Documentation for SEAWAVE–QEX, Version 2, March 2, 2018

The SEAWAVE-QEX functions and documentation described in the appendix of the main report (Vecchia, 2018) have been updated as described in this document. Several additional, optional arguments have been added to the swaveqexMerge and swaveqexFit functions to allow more flexibility in the data processing and diagnostic output. This documentation should be used in place of Vecchia (2018; appendix). Some additional examples are provided to illustrate the changes for version 2, particularly with respect to the graphical output.

SEAWAVE–QEX Functions, Version 2

Function: swaveqexMerge

Purpose: Merges daily discharge and pesticide concentration data, completes data screening steps, produces rough data plots, and creates object for input to swaveqexFit.

Required R libraries: waterData

Usage

```
> qexfitinput <– swaveqexMerge( cdatin, qwstnum, ddstnum, yrbeg, yrend, getdd="WD", minss=3,
                                runname=" ", redconc=0)
```

cdatin is a data frame with the pesticide concentration data

- the first column should be the station number (character)
- the second column should be the date, in "yyyy-mm-dd" format (character)
- the concentration value should be in a column named "final_value" (numeric)
- the remark should be in a column named "final_remark" (character)

qwstnum is the station number from cdatin to analyze (character).

ddstnum is the station number for daily discharge (character, usually the same as qwstnum). If getdd is omitted or gettdd= "WD", the waterData package is used to download daily discharge for the specified U.S. Geological Survey station number (ddstnum). If getdd= "File", the daily discharge data are assumed to be in a tab-delimited text file called dd_ddstnum.txt in the current working directory. The first column of the text file for daily discharge should be the date (in yyyy-mm-dd format) and the second column should be the discharge value. The text file should not have a header and there should be no missing values.

yrbeg and yrend are the beginning and ending calendar years for analysis (numeric). If unknown, the entire period of record can be analyzed by setting yrbeg=0 and yrend=3000.

minss is the minimum spacing allowed between successive samples, in days. The value for minss can be any integer between 1 and 15, and the default value is 3.

runname is an optional character identifier to append to the file names of the output produced by swaveqexFit (see the following description of output for swaveqexFit).

redconc ("reduce concentration") is a constant value, in micrograms per liter, to be subtracted from the observed concentrations before fitting the model. redconc can be positive (in which case concentrations are reduced) or negative (in which case concentrations are increased). The value of redconc should be less than the minimum of the observed concentrations.

Examples

qexfitinput <- swavqexMerge(SWqexAtrazineData, "03353637","03353637",1993,2002)

This command prepares the atrazine data for Little Buck Creek near Indianapolis, Indiana (USGS station number 03353637 in the SWqexAtrazineData dataframe) and saves the result in an object called qexfitinput.

qexfitinput <- swavqexMerge(SWqexAtrazineData, "03353637","03353637",1993, 2002, getwd="File")

This command looks for discharge data in a file called dd_03353637.txt in the current working directory.

qexfitinput <- swavqexMerge(SWqexAtrazineData, "03353637","03353637",1993, 2002, minss=1, runname="SS1R004", redconc=0.004)

This command assumes a minimum sample spacing of 1 day instead of the default (minss=3), subtracts 0.004 micrograms per liter from the observed concentrations, and specifies that the identifier "SS1R004" should be appended to the output files produced by swaveqexFit.

Output

Rough data plots (sent to the default plot device). These plots can be used to adjust yrbeg and yrend and see if data are sufficient for analysis.

An object (list), named qexfitinput or any other user-specified name, for input to swaveqexFit.

Function: swaveqexFit

Purpose: Uses input object prepared by swaveqexMerge to estimate the model parameters, produce diagnostic plots, and generate conditional simulations of daily concentration.

Required R libraries: tmvtnorm, survival

Other functions required (described later): swaveqexPESTpdo, swaveqexCSIM

Usage

> qexfitout <- swaveqexFit(qexfitinput, outfolder, ncs=100, samcov="full", itrans=1, loc=0, stnd="no")

qexfitinput is an object (list) produced by swaveqexMerge.

outfolder is a character name specifying the name of the folder to save the diagnostic plots and conditional simulations produced by swaveqexFit. The folder needs to be created ahead of time. For example, "outatrazine\\" will save the results in a folder called outatrazine in the default working directory.

ncs is the number of conditional simulations to generate (default is 100, maximum is 250).

samcov specifies whether samples are collected year-round (samcov= "full") or only during a shorter sampling season (samcov= "partial"). It is recommended to use samcov= "full" if the

sampling season is 10 months or more, samcov= "partial" if the sampling season is 6 months or less, and user discretion is advised if the sampling season is between 6 and 10 months. If samcov= "partial", only seasonal waves with a single application season are allowed.

itrans (integer) indicates the transformation to use for the pesticide concentration data. If itrans=1 (the default), a base-10 logarithmic transformation is used; if itrans=2, a square-root transformation is used; and if itrans=3, a cube-root transformation is used. Note that the SEAWAVE-QEX model methodology was developed and verified using itrans=1. Therefore, use of alternate transformations is not recommended without careful consideration of the model assumptions and careful interpretation of model output.

loc (numeric) specifies whether or not to indicate a "limit-of-concern" on the graphical output. If loc=0 (the default), no loc is shown. If loc>0, a specified limit-of-concern is indicated. For example, loc=10 specifies that the limit-of-concern is 10 micrograms per liter.

stnd indicates whether or not to show the fitted trend line on the graphical output. If stnd= "no" (the default), no trend line is shown, and if stnd= "yes", the trend line is shown.

Output

A Portable Document Format (PDF) file called "PlotsXXXYYY.pdf" is produced in the folder specified by the outfolder argument, where "XXX" is the value of the qwstnum argument and "YYY" is the value of the optional runname argument used in the call to swaveqexMerge. This file contains diagnostic plots similar to figures 5–8 of the main report (Vecchia, 2018).

For version 2, the diagnostic plots have been enhanced and several additional plots have been included to improve the model diagnostic output. Examples of the new plots are provided in the "Instructions for running SEAWAVE-QEX" section of this documentation.

A tab-delimited text file called "CSIMSXXXYYY.txt" (where "XXX" is the value of the qwstnum argument and "YYY" the value of the runname argument) is produced in the folder specified by the outfolder argument. This file contains daily pesticide concentrations and other information produced by the model. Another tab-delimited text file with summary statistics of the output, called "STATSXXXYYY.txt", also is produced. See the "Instructions for running SEAWAVE-QEX" section of this documentation for a description of these files.

Additional output produced by swave consists of list with 3 elements (assigned to an object named qexfitout or any other user specified name). The elements of the list contain the following information:

qexfitout[[1]] is the station number with pesticide concentration data (character)

qexfitout[[2]] is a vector of length 25 with the output variable names (character)

qexfitout[[3]] is a vector of length 25 with the output values (numeric)

The output names and descriptions are as follows:

yrbeg	beginning year of record
yrend	ending year of record
rlen	record length
nobs	number of observations
nucen	number of uncensored observations
prucen	proportion of uncensored observations
int	regression intercept

cswave	regression coefficient for seasonal wave
pswave	approximate *p*-value for cswave
cmtfa	regression coefficient for mid-term flow anomaly
pmtfa	approximate *p*-value for cmtfa
cstfa	regression coefficient for short-term flow anomaly
pstfa	approximate *p*-value for cstfa
ctnd	regression coefficient for trend term
ptnd	approximate *p*-value for ctnd
wmcls	wave model class (1 or 2)
wmodno	pulse input model number (1 through 6)
hlife	modeled "half-life" (1 through 4)
wshft	phase shift
sigma	estimated error standard deviation
alph	estimated value of alpha
cts	estimated correlation time scale
n2LLIK	negative 2 times the log-likelihood value
sdmtfa	standard deviation of the mid-term flow anomaly
sdstfa	standard deviation of the short-term flow anomaly

Function: **swaveqexPESTpdo**

Purpose: Selects the best-fit seasonal wave model, computes estimates of the regression coefficients, and computes maximum pseudo-likelihood estimates of the seasonal standard deviation parameters and the correlation time scale.

Required R libraries: tmvtnorm, survival

Other functions used (described later): estsigxx, evalmodlikxx, compwaveconvxx

This function is called internally from swaveqexFit. User does not need to call this function.

Additional details: With highly censored data, exact maximum likelihood estimation is intractable. An alternative method, based on the pseudo-likelihood function is used. This method has been determined to be comparable (in terms of bias and efficiency) to exact maximum likelihood while being much simpler to compute (Besag, 1977; Zeger and Brookmeyer, 1986).

Function: **estsigxx**

Purpose: Finds iterative solution for sigma to maximize the pseudo-likelihood given values for alpha and cts. User does not need to call this function.

Function: **evalmodlikxx**

Purpose: Computes value of negative 2 times the log-pseudo-likelihood. User does not need to call this function.

Function: **compwaveconvxx**

Purpose: Computes the seasonal wave given the model class, model number, model half-life, and phase shift. Function used internally.

Function: **swaveqexCSIM**

Purpose: Computes conditional simulations of daily pesticide concentration given estimated model parameters and other information passed from swaveqexFit. Input and output are processed within swaveqexFit.

Required R libraries: tmvtnorm

Other functions required (described later): impcenvals, condsim

Function: **impcenvals**

Purpose: Imputes values for censored normalized residuals. Input and output are processed within swaveqexCSIM. Function used internally.

Required R libraries: tmvtnorm

Additional details: The imputed values for each block of consecutive censored residuals are generated at random from a truncated conditional multivariate normal distribution for the censored residuals given the closest uncensored values before and after the block. This process relies on the assumption of an exponential correlation function, for which the residuals have a first-order Markov dependence structure and, thus, only the closest uncensored values are required.

Function: **condsim**

Purpose: Computes a conditional trace for the normalized residuals given the uncensored residuals and the imputed censored residuals. Input and output are processed within swaveqexCSIM and the user does not need to call this function.

Required R libraries: tmvtnorm

Additional details: The values for each block of days in between the observed/imputed values are generated at random from a conditional multivariate normal distribution given the closest values before and after the block. This process relies on the assumption of an exponential correlation function, for which the residuals have a first-order Markov dependence structure and, thus, only the closest values before and after the block are required.

Model Archive Files, Version 2

The following files are provided in the Model Archive to the main report (Vecchia, 2018):

- swaveqexFunctionsV2.R
  This text file contains the R code required to create the SEAWAVE–QEX (version 2) functions.
- SEAWAVEQEX_V2_Readme.pdf
  This is the documentation for version 2
- swaveQEX.Rdata

This is an R workspace containing the following dataframes:

- *SWqexAtrazineData, SWqexCarbarylData, SWqexChlorpyrifosData, SWqexFipronilData*

  These dataframes contain the atrazine, carbaryl, chlorpyrifos, and fipronil data used for the applications described in Vecchia (2018).

- *SWqexAtrazineSites, SWqexCarbarylSites, SWqexChlorpyrifosSites, SWqexFipronilSites*

  These dataframes contain the site lists and other information for each pesticide (see tables 2–5 of Vecchia, 2018).

- *SWqexAtrazinePest, SWqexCarbarylPest, SWqexChlorpyrifosPest, SWqexFipronilPest*

  These dataframes contain the SEAWAVE–QEX parameter estimates for each pesticide/site using the period of record (yrbeg, yrend) that is specified in the site list dataframes. See the previous description of output for the swaveqexFit function for the variable names.

Instructions for Running SEAWAVE–QEX, Version 2

A recent version of R (v.3.3.0 or later) is required, and installing Rstudio is recommended. The user libraries waterData, tmvtnorm, and survival also need to be installed.

Step 1. Open the swaveQEX.Rdata workspace.

Step 2. Create the SEAWAVE–QEX (Version 2) functions in your user environment using the source command:

```
> source("swaveqexFunctionsV2.R")
```

Step 3. Before running the model, attach the following required libraries:

```
> library("survival")
> library("tmvtnorm")
> library("waterData")
```

SEAWAVE–QEX should now be fully functional. Parameter estimation results provided in the dataframes can be verified. For example, to reproduce the estimation results for the second site in the SWqwsCarbarySites dataframe (see also Vecchia, 2018; figs. 9-10 and table 3), the data first need to be prepared using swaveqexMerge:

```
> Kisco <- swaveqexMerge(SWqexCarbarylData,   "01374987","01374987",2000,2008)
```

This command will create an object named Kisco in your workspace that is ready for swaveqexFit. Before running swaveqexFit, create a folder in the default directory called CarbarylOutput (or any other name of your choosing) for storing the output files. Also, make sure that no plotting devices are open. Then, swaveqexFit can be called:

```
> KiscoPest <- swaveqexFit(Kisco,   "CarbarylOutput\\", ncs=100)
```

This command will create the object KiscoPest in the current workspace with the SEAWAVE–QEX parameter estimates. [Note that this is for a Windows computer. For an Apple computer, "CarbarylOutput\\" should be replaced by "CarbarylOutput/"]. To make the results easier to view, a dataframe can be created with the first column consisting of the parameter names and the second column consisting of the parameter estimates:

> KiscoPest <- data.frame(KiscoPest[[2]],KiscoPest[[3]])

The parameter estimates should be identical to the values in the second row of the SWqexCarbarylPest dataframe.

There should be three output files produced by the previous commands: "Plots01374987.pdf", "CSIMS01374987.txt", and "STATS01374987.txt". These files should be in the folder called CarbarylOutput (or other designated name).

Plots01374987.pdf has diagnostic plots similar to figures V2Doc-1 through V2Doc-8, which are included at the end of this documentation. Fig. V2Doc-1 is similar to fig. 9 from Vecchia (2018). The points labeled as "uncensensored observations" should be identical to the observed concentrations shown in fig. 9 of the report because those points do not change depending on the conditional simulation. However, the conditional trace (including the points labeled "simulated censored observations") are randomly generated and will differ for each plot. The estimated annual maximum concentrations (which are the average of ncs=100 values) may differ slightly for each plot. Note that the 80-pct error bounds are included in Fig. V2Doc-1, whereas only the estimated annual maxima are shown in fig. 9 of Vecchia (2018).

The second plot should look like fig. V2Doc-2, and is similar to fig. 10 of Vecchia (2018). The curves showing the fitted seasonal wave, plus and minus two seasonal standard deviations, as well as the points labeled "uncensored observations" should be identical to fig. V2Doc-2. However, the points labeled "simulated censored observations" are from a conditional trace and will be different for each plot. Shading is used to indicate the "high concentration season", which is the season during which the seasonal wave is highest.

Figures V2Doc-3 through V2Doc-8 are diagnostic plots used for examining the model fit. Fig. V2Doc-3 shows the short-term flow anomaly (STFA) versus adjusted concentrations obtained by removing variability due to seasonality, trend, and the mid-term flow anomaly (MTFA) from the observations. Fig. V2Doc-4 is a similar plot for MTFA. Fig. V2Doc-5 shows the estimated long-term trend along with adjusted concentrations obtained by removing variability due to seasonality and both of the flow anomalies. Fig. V2Doc-6 shows the normalized residuals plotted with respect to the time-of-year. The shaded region showing the high concentration season is the same as Fig. V2Doc-2. Fig. V2Doc-7 shows the normalized residuals with respect to year. Finally, Fig. V2Doc-8 shows the fitted exponential correlation function of the normalized residuals along with the empirical correlogram, and is similar to fig. 8 of Vecchia (2018), except that confidence bounds for the empirical correlogram are shown. The confidence bounds are computed using a bootstrapping technique. Also included is a vertical line showing the average time between successive observations, which for this example is about 17 days.

The output file CSIMS01374987.txt file is a tab-delimited text file with the conditional simulations produced by the model. The file has the following format:

date	year	jday	qobs	cobs	crem	estreg	estcmu	csim1	csim2	csim3
2000-01-01	2000	1	24.0			1.11	2.14	0.20	0.10	2.38
2000-01-02	2000	2	24.0			1.11	3.48	0.22	0.41	1.05
2000-01-03	2000	3	29.0			1.28	3.11	0.58	1.08	1.90
2000-01-04	2000	4	32.0			1.37	3.23	0.91	2.88	1.85

There is a row for each day of the period analyzed and the columns consist of the following:

date (column 1): YYYY–MM–DD format

year (column 2): the calendar year

jday (column 3): integer day

qobs (column 4): observed daily discharge (cubic foot per second)

cobs (column 5): observed concentrations (micrograms per liter, blanks for missing values)

crem (column 6): concentration remark (< for censored value, blank for missing or uncensored values)

The remaining columns (7 through ncs+8) contain transformed model generated concentrations (TC):

TC = Round(1000C,2), were C is concentration, in micrograms per liter. TC is obtained by multiplying the model generated concentration by 1000 and rounding to two decimal places. To obtain concentration, in micrograms per liter, out to 5 decimal places, divide TC by 1000.

Note that missing values of TC are coded as numeric value -9

estreg (column 7): Fitted value of TC from the regression model

estmu (column 8): Mean of the ncs (for example, ncs=100) conditional traces

csim1, csim2, csim3, … (columns 9 through ncs+8): conditional traces of TC.

The output file STATS01374987.txt is a tab-delimited text file containing summary statistics computed from the conditional simulations. There is a row for each year of the simulation period (2000-2008 for this example), and columns correspond to the number of observations (nobs), average spacing (days) between successive observations (ssave), minimum spacing (days) between successive observations (ssmin), maximum spacing (days) between successive observations (ssmax), the maximum observed concentration (obsmax, in micrograms per liter), the estimated annual maximum concentration (estmax), and the 80-percent error bounds for the annual maximum concentration (p10max and p90max).

Several optional arguments to the swaveqexMerge and swaveqexFit functions can be used to customize the analysis and graphical output. For example, the following commands will produce the output shown in Figs. V2Doc-9 and V2Doc-10 in a file called "PlotsSopeChlorpyrifos_02335870.pdf" in the folder called "Myoutputfolder":

```
Sope  <- swaveqexMerge(SWqexChlorpyrisosData,  "02335870","02335870",1993,2002,
          runname= "SopeChlorpyrifos_")
SopePest <- swaveqexFit(Sope, "Myoutputfolder", ncs=200, stnd="yes",loc=0.5)
```

These figures can be compared to figs. 11 and 12 of Vecchia (2018). In Fig. V2Doc-9, the trend line (stnd= "yes") and a hypothetical "limit-of-concern" (loc=0.5 micrograms per liter) are shown, and the estimated annual maxima and error bounds are computed based on 200 conditional simulations (ncs=200) instead of the default (ncs=100). Fig. V2Doc-10 is similar to fig. 12 of the main report, except two high-concentration seasons are shaded because the best-fit seasonal wave was from wave class 2.

References Cited

Besag, Julian, 1977, Efficiency of Pseudoliklihood estimation for simple Gaussian fields: Biometrika, v. 64, no. 3 p. 616–618.

Vecchia, A.V., 2018, Model methodology for estimating pesticide concentration extremes based on sparse monitoring data: U.S. Geological Survey Scientific Investigations Report 2017-5159, 47p, https://doi.org/10.3133/sir20175159.

Zeger, S.L., and Brookmeyer, Ron, 1986, Regression analysis with censored autocorrelated data: Journal of the American Statistical Association, v. 81, no. 395, p. 722–729.

Examples of diagnostic output for SEAWAVE-QEX, Version 2 (Figs. V2Doc-1 through V2Doc-10)

Fig. V2Doc-1

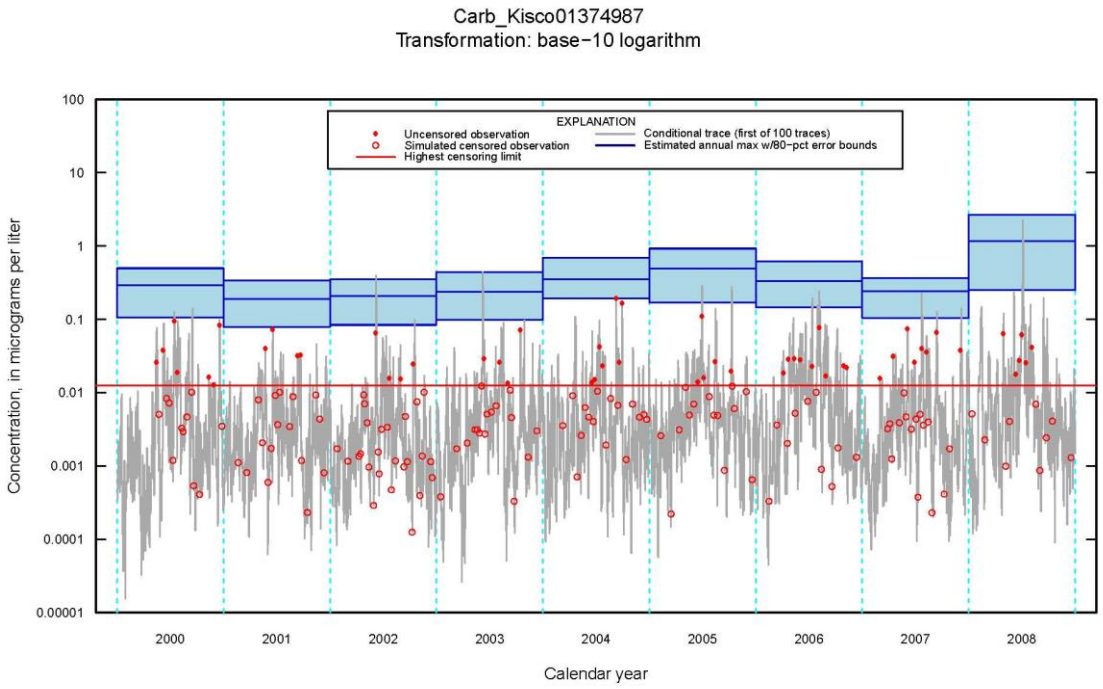

Fig. V2Doc-2

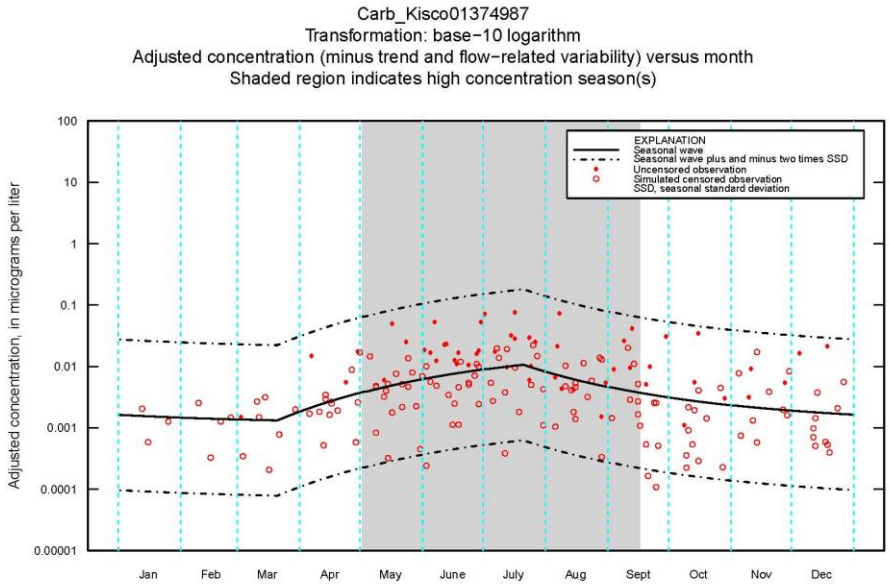

Fig. V2Doc-3

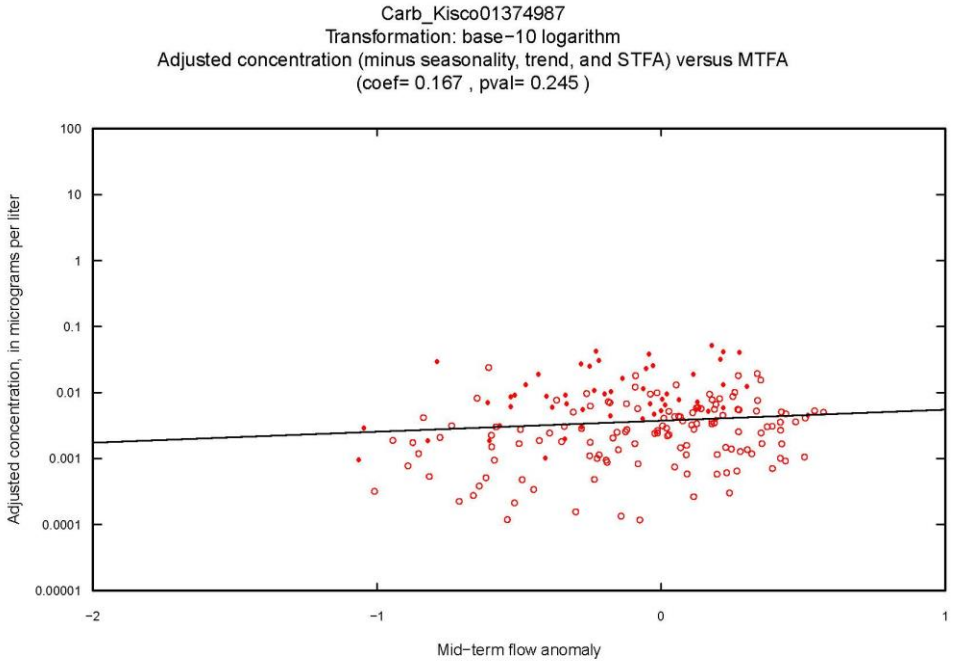

Fig. V2Doc-4

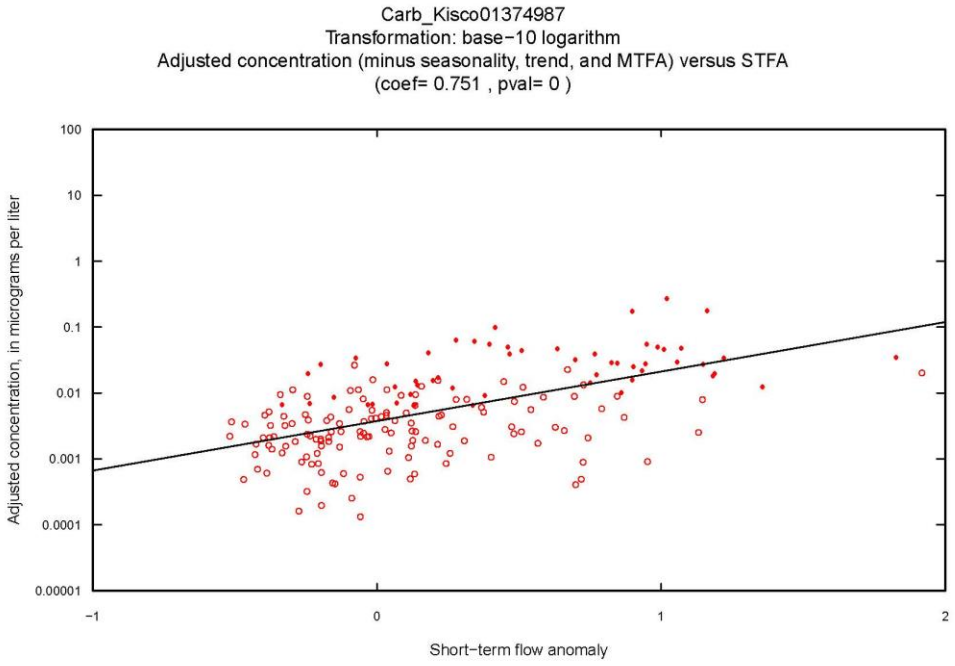

Fig. V2Doc-5

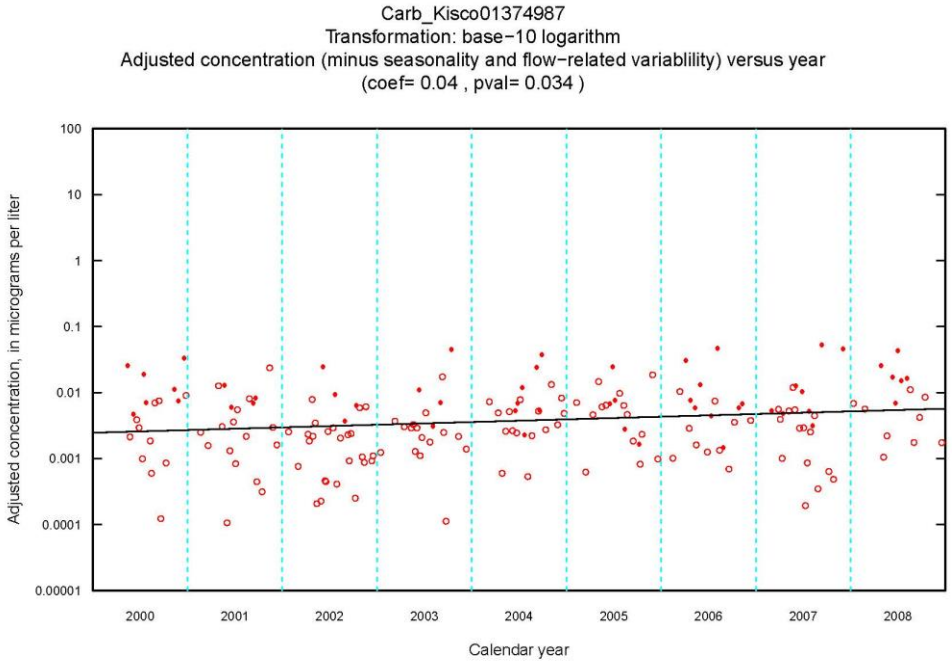

Fig. V2Doc-6

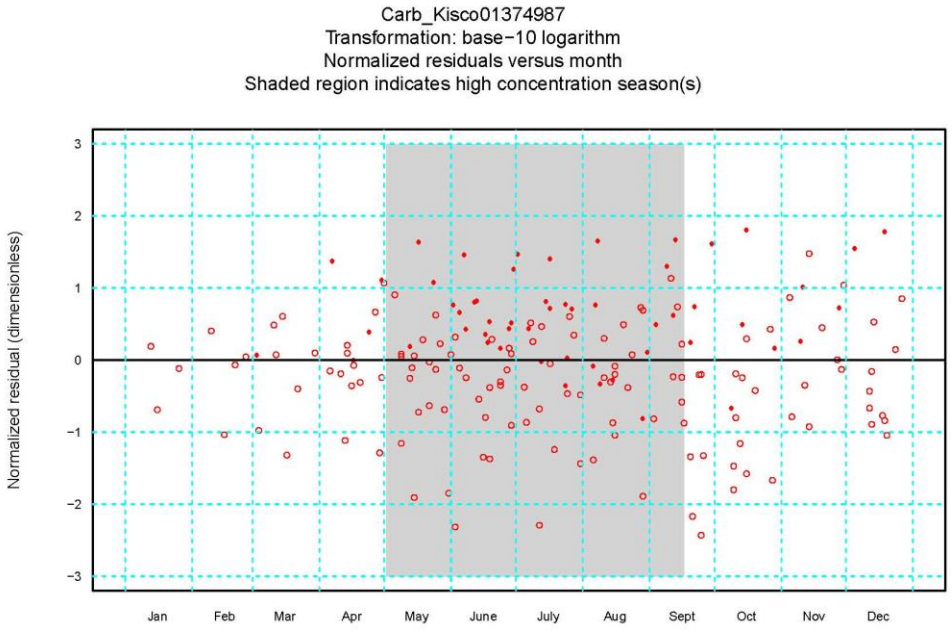

Fig. V2Doc-7

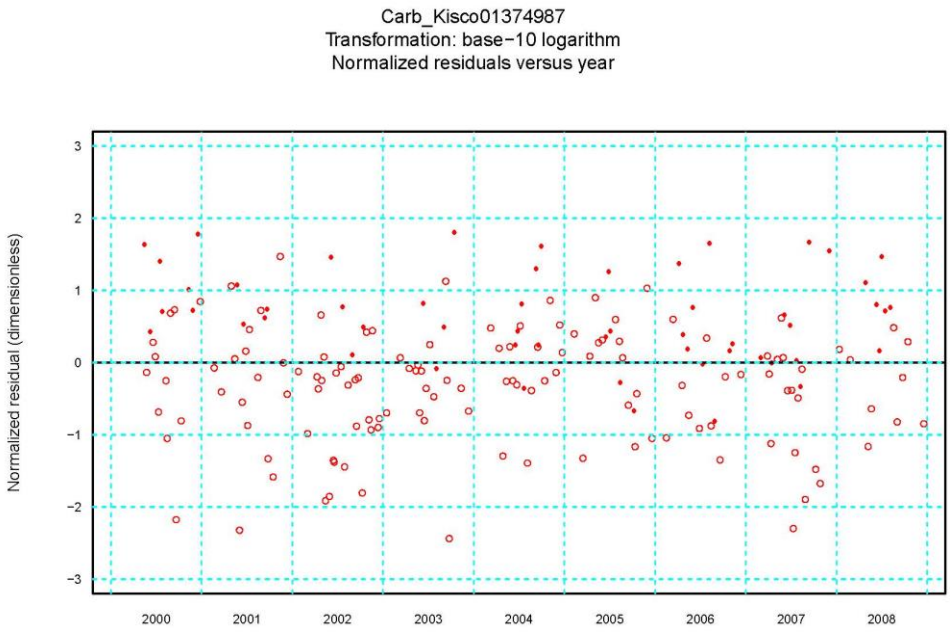

Fig. V2Doc-8

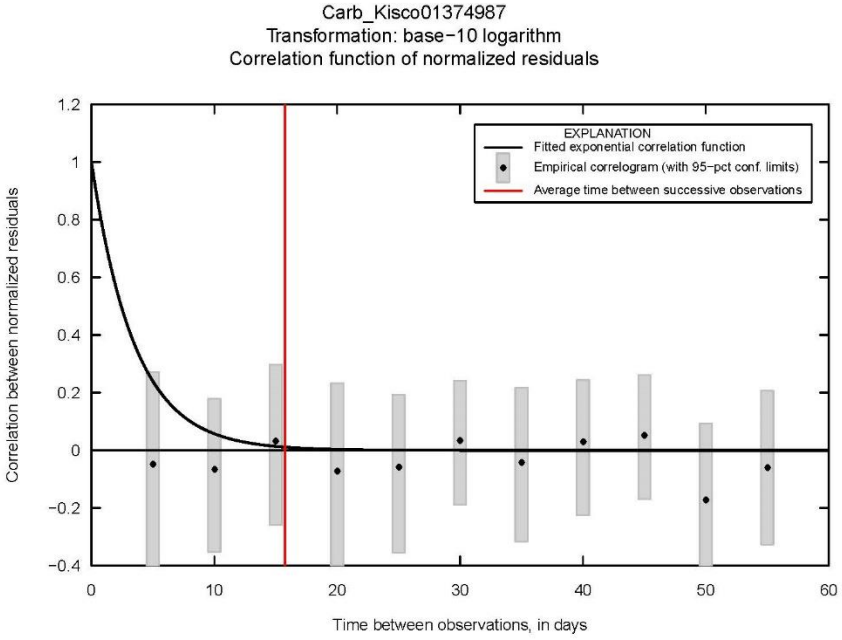

Fig. V2Doc-9

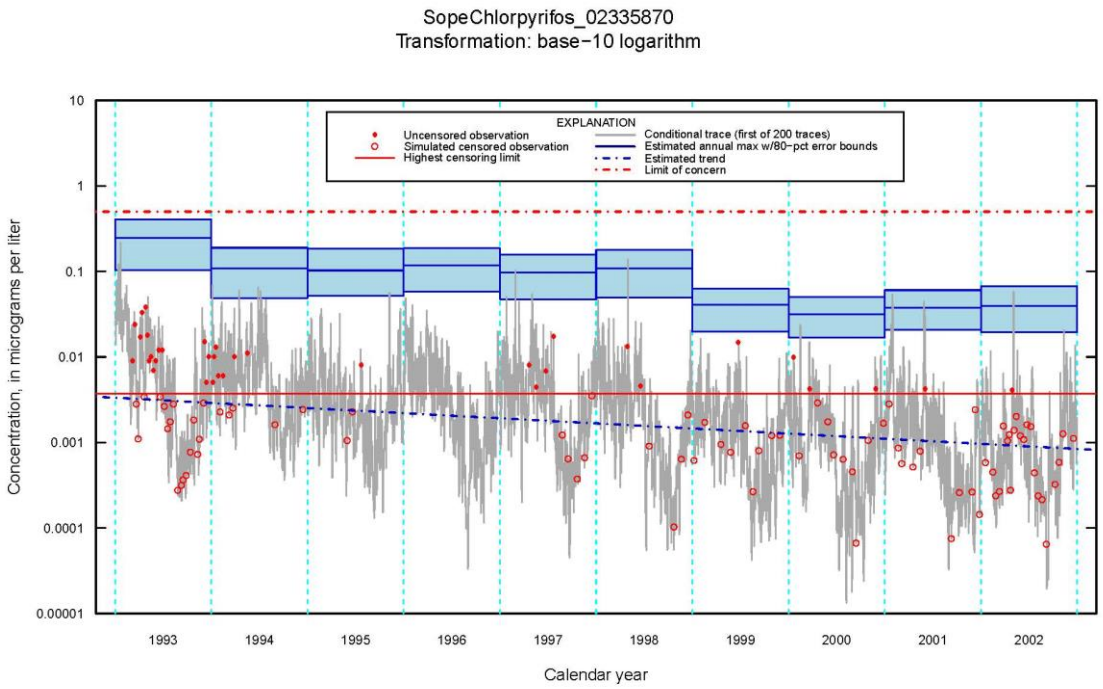

Fig. V2Doc-10