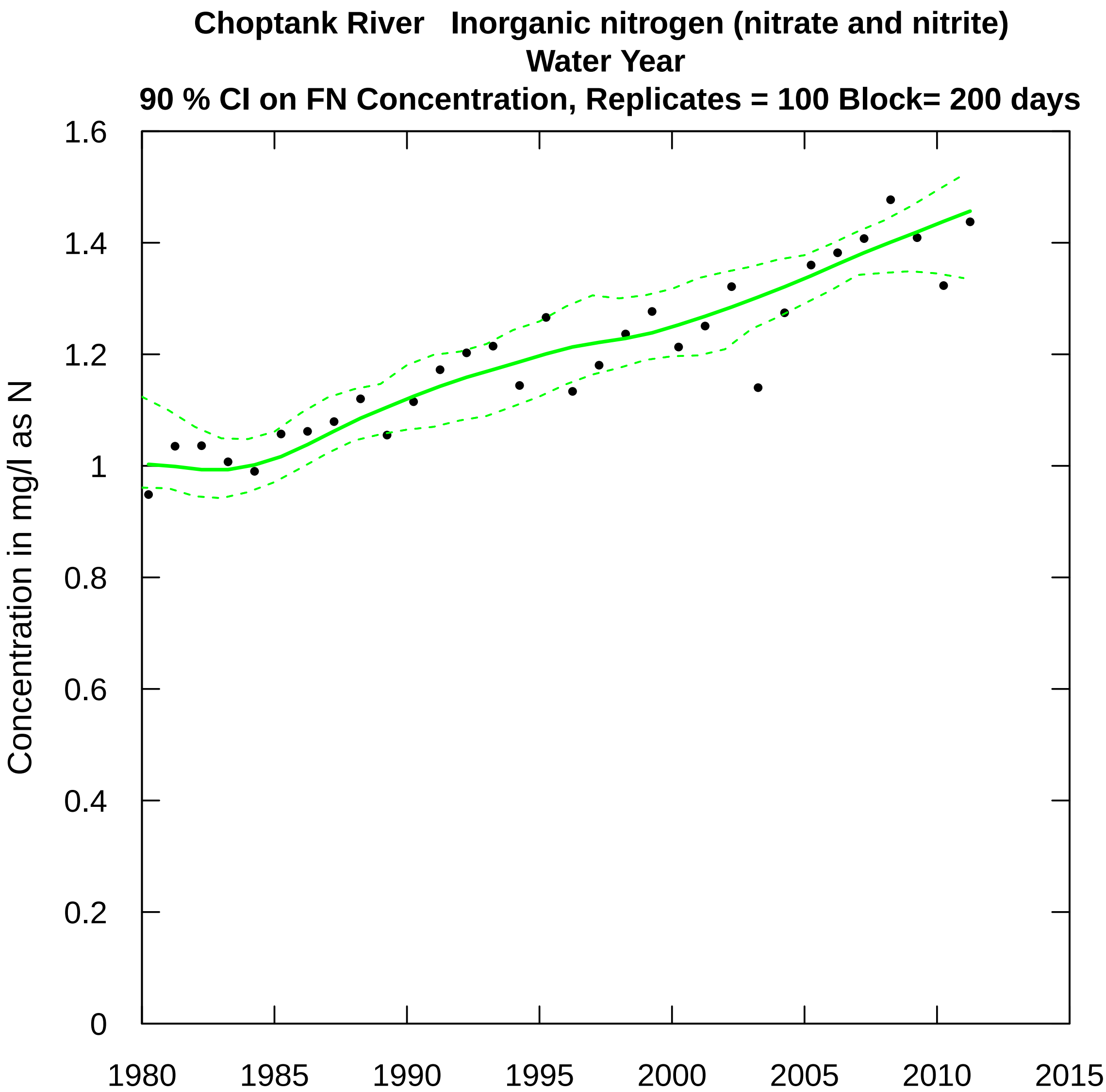

Figure 4. plotConcHistBoot

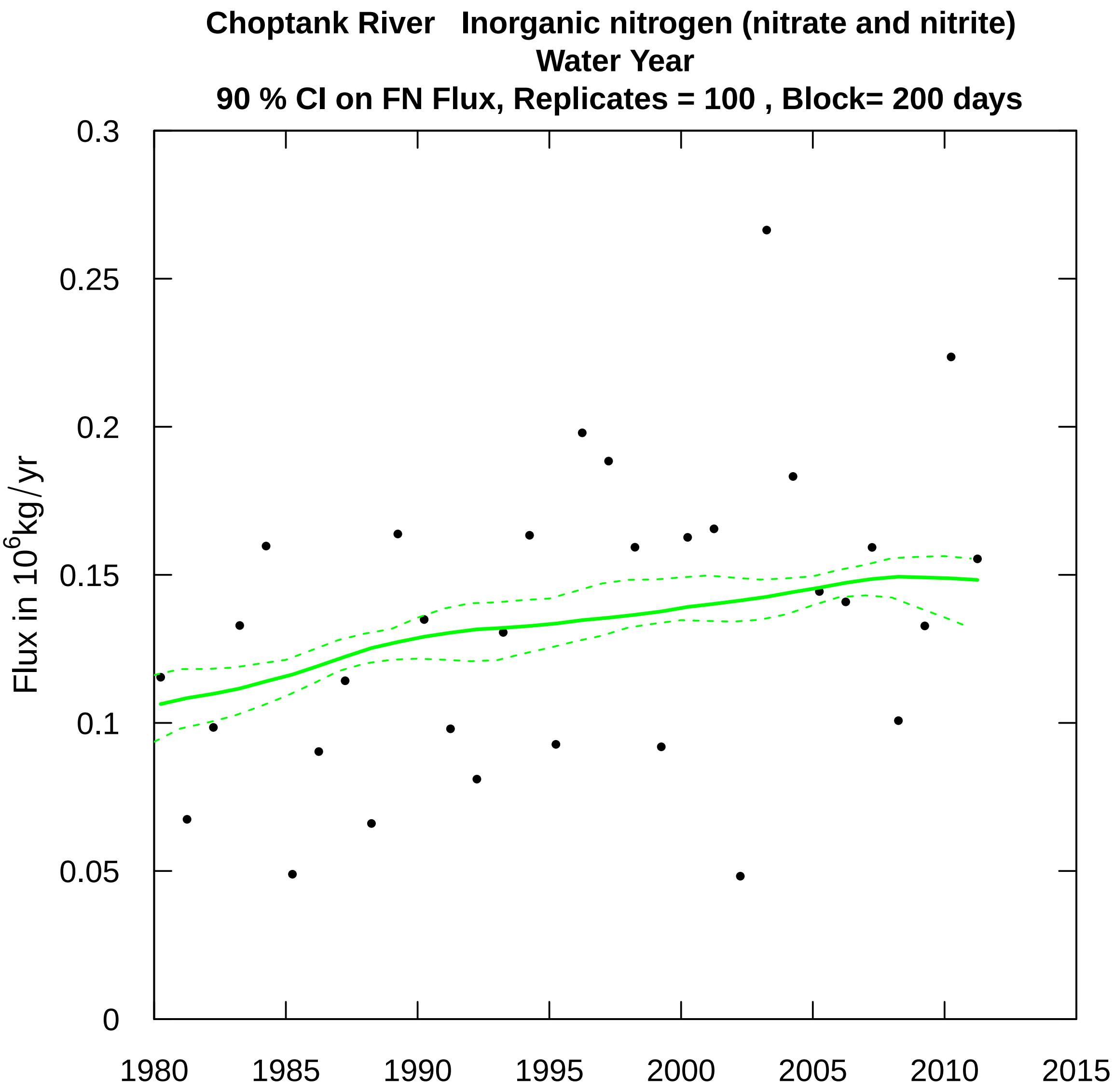

Figure 5. plotFluxHistBoot