Manual tutorial for the cleaning of continuous biological parameters from the Flemish Hydrological Information Center (HIC) databases

Funded by De Vlaamse Waterweg (Flemish Waterways)

For the Hydrological Information Center (HIC) of Flanders Hydraulics Research

Written and developed by Pali Felice Gelsomini with support from Tom Maris Ecosystem Management Research Group (ECOBE) University of Antwerp

Contact palifelice.gelsomini@uantwerpen.be

November 8, 2021

Contents

Data cleaning, validation and calibration methodology

The cleaning and validation methodology was adapted from the methodology utilized by the Hydrological Information Center (HIC) for processing continuous hydrological data. The one major difference is we used median and median absolute deviations (MAD), instead of mean and standard deviation, as per recommendation of the HIC. The chlorophyll-a florescence measurements additionally must be post-calibrated using lab-tested point samples and is done following the protocol given by the sensor manufacturer YSI.

There are two steps to the data cleaning and validation process: first an automated spike removal and second a manual check. Following these two steps, the data is post calibrated if need be (e.g. the chlorophyll-a florescence must be post calibrated).

The R package "HICbioclean" was developed to automate this process. See that section for a tutorial and more details.

Continuous monitoring stations, reference site and sensor maintenance data

For the validation and calibration, the closest available point-sampled water-quality monitoring stations from OMES were selected as reference sites (see figure 1 and table 1 to see which point sampled site is reference to with which continuous site, for the location of these sites and the distances between the reference and continuous sites). The OMES monitoring campaign is funded by the Flemish Waterways. The water-quality monitoring stations used as references are sampled biweekly during the growing season April to September and monthly the rest of the year. Sensor maintenance data on cleaning and sensor changes was provided by the HIC and was used during the manual validation procedure to asses miscalibrations and sensor drifts (see table 1 for the file names and to which continuous monitored sites they belong).

Figure 1: Map of continuous monitoring sites and their respective point sampled reference sites which were used for validation and calibration of the continuous data. The point sampled site Grens was evaluated as a possible reference site for Lillo, but was not chosen. The other available point sampled sites which were not used as reference sites are displayed but are not labeled.

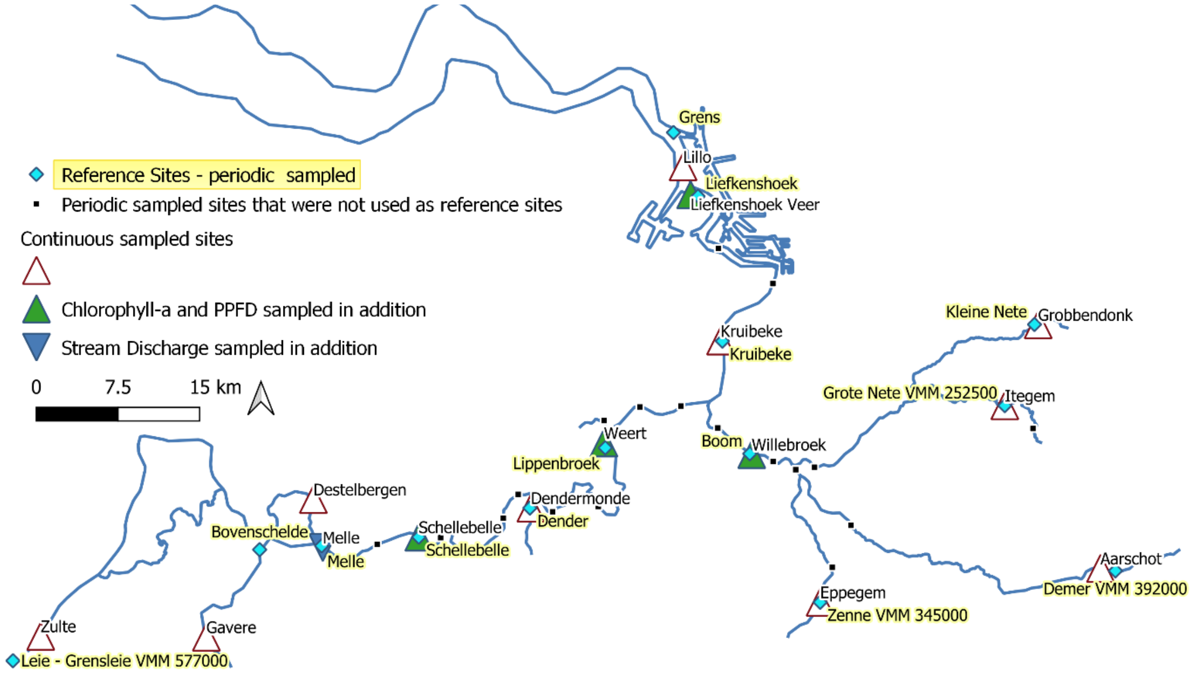


Table 1: Metadata for continuously measured biological data cleaning, calibration and validation. Continuous monitoring site names, location information, site number, the continuously measured biological parameters at each site that were cleaned and validated, the respective file for sensor maintenance information on cleanings and sensor changes, the point sampled site from the OMES campaign which was used for validating and calibrating the continuous data, the distance between the continuous measurement site and the point sampled reference site.

Continuous data site name	Site number	river	km from mouth	2018 data	2019 data	2020 data	Sensor maintenance file	Point sampled reference site from OMES	Distance from reference site (m)
Gavere SF/Bovenschelde	bos02a-SF-CM	Schelde	167			DO, pH		Bovenschelde	11440
Aarschot Afwaarts SF/Demer	dem02a-SF-CM	Demer	142	DO, pH	DO, pH	DO, pH	WISKI_MPS_Aarschot_	VMM 392000	1400
Dendermonde SF/Dender	den02a-SF-CM	Dender	125	DO, pH	DO, pH	DO, pH	WISKI_MPS_Appels_	Dender	0
Itegem Hullebrug SF/Grote Nete	gnt03a-SF-CM	Grote Nete	118	DO, pH	DO, pH	DO, pH	WISKI_MPS_Itegem Hullebrug_		
Grobbendonk Troon SF/Kleine Nete	knt03a-SF-CM	Kleine Nete	119	DO, pH	DO, pH	DO, pH	WISKI_MPS_Grobbendonk_	Kleine Nete	345
Zulte SF/Leie	lei05a-SF-CM	Leie	194			DO, pH		VMM 577000	3448
Klein Willebroek SF/Rupel	rup02e-SF-CM	Rupel	99	DO, pH, chl-a	DO, pH, chl-a, PPFD	DO, pH, chl-a, PPFD	WISKI_MPS_Boom_	Boom	193
Eppegem SF/Zenne	zen03a-SF-CM	Zenne	116	DO, pH	DO, pH	DO, pH	WISKI_MPS_Eppegem_	VMM 345000	0
Lillo Meetpaal-Onder SF/Zeeschelde	zes07g-SF-CMO	Zeeschelde	60	DO	DO	DO	WISKI_MPS_Gavere_	Liefkenshoek	3032
Liefkenshoek Veer SF/zeeschelde	zes09x-SF-CM	Zeeschelde	63		DO, pH, chl-a, PPFD	DO, pH, chl-a, PPFD	WISKI_MPS_Liefkenshoek_	Liefkenshoek	698
Kruibeke SF/Zeeschelde	zes24a-SF-CM	Zeeschelde	85	DO, pH	DO, pH	DO, pH	WISKI_MPS_Kruibekeveer_	Kruibeke	163
Weert SF/Zeeschelde	zes39c-SF-CM	Zeeschelde	103	DO, pH, chl-a	DO, pH, chl-a, PPFD	DO, pH, chl-a, PPFD	WISKI_MPS_Weert_	Lippenbroek	462
Lippenbroek UIT SF/Zeeschelde	zes40a-SF-CM	Zeeschelde	104	DO, pH	DO, pH	DO	WISKI_MPS_LippenbroekUIT_		
Schellebelle SF/Zeeschelde	zes54m-SF-CM	Zeeschelde	140	DO, pH, chl-a	DO, pH, chl-a, PPFD	DO, pH, chl-a, PPFD	WISKI_MPS_Schellebelle_	Schellebelle	0
Melle SF/Zeeschelde	zes57a-SF-CM	Zeeschelde	150	DO, pH	DO, pH	DO, pH	WISKI_MPS_Melle_	Melle	185
Destelbergen SF/Zeeschelde	zes57n-SF-CM	Zeeschelde	153	DO, pH	DO, pH		WISKI_MPS_Zulte_		

Automated spike removal

1. **Min/max filter:** Data was first passed through a min/max filter to remove unreasonably large and small values. 0 was the minimum value for all parameters. Dissolved oxygen and pH had a maximum of 15, chlorophyll a 1000 and PPFD 2000.

2. **Spike removal:** All sample points that were more than 3 MAD (the scale factor 1.4826 was used assuming normal distribution; this is the default in R) from the median of the 10 surrounding data points (5 before and 5 after) are automatically deleted. Median was used because it is a robust statistic, being more resilient to outliers. There needed to be a minimum of 5 surrounding data points, otherwise the point was not evaluated. Given the 5-minute sampling interval of the continuous data, 5 data points before and after was interpreted as 25 minutes before and 25 minutes after; this means that the algorithm would look no farther than 25 minutes, even if data was missing. All evaluated data points that passed the test were flagged as “automatic good”.

3. **Gap linear interpolation**: All data less than or equal to 1 hour were interpolated linearly. All interpolated data points were flagged as “automatic good”

The PPFD data has a different auto validation workflow. PPFD data was from paired sensors at a fixed distance from each other for calculating the light attenuation coefficient kd. This means that the two sensors needed to be auto-validated in tandem with each other.

1. **Min/max filter on the PPFD data** for the upper and lower sensors.
2. **Spike removal on the PPFD data** for the upper and lower sensors.
3. Generally both the upper and lower sensors show the same trends with spikes occurring at the same times. If a spike was deleted and then interpolated from one sensor but not the other sensor, then that will lead to an artificially large or small difference between sensors, thus creating an artificially large or kd value. Therefor **if a point was deleted from one sensor, it must be deleted in the other sensor as well.**
4. Many spikes were registered both in the upper and lower sensors. These spikes were most likely not sensor errors, but may have been passing clouds or debris and this information is very important for understanding the total light climate**. All spikes that are registered in both the upper and lower sensors will not be deleted.**
5. **PPFD data is linear interpolated for data gaps of max one hour.**
6. **Light attenuation coefficient kd is calculated.** kd = 1/Δz*ln(E1/E2) where Δz is the distance between sensors in meters (0.4m) and E1 is the upper sensor PPFD and E2 is the lower sensor PPFD.
7. **Delete all PPFD values where kd is negative.** Light cannot be greater when deeper in the water column.
8. **Remove all kd values where PPFD is below the detection limits** (1 µmol/s/m^2 for the upper sensor and 0.25 µmol/s/m^2 for the lower sensor). When the light levels approach zero, it becomes too difficult to accurately measure the difference between the upper and lower sensors.
9. **Spike removal on kd. Remove those spikes also from the PPFD data.**
10. **Interpolate PPFD data again for data gaps of max one hour.**
11. **Recalculate kd and remove kd values that are outside the detection limit again.**

Manual validation and calibration

1. **The time series is plotted** along with the reference site values and the sensor maintenance data on cleaning and sensor changes.

2. **The data is then visually evaluated for the following issues:**
   Sudden jumps relating to sensor replacements: There can be sudden shifts in the data following a sensor change due to miscalibrations or sensor drift or jumps relating to sensor start up (particularly an issue with pH where the value right after the sensor is placed out in the field is extremely low and then slowly rises back up over the course of the next day). Sensor miscalibration can be recalibrated linearly either two sided or one sided. The data may be sifted (+ -) and/or scaled (* /), with an attempt to match both the values and the signal amplitude to the previous and following data sections. The reference site values were used for the recalibration. When no reference site is available, then trends seen in the other measured parameters and the surrounding sites can be used as a guide. Recalibrated data is quality flagged as "estimate". If the recalibration is very untrustworthy, as in the data section still doesn't match up with the previous and following data or the amplitude of the data section does not match the surrounding data, it is flagged as "suspect". Issues related to sensor start up (e.g. pH) are simply deleted and quality flagged as "missing".
   Data noise and sensor error: Some sections of data are simply signal noise and are clearly sensor error, they are deleted and quality flagged as "missing". Some sections are sensor error (e.g. flatlines), they are deleted and flagged as "missing". Some sections seem to have an error in the sensor, but useful information may still be derived from the data, they are quality flagged as "suspect".
   Spikes in very irregular timeseries and prolonged spikes: The automated spike removal is not effective for timeseries with very high variance. Also, the automated spike removal will not remove spikes that are cause by prolonged disturbances such as debris trapped on the sensors. A visual evaluation must be made of the remaining data spikes. If the spikes are cyclical and follow the general data trends then they are left as is. If the spikes severally deviate from the general data trend or are associated with a sensor failure then they are deleted and flagged as "missing". If it is unclear, then they are flagged as "suspect".
3. **Gaps created by deleting data during the manual cleaning may be linear-interpolated.** Gaps of up to one hour large may be filled automatically and are quality flagged as "estimate". If the data gap is in a very simple signal shape and a linear-interpolation of greater than hone hour will not alter the signal shape, then a linear interpolation of greater than one hour may be done, and the points are quality flagged as "estimate"

4. **The chlorophyll-a data is then post-calibrated** after the manual cleaning and validation. Only non-suspect data is used for calculating the calibration. The calibration is calculated using linear regression between the lab-tested reference-site-data and the nearest, continuous-measured florescence-data. A y-intercept of 0 is used for the linear-regression as recommended by YSI, the sensor manufacturer. After the calibration has been applied to the data, no special quality flags are given to the data.

Data quality flagging rule summary:

- Data that is clearly sensor error is deleted and flagged as “missing” and then gaps of up to one hour are interpolated and flagged as “estimate”. If the resulting data gap is in a very simple signal shape and a linear interpolation of greater than hone hour will not alter the signal shape, then a linear interpolation of greater than one hour may be done, and the points are flagged as “estimate”
- Data that looks to be sensor error but may still provide information is flagged as “suspect”
- Data that where the sensor is mis-calibrated and the data must be recalibrated is flagged as “estimate”
- Data that was recalibrated, but the recalibration is very untrustworthy, as in the data section still doesn’t match up with the previous and following data or the amplitude of the data section does not match the surrounding data, is flagged as “suspect”
- When data is post-calibrates such as (such as the chlorophyll-a) then no data flag is applied. Post-calibration is defined as calibration needed for the sensor sampling protocol and it is not used for fixing an incorrectly calibrated sensor.
- At the end of the manual validation and calibration, all data points that are not already flagged as either “estimate” or “suspect” will be flagged as “good”

Final file formatting

As per request of the HIC, the data will be given the state of value codes 11 good measurements, 21 good calculation, 31 estimate measurements, 41 estimate calculations, 61 suspect measurements, 71 suspect calculations. Generally only the values 11, 31 and 61 will be used.

All data that was originally missing will receive their original NA value of -777 and the state of value flag of 255. All data that was deleted during this data cleaning process will receive the NA value of -88888 and the state of value flag 61. The file format will be a .zrx file with the below syntax. The highlight text is the sample location and parameter code.

```
#REXCHANGE1013plu15a-1066VAL|*|RINVAL-777.0|*|
#TZUTC+1|*| CUNITmm|*|
#LAYOUT(timestamp,value,primary_status)|*|
201708231515         0.8  <kwaliteitscode>
201708231520         0.7  <kwaliteitscode>
201708231525         -777.0     <kwaliteitscode>
201708231530         1.6  <kwaliteitscode>
201708231535         0.2  <kwaliteitscode>
201708231540         0.4  <kwaliteitscode>
201708231545         0.6  <kwaliteitscode>
```

R package HICbioclean

An R package specifically designed for automating the process of cleaning, calibrating and validating continuous biological water quality data from the Flemish HIC (Hydrological Information Centre) database.

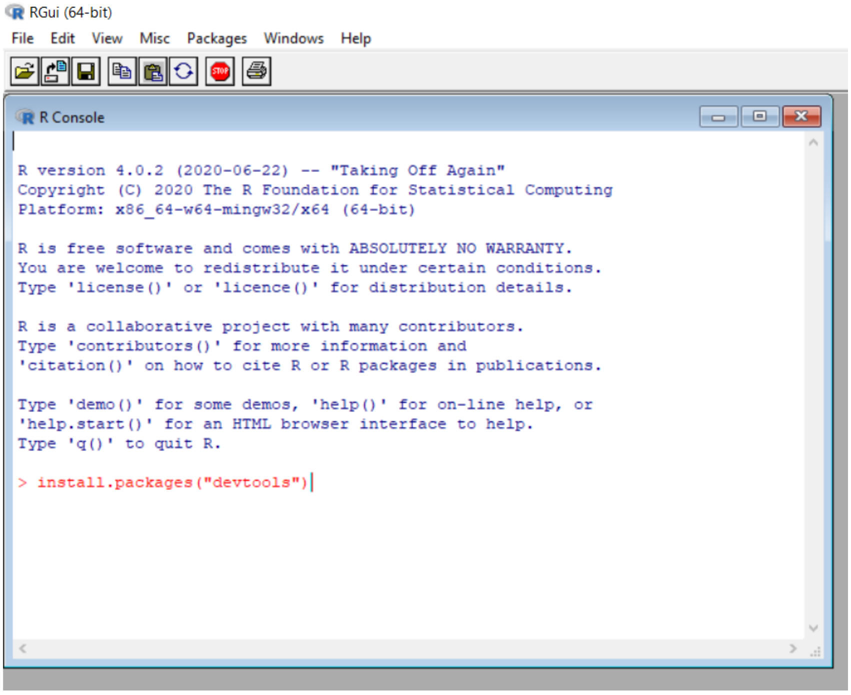
It provides both R functions for integration into R scripts and easy to use R Shiny graphical apps for intuitive data cleaning, validation and calibration without the need for coding.

Installation

Download the latest version of R from https://cran.r-project.org/ if not already installed.

Open the program R.

To download the HICbioclean package from github, you will first have to install the package devtools if you don't already have it. Copy the following code into the R console and press enter and follow the onscreen instructions:
install.packages("devtools")



```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> install.packages("devtools")
```

Install the HICbioclean package into R. Copy the following code into the R console and press enter:
devtools::install_github("pgelsomini/HICbioclean", build_vignettes = TRUE)

Open the package library for HICbioclean. Copy the following code into the R console and press enter:
library(HICbioclean)

Tutorial

The work flow for cleaning the continuous biological water quality data is split into three main steps:

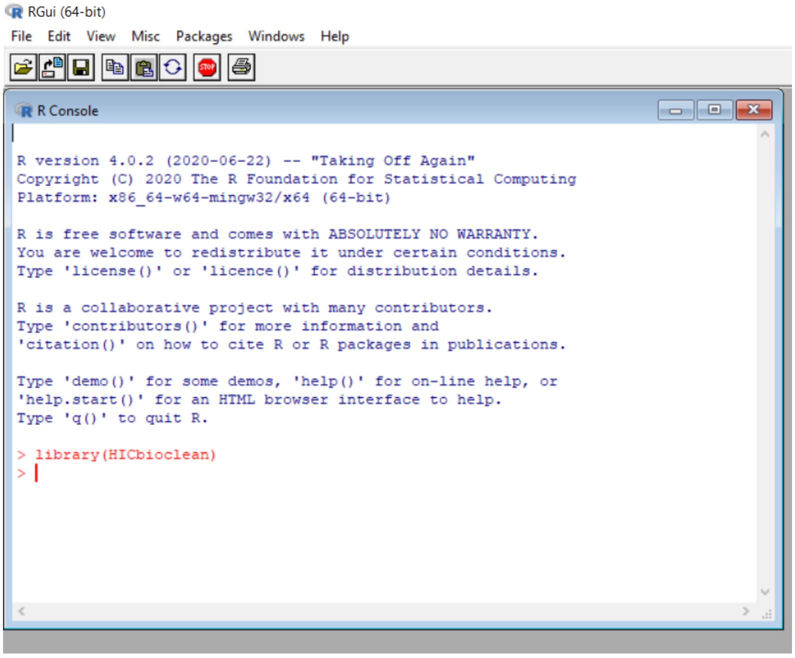
1. Data download and formatting
2. Auto-validation
3. Manual-validation and final export

Data download and formatting

Data download:

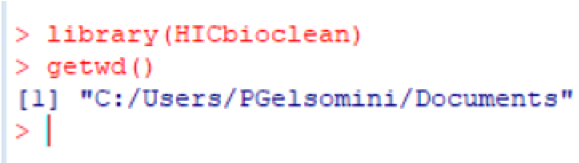
The continuous biological data (oxygen, pH, chlorophyll, and PPFD) can be downloaded from the HIC database using the function **HICwebservicesBioDownload**. This is a wrapper function for functions found within the **HICwebservices package**. The HICwebservices package allows data on the HIC database to be downloaded via an internet connection into the R environment on your computer. The HICwebservices package needs to be installed and configured on your computer before you can download the data. Contact the HIC for details on how to do this. Once you have the HICwebservices package installed and configured on your computer you can use the following code in R to download the data for a given year.

1. Open R and enter **library(HICbioclean)** into the R Console and press enter. This loads this package into R.



```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(HICbioclean)
> 
```

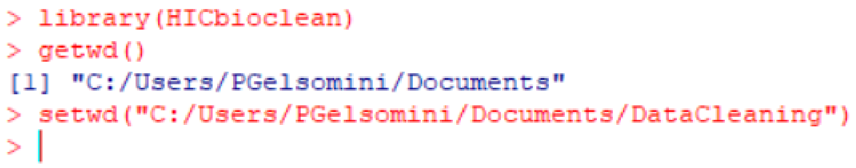
2. Check where your working directory is. This is the folder where all your data will be saved into. Type the function **getwd()** and press enter.



```
> library(HICbioclean)
> getwd()
[1] "C:/Users/PGelsomini/Documents"
> 
```

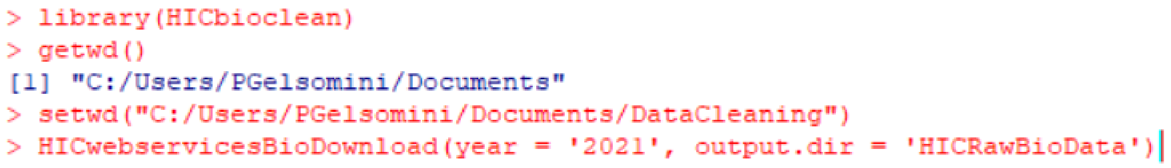
3. If you want to change the working directory, you can use the function **setwd('PathName').** Always type paths with forward-slashes (/) and not back-slashes (\). Windows uses back-

slashes so be careful when copying and pasting a path.



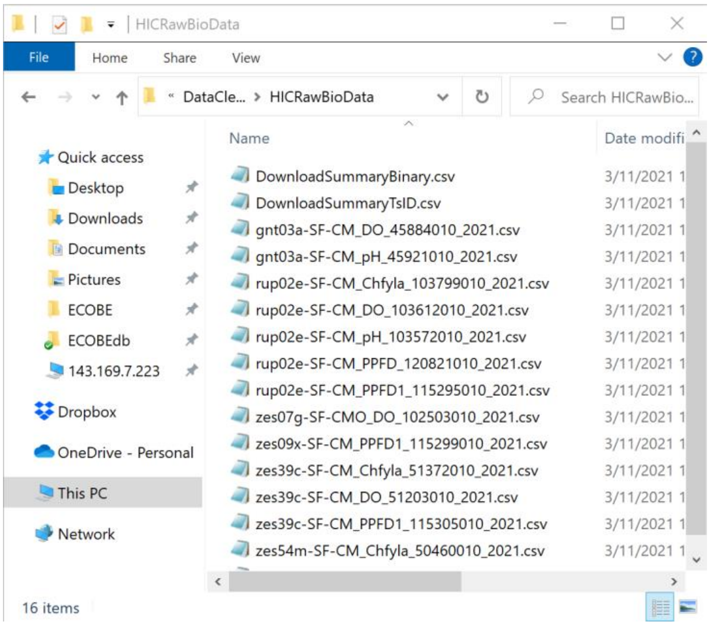
```
> library(HICbioclean)
> getwd()
[1] "C:/Users/PGelsomini/Documents"
> setwd("C:/Users/PGelsomini/Documents/DataCleaning")
> 
```

4. Use the function **HICwebservicesBioDownload(year = '2021', output.dir = 'FolderName')** to download the biological data from the HIC database. Enter the year you wish to download and the output directory folder name where you wish to save the data into. This folder will be created inside your working directory. Then press enter. The download can take some time so be patient.



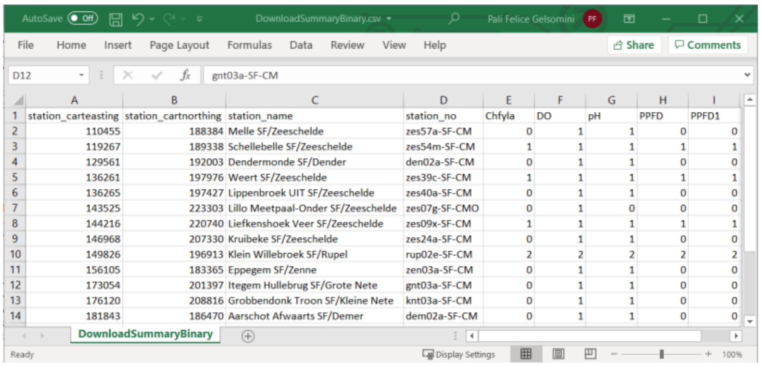
```
> library(HICbioclean)
> getwd()
[1] "C:/Users/PGelsomini/Documents"
> setwd("C:/Users/PGelsomini/Documents/DataCleaning")
> HICwebservicesBioDownload(year = '2021', output.dir = 'HICRawBioData')
```

5. The data will be saved as CSV files inside the specified output directory folder. The first part of the CSV file name is the site number, the second part is the time series ID number (from the HIC database) and the third part is the year.



Name	Date modifi
DownloadSummaryBinary.csv	3/11/2021 1
DownloadSummaryTsID.csv	3/11/2021 1
gnt03a-SF-CM_DO_45884010_2021.csv	3/11/2021 1
gnt03a-SF-CM_pH_45921010_2021.csv	3/11/2021 1
rup02e-SF-CM_Chfyla_103799010_2021.csv	3/11/2021 1
rup02e-SF-CM_DO_103612010_2021.csv	3/11/2021 1
rup02e-SF-CM_pH_103572010_2021.csv	3/11/2021 1
rup02e-SF-CM_PPFD_120821010_2021.csv	3/11/2021 1
rup02e-SF-CM_PPFD1_115295010_2021.csv	3/11/2021 1
zes07g-SF-CMO_DO_102503010_2021.csv	3/11/2021 1
zes09x-SF-CM_PPFD1_115299010_2021.csv	3/11/2021 1
zes39c-SF-CM_Chfyla_51372010_2021.csv	3/11/2021 1
zes39c-SF-CM_DO_51203010_2021.csv	3/11/2021 1
zes39c-SF-CM_PPFD1_115305010_2021.csv	3/11/2021 1
zes54m-SF-CM_Chfyla_50460010_2021.csv	3/11/2021 1

6. Two CSV files were also made called DownloadSummaryBinary.csv and DownloadSummaryTsID.csv which provide a summary of the data that was downloaded and the coordinates of the locations (coordinate system: Belgian Lambert 72). These should be taken out of this folder before you move to the auto-despiking step.



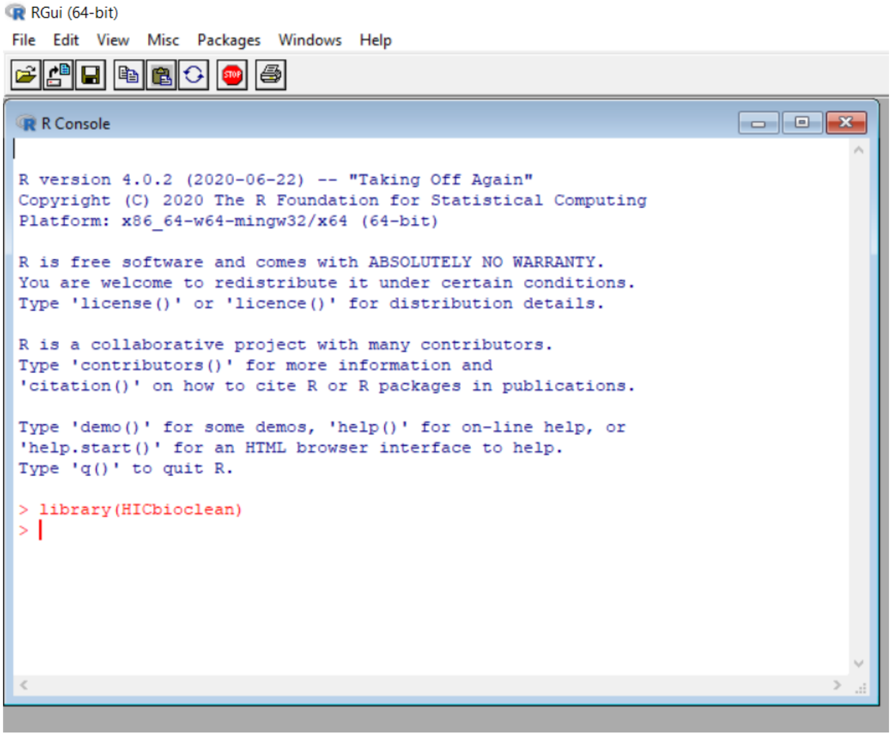
	A	B	C	D	E	F	G	H	I
1	station_carteasting	station_cartnorthing	station_name	station_no	Chfyla	DO	pH	PPFD	PPFD1
2	110455	188384	Melle SF/Zeeschelde	zes57a-SF-CM	0	1	1	0	0
3	119267	189338	Schellebelle SF/Zeeschelde	zes54m-SF-CM	1	1	1	1	1
4	129561	192003	Dendermonde SF/Dender	den02a-SF-CM	0	1	1	0	0
5	136261	197976	Weert SF/Zeeschelde	zes39c-SF-CM	1	1	1	1	1
6	136265	197427	Lippenbroek UIT SF/Zeeschelde	zes40a-SF-CM	0	1	1	0	0
7	143525	223303	Lillo Meetpaal-Onder SF/Zeeschelde	zes07g-SF-CMO	0	1	0	0	0
8	144216	220740	Liefkenshoek Veer SF/Zeeschelde	zes09x-SF-CM	1	1	1	1	1
9	146968	207330	Kruibeke SF/Zeeschelde	zes24a-SF-CM	0	1	1	0	0
10	149826	196913	Klein Willebroek SF/Rupel	rup02e-SF-CM	2	2	2	2	2
11	156105	183365	Eppegem SF/Zenne	zen03a-SF-CM	0	1	1	0	0
12	173054	201397	Itegem Hullebrug SF/Grote Nete	gnt03a-SF-CM	0	1	1	0	0
13	176120	208816	Grobbendonk Troon SF/Kleine Nete	knt03a-SF-CM	0	1	1	0	0
14	181843	186470	Aarschot Afwaarts SF/Demer	dem02a-SF-CM	0	1	1	0	0

Formatting site maintenance data:

You need to ask the HIC directly for the records of the sensor changes and the cleanings. They will give the files to you as excel files. These need to be formatted before they can be used in R.

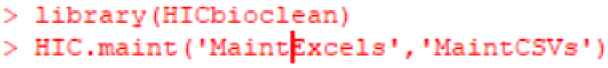
The data must first be formatted because the raw text files exported from the HIC database cannot be loaded into R and used as is. In this step the date and time is also formatted into a datetime column and a numeric datetime column for easier processing in R. The metadata in the header of the text files are placed in columns next to the data. The value column name is renamed “Value”.

7. Open R and enter **library(HICbioclean)** into the R Console and press enter. This loads this package into R.



```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(HICbioclean)
> 
```

8. Place all the excel files into one folder. Use the function **HIC.maint()** to format these excel files. Enter into the function first the path to the folder where the excels are saved (if this folder is inside your working directory then the folder name is needed) and the name of the folder where you would like to have the formatted files saved to. Then press enter.



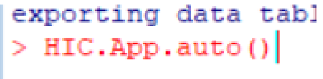
```
> library(HICbioclean)
> HIC.maint('MaintExcels','MaintCSVs')
```

9. The files should now all be formatted and saved as CSV files in the new specified folder.

Auto-validation

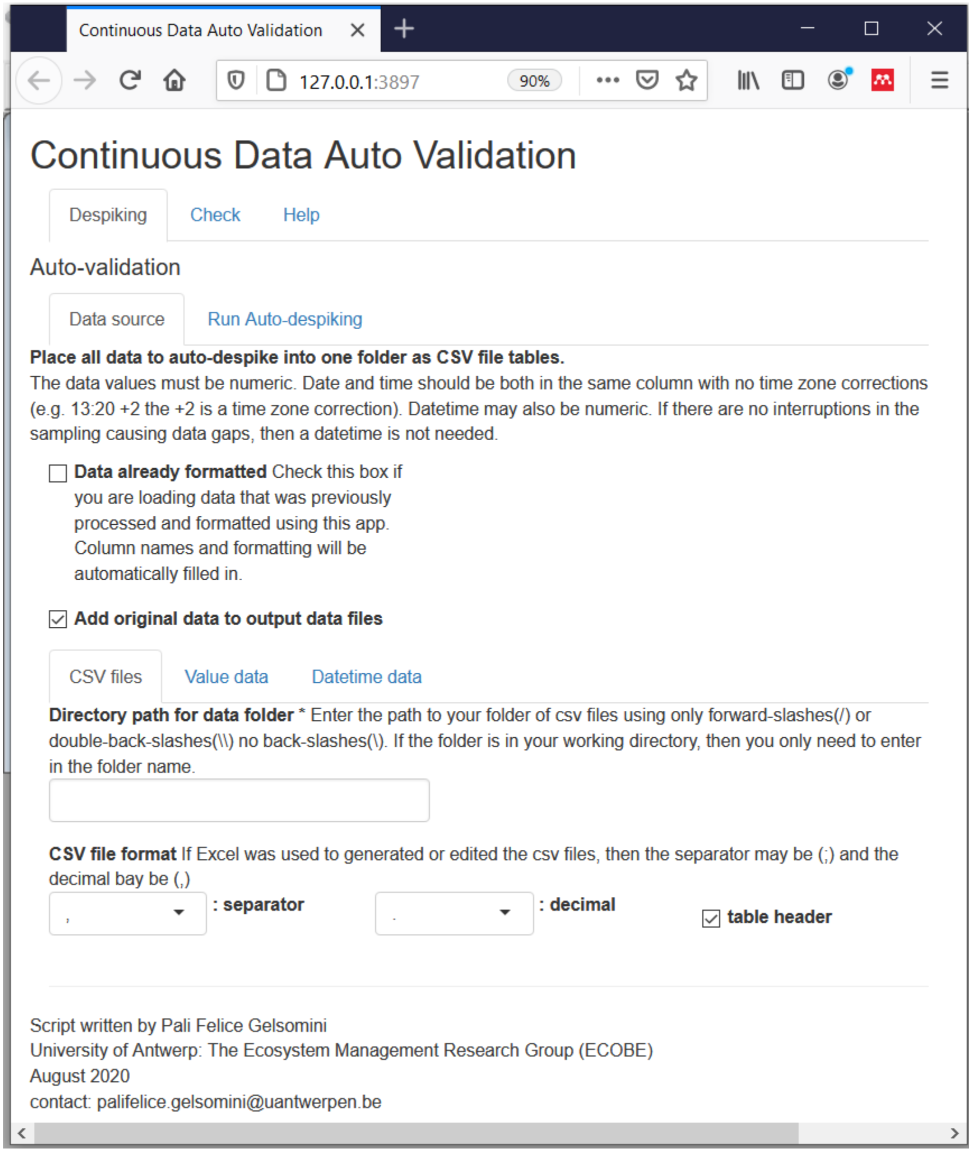
In this step the data runs through a series of automated processes to clean and despike the data.

10. Put your PPFD data all in one folder and place all the other data into another folder. PPFD has a special protocol and can’t be batch processed with the rest of the data. Do the below steps 11 through 34 to batch process the folder containing the rest of the data (no PPFD data).
11. If this is a new R session (you’ve closed the program and restarted it) you will have to enter **library(HICbioclean)** into the R console and press enter again first. Otherwise move onto the next step.
12. Enter **HIC.App.auto()** into the R console and press enter to open the auto-validation app.



```
exporting data tabl
> HIC.App.auto()
```

13. The R Shiny app should now be open in your web browser.



Continuous Data Auto Validation

Auto-validation

Place all data to auto-despike into one folder as CSV file tables.
The data values must be numeric. Date and time should be both in the same column with no time zone corrections (e.g. 13:20 +2 the +2 is a time zone correction). Datetime may also be numeric. If there are no interruptions in the sampling causing data gaps, then a datetime is not needed.

☐ **Data already formatted** Check this box if you are loading data that was previously processed and formatted using this app. Column names and formatting will be automatically filled in.

☑ **Add original data to output data files**

Directory path for data folder * Enter the path to your folder of csv files using only forward-slashes(/) or double-back-slashes(\\) no back-slashes(\). If the folder is in your working directory, then you only need to enter in the folder name.

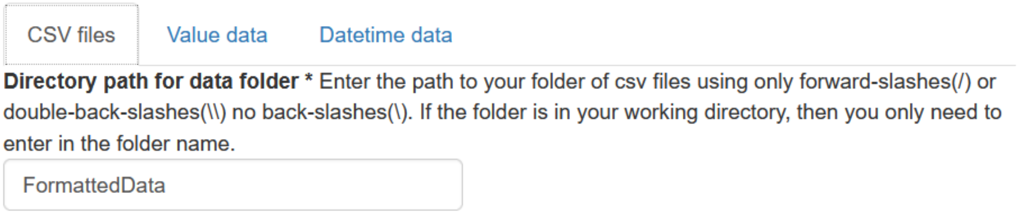
CSV file format If Excel was used to generated or edited the csv files, then the separator may be (;) and the decimal bay be (,)

, **: separator** . **: decimal** ☑ **table header**

Script written by Pali Felice Gelsomini
University of Antwerp: The Ecosystem Management Research Group (ECOBE)
August 2020
contact: palifelice.gelsomini@uantwerpen.be

14. All the default settings in the app are selected so that you should be able to process the HIC data without making any changes to them. If need be you can always change the settings in the app. See the in app descriptions of each setting.

15. Enter folder where you saved your formatted data from the previous step into the field "Directory path for data folder". For me that is "FormattedData" and because I saved that folder in my working directory I don't need a full path.



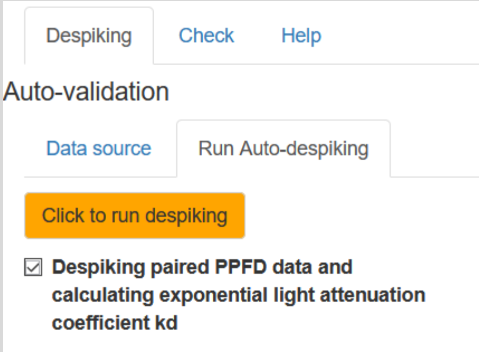
Directory path for data folder * Enter the path to your folder of csv files using only forward-slashes(/) or double-back-slashes(\\) no back-slashes(\). If the folder is in your working directory, then you only need to enter in the folder name.

FormattedData

16. The other tabs "value data" and "datetime data" contain the column names and data format of the values and datetime. If you are working with downloaded HIC database CSV files then the settings are already filled in for you correctly by default. There is no need to make any changes to these tabs.
17. Open the "Run Auto-despiking" tab.

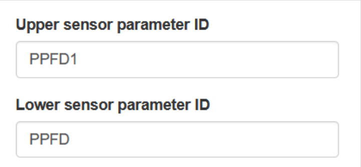
PPFD data specific instructions

a. If you are auto validating paired PPFD data for calculating light attenuation coefficient kd, the steps are all generally identical. But you have to tell the app that you are auto validating PPFD data. Under the tap "Despiking">>"Run Auto-despiking" check the box "Despiking paired PPFD data…".



☑ **Despiking paired PPFD data and calculating exponential light attenuation coefficient kd**

b. Make sure that in your data in your data the upper sensor is labeled PPFD1 and the lower sensor is labeled PPFD. You can change these options if need be.



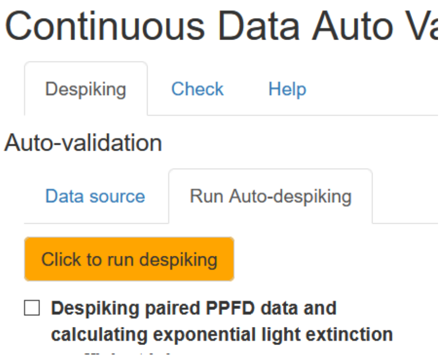
Upper sensor parameter ID

PPFD1

Lower sensor parameter ID

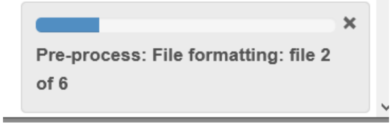
PPFD

18. and click the orange "Click to run despiking" button. This will run the full work flow with the the default settings.



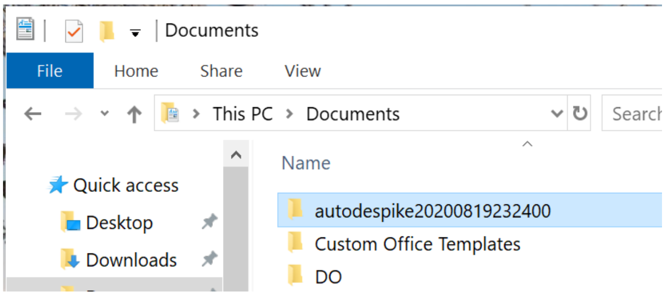
☐ **Despiking paired PPFD data and calculating exponential light extinction**

19. A progress bar will appear in the corner of the screen wile the process is running

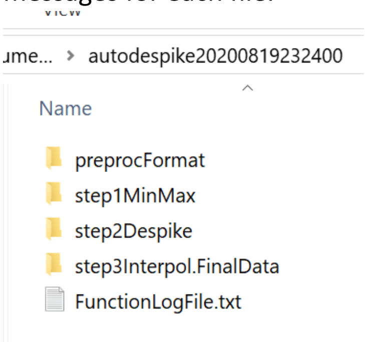


Pre-process: File formatting: file 2 of 6

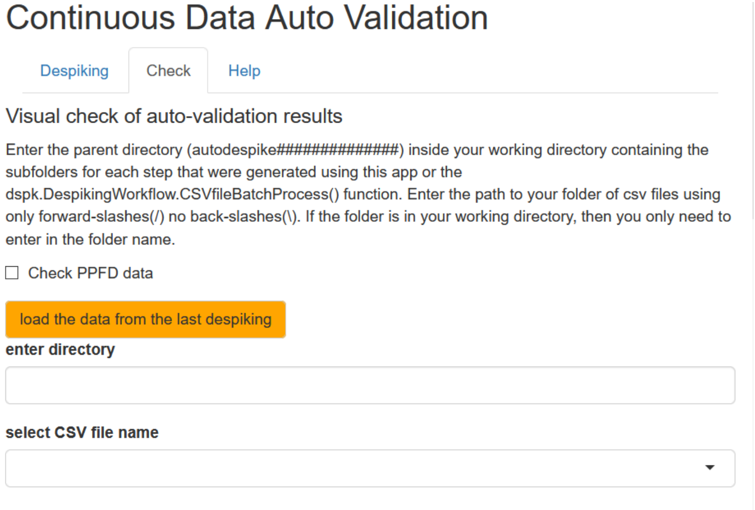
20. The auto validated data is saved in your working directory in the autogenerated folder “autodespikeYYYYMMDDHHMMSS”. If it is PPFD data you are processing it will be saved in “autoPPFDdespikeYYYYMMDDHHMMSS”.



21. Inside that folder you have 4 subfolders with the data generated at each step. The last folder labeled “FinalData” that the fully auto-validated data. There is a text file “FunctionLogFile.txt” that stores the function arguments for later reference and reproducibility, any calculated function arguments such as sampling interval and any error messages for each file.



22. After the process is complete you can move from the “Despiking” tab to the “Check” tab and click the orange “load the data from the last despiking” button.



Continuous Data Auto Validation

Visual check of auto-validation results

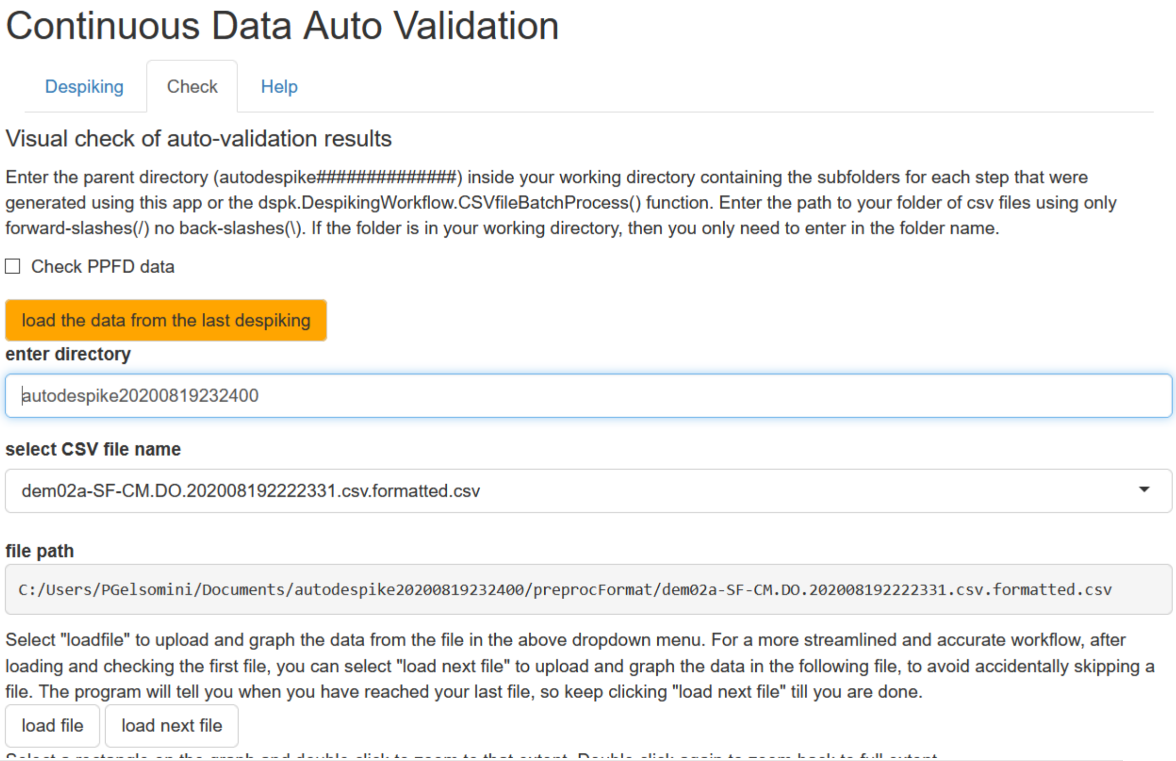
Enter the parent directory (autodespike##############) inside your working directory containing the subfolders for each step that were generated using this app or the dspk.DespikingWorkflow.CSVfileBatchProcess() function. Enter the path to your folder of csv files using only forward-slashes(/) no back-slashes(\). If the folder is in your working directory, then you only need to enter in the folder name.

□ Check PPFD data

enter directory

select CSV file name

23. This will populate the “enter directory” field with the folder that was just auto generated during the last auto-validation process. In the drop down menu you can select the different files, but don’t do this just yet. You can enter a folder name into the “enter directory” field but this folder must be in your working directory.



Continuous Data Auto Validation

Visual check of auto-validation results

Enter the parent directory (autodespike##############) inside your working directory containing the subfolders for each step that were generated using this app or the dspk.DespikingWorkflow.CSVfileBatchProcess() function. Enter the path to your folder of csv files using only forward-slashes(/) no back-slashes(\). If the folder is in your working directory, then you only need to enter in the folder name.

☐ Check PPFD data

enter directory: autodespike20200819232400

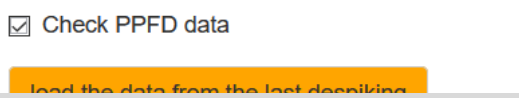
select CSV file name: dem02a-SF-CM.DO.202008192222331.csv.formatted.csv

file path: C:/Users/PGelsomini/Documents/autodespike20200819232400/preprocFormat/dem02a-SF-CM.DO.202008192222331.csv.formatted.csv

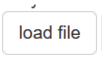
Select "loadfile" to upload and graph the data from the file in the above dropdown menu. For a more streamlined and accurate workflow, after loading and checking the first file, you can select "load next file" to upload and graph the data in the following file, to avoid accidentally skipping a file. The program will tell you when you have reached your last file, so keep clicking "load next file" till you are done.

Select a rectangle on the graph and double click to zoom to that extent. Double click again to zoom back to full extent.

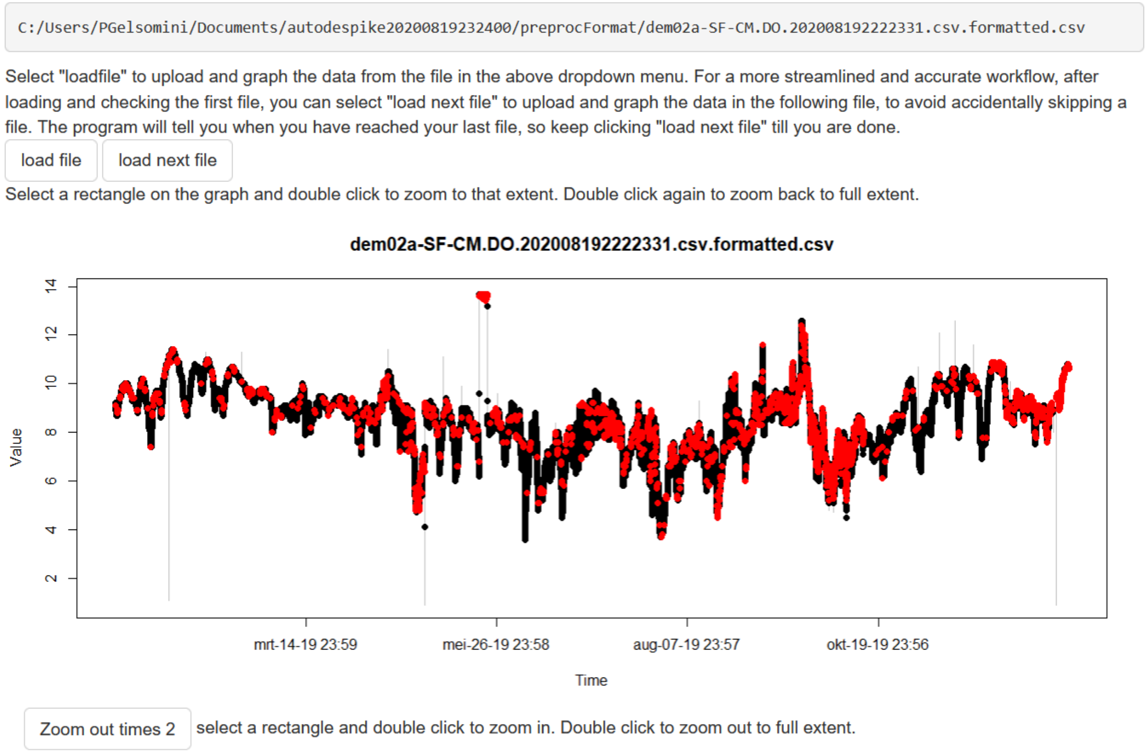
24. **If you are checking PPFD data** you must tell the app that it will be graphing paired PPFD data and light attenuation coefficient kd data. Check the box next to “Check PPFD data”.



☑ Check PPFD data

25. Click the load file button  to load the data from the first file.

26. The graph of the first dataset will appear just below.



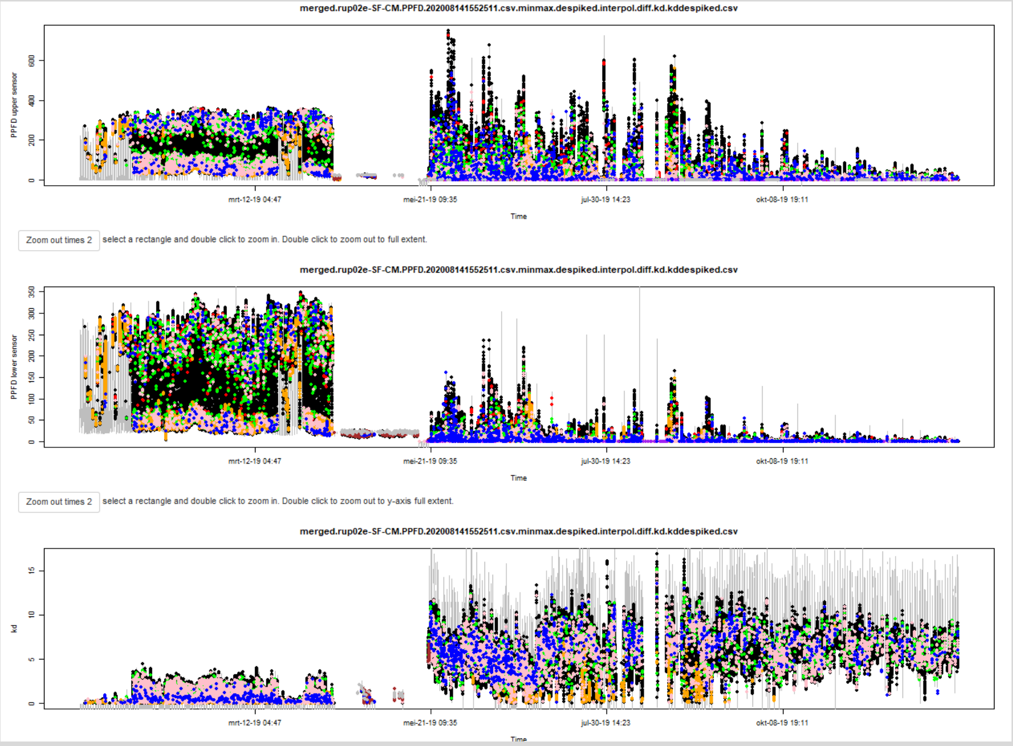
```
C:/Users/PGelsomini/Documents/autodespike20200819232400/preprocFormat/dem02a-SF-CM.DO.202008192222331.csv.formatted.csv
```

Select "loadfile" to upload and graph the data from the file in the above dropdown menu. For a more streamlined and accurate workflow, after loading and checking the first file, you can select "load next file" to upload and graph the data in the following file, to avoid accidentally skipping a file. The program will tell you when you have reached your last file, so keep clicking "load next file" till you are done.

Select a rectangle on the graph and double click to zoom to that extent. Double click again to zoom back to full extent.

select a rectangle and double click to zoom in. Double click to zoom out to full extent.

27. **If you are checking PPFD data** there will be three graphs. (upper sensor PPFD, lower sensor PPFD, and kd) and some extra state of value codes specific to PPFD validation procedures. The x-axis of all three graphs are linked to each other so the data will always be displayed perfectly aligned.

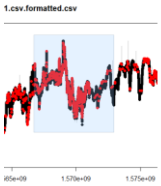


select a rectangle and double click to zoom in. Double click to zoom out to full extent.

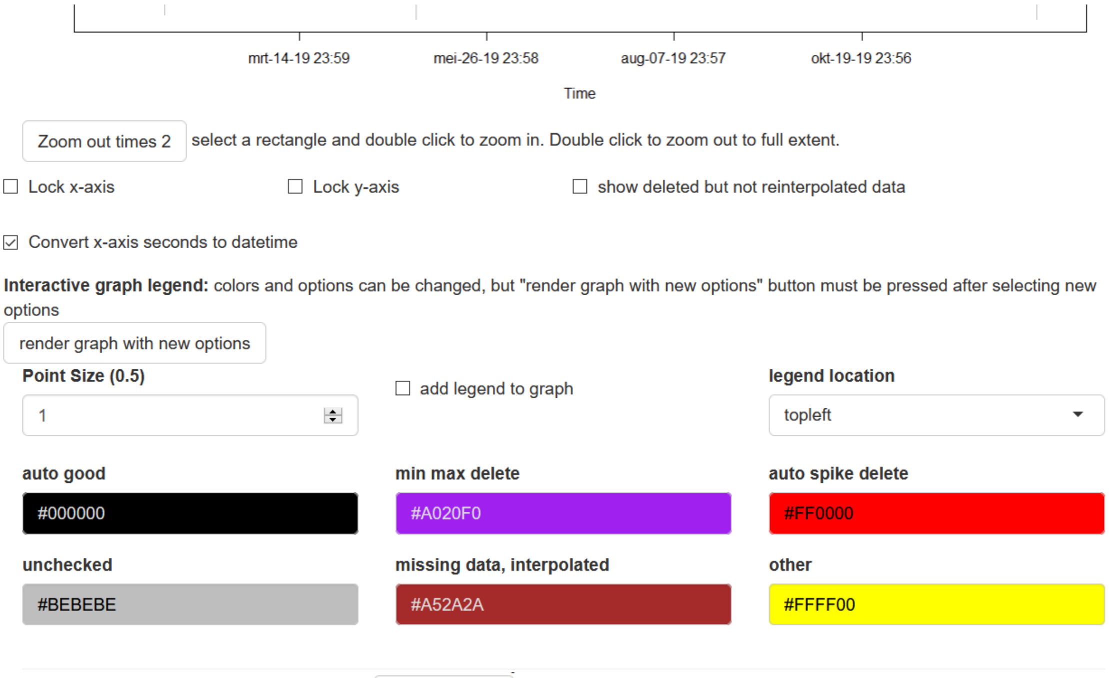
select a rectangle and double click to zoom in. Double click to zoom out to y-axis full extent.

merged.rup02e-SF-CM.PPFD.202008141552511.csv.minmax.despiked.interpol.diff.kd.kddespiked.csv
kd
0
5
10
15
mrt-12-19 04:47
mei-21-19 09:35
jul-30-19 14:23
okt-08-19 19:11
Time

28. The origin data is graphed in a thin gray line. The Validated and interpolated data is graphed as points. The points are color coded according to their state of value



29. Pull a box on the graph and double click the graph to zoom to that area. Double click again on the graph to zoom back out the full extent. Click the zoom out times 2 bottom Zoom out times 2 to increase your extent by 2. Use these methods to move around the graph to explore the data and see if the auto-despiking worked sufficiently.
30. Below the graphs there are more options and tools for exploring the data. You can lock the x and y axises while zooming. Show data that was deleted but didn't get reinterpolated. Convert x-axis from seconds to datetime. Point size. Legend colors. And you can add the legend to the graph. After you make changes to the interactive graph legend, you must click the button "render graph with new options". With large yearly datasets graph rendering is quite slow so it is best if this is done once you are done making changes to the legend.



Zoom out times 2 select a rectangle and double click to zoom in. Double click to zoom out to full extent.

□ Lock x-axis □ Lock y-axis □ show deleted but not reinterpolated data

☑ Convert x-axis seconds to datetime

Interactive graph legend: colors and options can be changed, but "render graph with new options" button must be pressed after selecting new options

render graph with new options

Point Size (0.5) 1

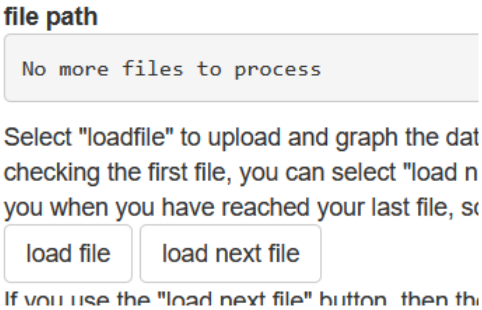
□ add legend to graph

legend location topleft

auto good	min max delete	auto spike delete
#000000	#A020F0	#FF0000
unchecked	**missing data, interpolated**	**other**
#BEBEBE	#A52A2A	#FFFF00

31. Click the "load next file" button load next file which is found next to the "load file" to move on to the next data file. This way you won't accidentally forget one. You may have to double click the graph to reset it to full extent to see all the data. Using the "load next file" button locks the dropdown menu, so you must first click the "reset file dropdown menu" button if you want to manually select a file, for example if you want to go back to a previous file.

32. Once you’ve reached your last file, the app will tell you “No more files to process”.



file path

```
No more files to process
```

Select "loadfile" to upload and graph the dat
checking the first file, you can select "load n
you when you have reached your last file, s

If you use the "load next file" button, then th

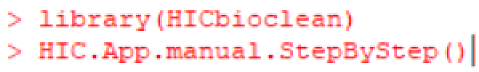
33. If you are not happy with the results you can change the settings of the auto-validation under the tab “Despiking”>>”Run Auto-despiking”. See the in app instructions on their use and for additional help and details on the algorithms see the “Help” tab.
34. Now either return to step 22 and enter a new parent directory in the “enter directory” field, return to step 14 to auto-validate a new set of data or close your app browser window and move onto the next step which is the manual-validation and final export.
35. Repeat steps 11 through 34 to batch process the folder containing the PPFD data. Pay special attention to the PPFD specific instructions.

To find your working directory and for more details into the algorithms see the ‘Help’ tab.

Manual-validation and final export

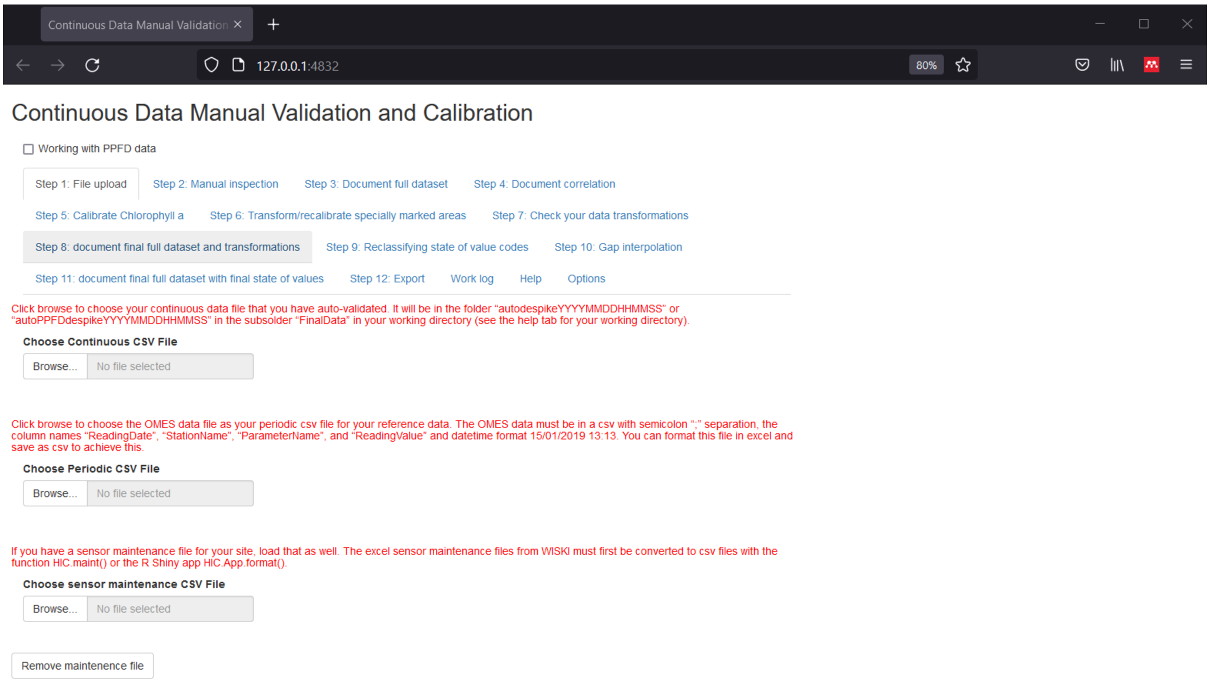
The last step after the formatting and the auto-validation. The other steps can be done in R code without the use of the Shiny apps, however this step must be done in the Shiny app because this is a manual check so it must be done visually.

36. If this is a new R session (you've closed the program and restarted it) you will have to enter **library(HICbioclean)** into the R console and press enter again first. Otherwise move onto the next step.
37. Type **HIC.App.manual.StepByStep()** into the R console and press enter. This will open an R Shiny app which will guide you through the manual cleaning process step by step. All these instructions are also provided inside the app.



```
> library(HICbioclean)
> HIC.App.manual.StepByStep()
```

38. The app should open in your web browser. The instructions are always written in the app in red to help guide you through the cleaning process.



Continuous Data Manual Validation
127.0.0.1:4832
80%

Continuous Data Manual Validation and Calibration

☐ Working with PPFD data

Step 1: File upload
Step 2: Manual inspection
Step 3: Document full dataset
Step 4: Document correlation
Step 5: Calibrate Chlorophyll a
Step 6: Transform/recalibrate specially marked areas
Step 7: Check your data transformations
Step 8: document final full dataset and transformations
Step 9: Reclassifying state of value codes
Step 10: Gap interpolation
Step 11: document final full dataset with final state of values
Step 12: Export
Work log
Help
Options

Click browse to choose your continuous data file that you have auto-validated. It will be in the folder "autodespikeYYYYMMDDHHMMSS" or "autoPPFDdespikeYYYYMMDDHHMMSS" in the subsolder "FinalData" in your working directory (see the help tab for your working directory).

Choose Continuous CSV File

Browse... No file selected

Click browse to choose the OMES data file as your periodic csv file for your reference data. The OMES data must be in a csv with semicolon ";" separation, the column names "ReadingDate", "StationName", "ParameterName", and "ReadingValue" and datetime format 15/01/2019 13:13. You can format this file in excel and save as csv to achieve this.

Choose Periodic CSV File

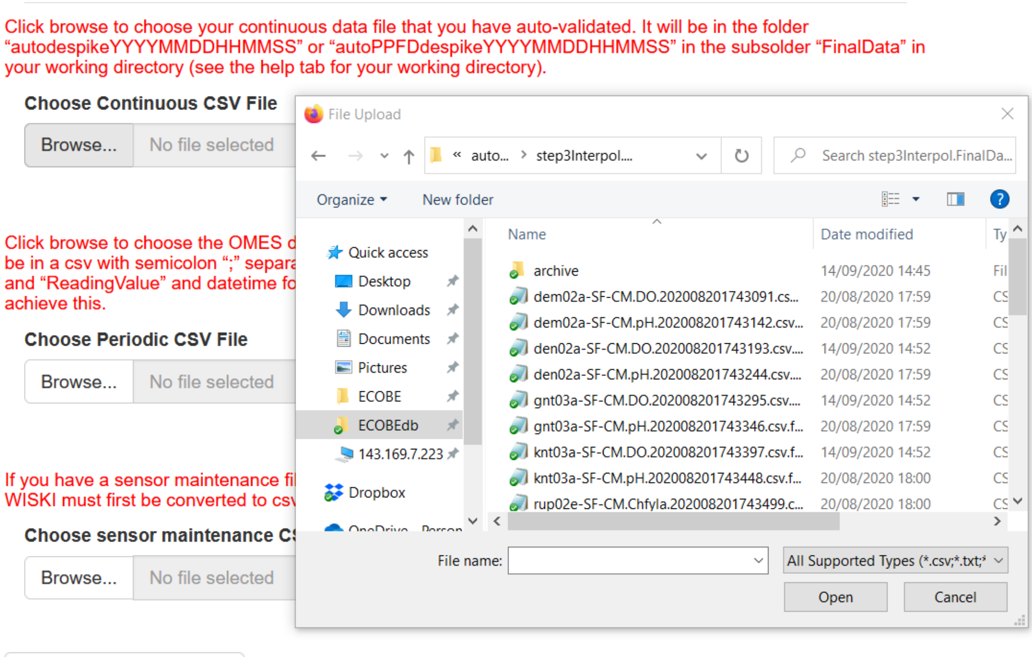
Browse... No file selected

If you have a sensor maintenance file for your site, load that as well. The excel sensor maintenance files from WISKI must first be converted to csv files with the function HIC.maint() or the R Shiny app HIC.App.format().

Choose sensor maintenance CSV File

Browse... No file selected

39. Click browse to choose your continuous data file that you have auto-validated. It will be in the folder “autodespikeYYYYMMDDHHMMSS” or “autoPPFDdespikeYYYYMMDDHHMMSS” in the subsolder “FinalData” in your working directory (see the help tab for your working directory).



Click browse to choose your continuous data file that you have auto-validated. It will be in the folder “autodespikeYYYYMMDDHHMMSS” or “autoPPFDdespikeYYYYMMDDHHMMSS” in the subsolder “FinalData” in your working directory (see the help tab for your working directory).

Choose Continuous CSV File

Browse... No file selected

Click browse to choose the OMES d
be in a csv with semicolon “;” separa
and “ReadingValue” and datetime fo
achieve this.

Choose Periodic CSV File

Browse... No file selected

If you have a sensor maintenance fi
WISKI must first be converted to csv

Choose sensor maintenance C

Browse... No file selected

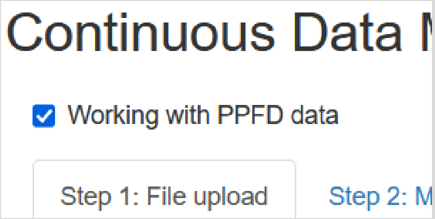
Name	Date modified	Ty
archive	14/09/2020 14:45	Fil
dem02a-SF-CM.DO.202008201743091.cs...	20/08/2020 17:59	CS
dem02a-SF-CM.pH.202008201743142.csv...	20/08/2020 17:59	CS
den02a-SF-CM.DO.202008201743193.csv....	14/09/2020 14:52	CS
den02a-SF-CM.pH.202008201743244.csv...	20/08/2020 17:59	CS
gnt03a-SF-CM.DO.202008201743295.csv....	14/09/2020 14:52	CS
gnt03a-SF-CM.pH.202008201743346.csv.f...	20/08/2020 17:59	CS
knt03a-SF-CM.DO.202008201743397.csv.f...	14/09/2020 14:52	CS
knt03a-SF-CM.pH.202008201743448.csv.f...	20/08/2020 18:00	CS
rup02e-SF-CM.Chfyla.202008201743499.c...	20/08/2020 18:00	CS

File name:

All Supported Types (*.csv;*.txt;*

40. Click browse to choose the OMES data file as your periodic csv file for your reference data. The OMES data must be in a csv with semicolon “;” separation, the column names “ReadingDate”, “StationName”, “ParameterName”, and “ReadingValue” and datetime format 15/01/2019 13:13. You can format this file in excel and save as csv to achieve this.

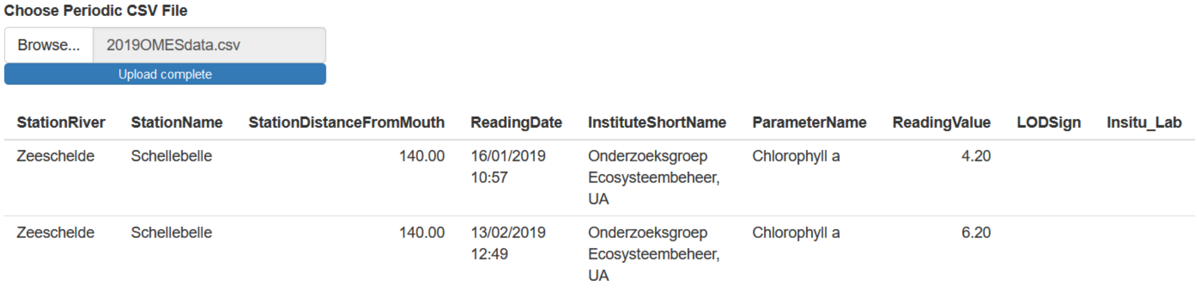
41. Just as before, here you must also say **if you are working with PPFD data** because it is graphed in three graphs instead of one and all three parameters (upper PPFD, lower PPFD and kd) are dealt with simultaneously.



Continuous Data I

☑ Working with PPFD data

42. Once the data is uploaded into the app, you will be able to see the first few lines of the files. The periodic OMES file gets subsetted based on the parameter and the location, so if you don’t see the periodic OMES file, then there is either no reference site for the continuous data that you have open, or there is no data on that parameter.



Choose Periodic CSV File

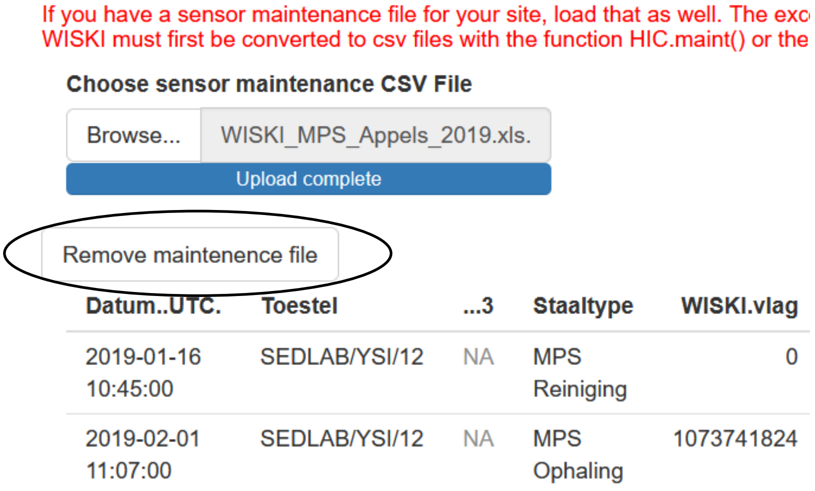
Browse... 2019OMESdata.csv

Upload complete

StationRiver	StationName	StationDistanceFromMouth	ReadingDate	InstituteShortName	ParameterName	ReadingValue	LODSign	Insitu_Lab
Zeeschelde	Schellebelle	140.00	16/01/2019 10:57	Onderzoeksgroep Ecosysteembeheer, UA	Chlorophyll a	4.20		
Zeeschelde	Schellebelle	140.00	13/02/2019 12:49	Onderzoeksgroep Ecosysteembeheer, UA	Chlorophyll a	6.20		

43. If you have a sensor maintenance file for your site, load that as well. The excel sensor maintenance files from WISKI must first be converted to csv files with the function HIC.maint() or the R Shiny app HIC.App.format(). See the earlier section on formatting.

44. If you don't have a maintenance file for the current site that you are cleaning, you can remove the currently uploaded file with the remove maintenance file button.



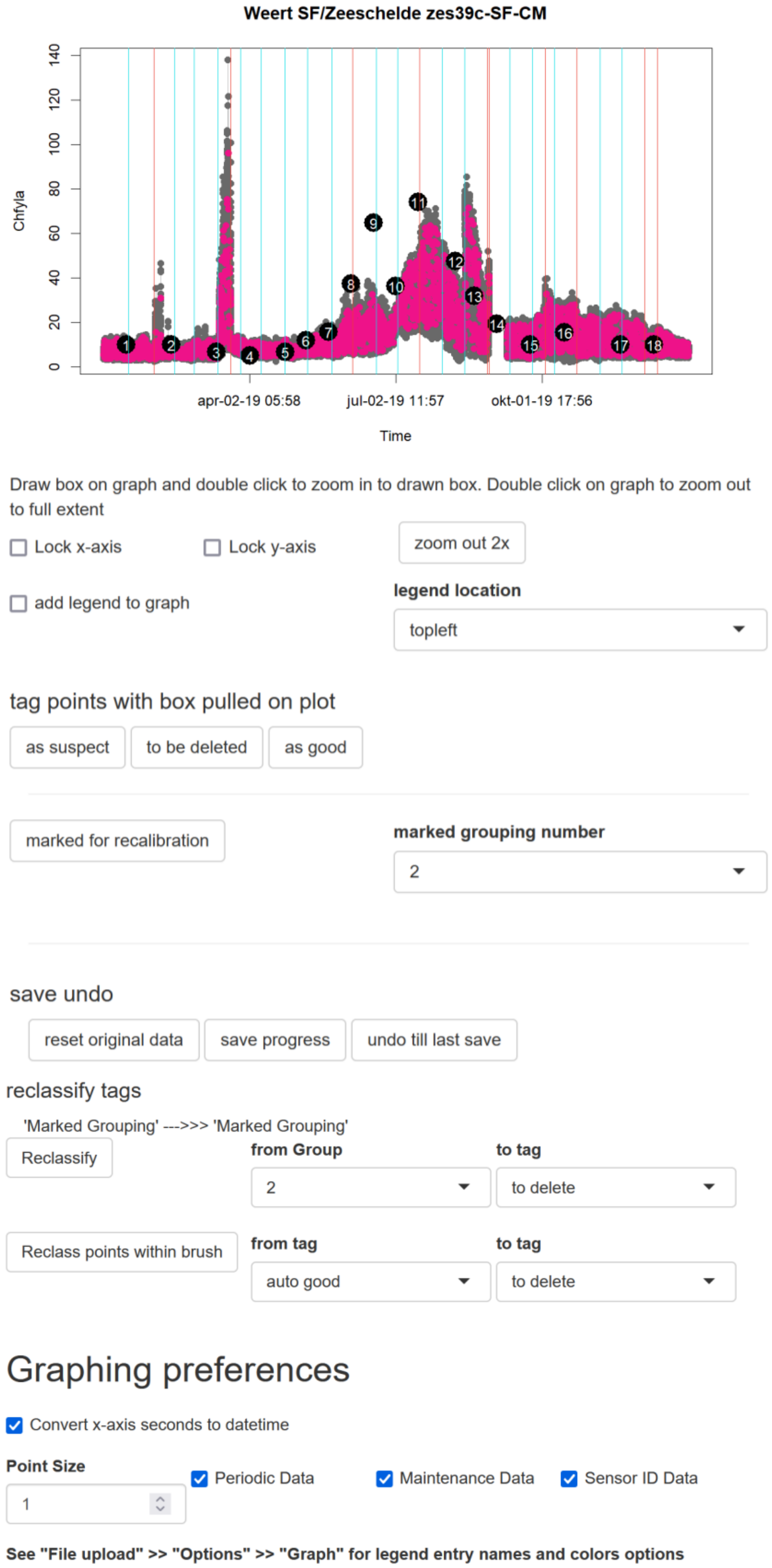
If you have a sensor maintenance file for your site, load that as well. The exc
WISKI must first be converted to csv files with the function HIC.maint() or the

Choose sensor maintenance CSV File

Browse... WISKI_MPS_Appels_2019.xls.

Datum..UTC.	Toestel	...3	Staaltype	WISKI.vlag
2019-01-16 10:45:00	SEDLAB/YSI/12	NA	MPS Reiniging	0
2019-02-01 11:07:00	SEDLAB/YSI/12	NA	MPS Ophaling	1073741824

6. Once the data is uploaded you can move to “Step 2: Manual inspection” tab. Here the data will be graphed. The continuous data is graphed colored by state of value. The reference site data is graphed over it. The sensor maintenance times are graphed as horizontal lines. You have tools for formatting the graph, exploring the graph and marking sections of data.



a

Draw box on graph and double click to zoom in to drawn box. Double click on graph to zoom out to full extent

a ☐ Lock x-axis ☐ Lock y-axis [zoom out 2x]

b ☐ add legend to graph

legend location

topleft ▾

tag points with box pulled on plot

c [as suspect] [to be deleted] [as good]

d [marked for recalibration]

marked grouping number

2 ▾

save undo

e [reset original data] [save progress] [undo till last save]

reclassify tags

'Marked Grouping' --->>> 'Marked Grouping'

f [Reclassify] **from Group** 2 ▾ **to tag** to delete ▾

g [Reclass points within brush] **from tag** auto good ▾ **to tag** to delete ▾

Graphing preferences

h ☑ Convert x-axis seconds to datetime

i **Point Size** 1 ☑ Periodic Data ☑ Maintenance Data ☑ Sensor ID Data

See "File upload" >> "Options" >> "Graph" for legend entry names and colors options

List of tools on the tab:

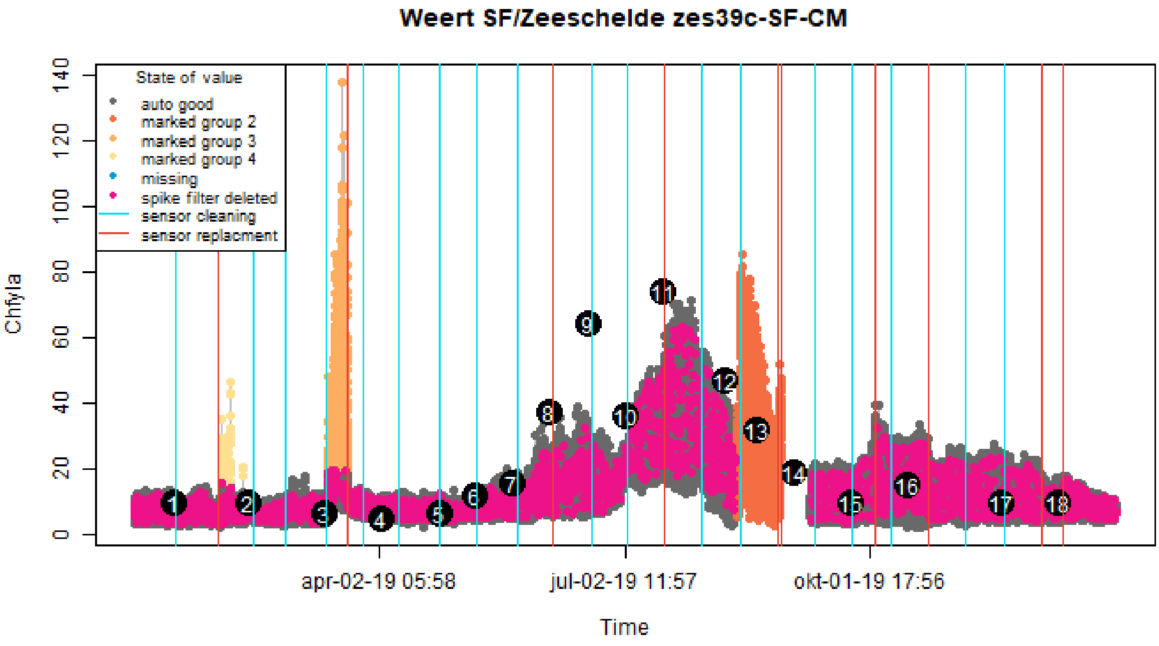
a. The graph can be zoomed in on by drawing a box on the graph and double clicking. Double click on the graph again to zoom out to full extent. Use the “Zoom out 2x” button to increase your extent by 2 times. You can lock the x and y axis while zooming.
b. You can add a legend and select it’s location so that it doesn’t overlap data points.
c. You can pull a box around sections of your data on the plot and mark them as “suspect”, “good” or “to be deleted”.
d. You can pull a box around a section of data on the graph and mark that section as a “marked for recalibration”. These marked groups can be mathematically transformed later, or have their state of value set to a non-work-class state of value such as “good”, “suspect”, or “estimate”. Upon export, only non-work-class state of values will be kept, all others will be set to state of value “good”. Mathematical transformations get applied to marked groups and you have 8 groups to works with numbered 2 through 9.
e. Your work can be saved, reverted back to the previous save or reset to the original data.
f. You can reclassify a specific marked group to another marked group, “to be deleted” or “good”, “suspect”, or “estimate”.
g. You can reclassify state of values of data within the box that you pulled on the plot form one state of value to a marked group, “to be deleted”, “good”, “suspect”, or “estimate”.
h. This is the preference to convert the UNIX numeric time x axis to date time format.
i. This is the option for point size and to turn layers on and off.

47. The recommended workflow is:
    1. At full extent (seeing all the years data) mark in different groups with the “mark for recalibration” button (d) all the sections of data that you want to delete, recalibrate and/or mark as suspect. You don’t have to mark anything as “good” since the data will be automatically marked as “good” upon export.
    2. Zoom in on each section of your data to inspect it. Polish up the boarders of those marked groups you just made. To unmark a point it is easiest to just highlight it and use the “as good” button (c). Data noise and spikes can be immediately marked “to be deleted” by highlighting it and using the respective button (c). Suspect data that doesn’t need to be recalibrated or transformed can be immediately marked “as suspect” by highlighting it and using the respective button (c). Data that needs to be transformed or recalibrated needs to be part of a marked group (d).
       Upon close inspection of the marked groups that you made, decide if you need to transform or recalibrate those data. For marked groups that don’t need to be transformed or recalibrated, you should use the “reclassify” button (f) to change your marked groups into “to be deleted”, “suspect” or “good” before you move on. For the next step you only want the marked groups for the data you want to transform or recalibrate. Make a note of each of the transformations you wish to perform on each remaining marked group.
    3. Make sure you zoom in on each month to inspect the data to make sure you didn’t miss anything.
    4. Save your work (e) before moving on to the data transformations so that you can always go back to “step 2” and undo your changes and then try and redo your data transformations if you mess up too badly.

48. **Make sure you record everything you did with copies of the plot in a word document for records and quality control!!**

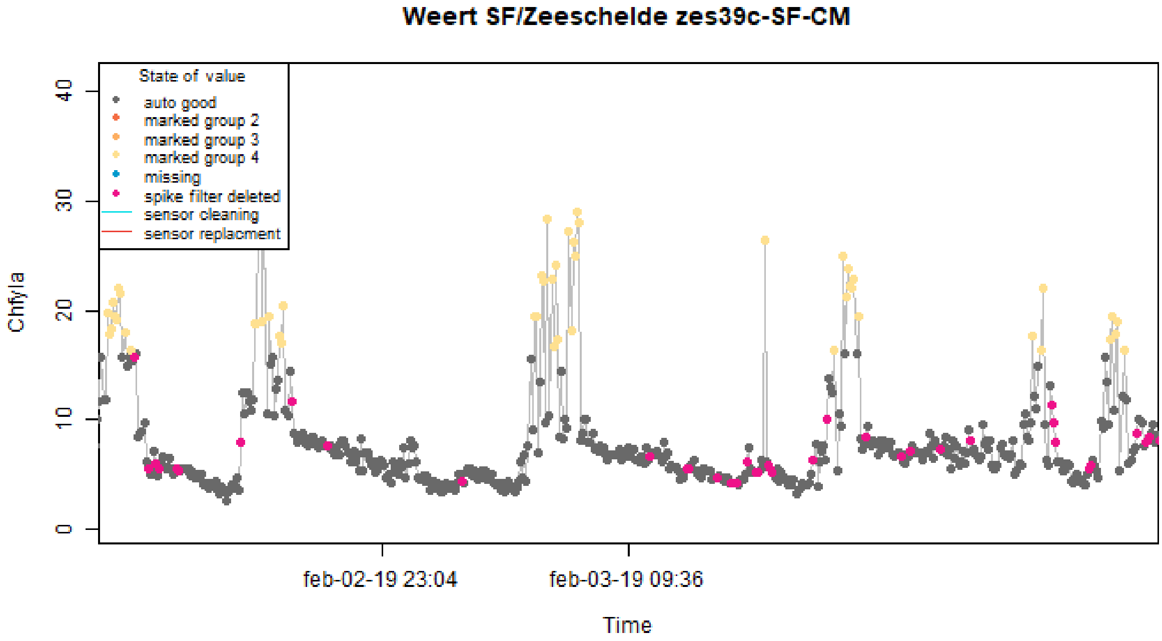
Workflow example:

Step 1:

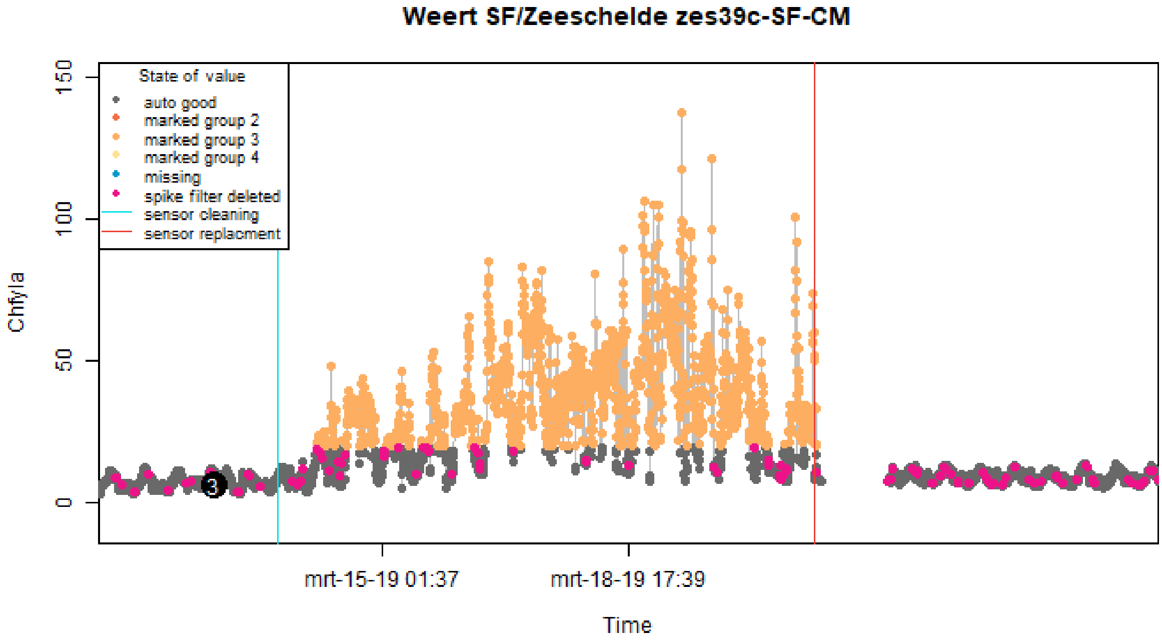


At full extent I marked **(d)** three areas that seem out of the ordinary compared to the yearly trends. Two periods of high chlorophyl in the early season and a period in summer that doesn't follow the reference data trend.

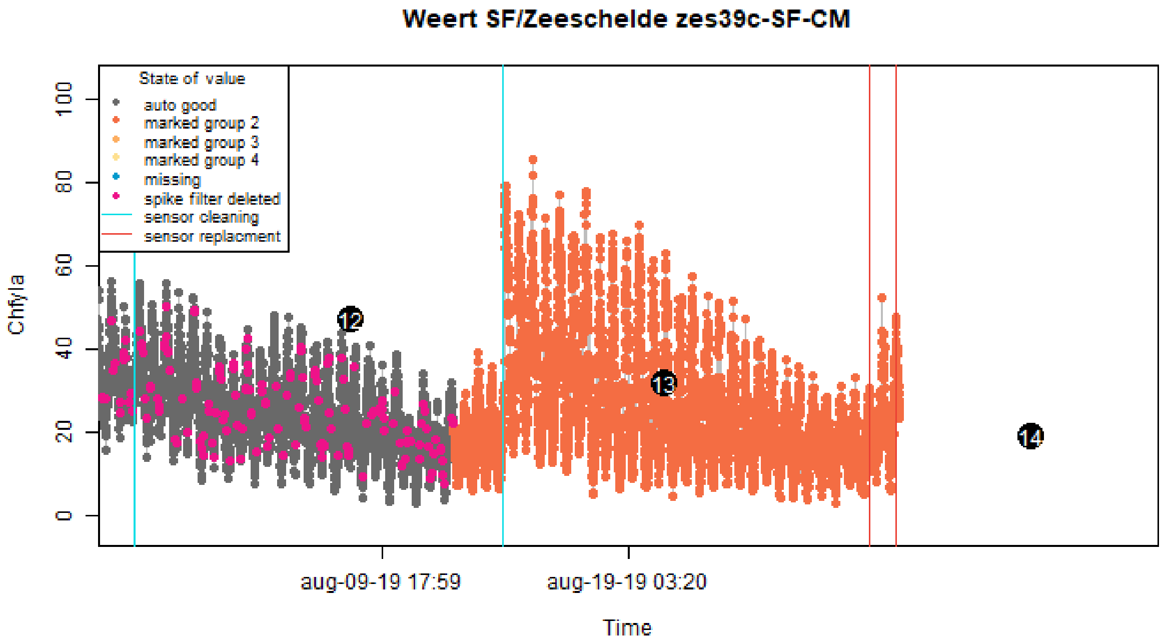
Step 2:



Zooming in on the early season peaks it seems as if these are simply spikes in the data possibly due to sensor interference. The chlorophyl spikes follow a day night cycle which isn't unusual however the shape of the curve is not normally seen and the values are much too high compared to the surrounding data and especially for that season. The marked group 4 will be set as "to delete" with the "reclassify" button **(f)** and the remaining points that weren't marked before will be highlighted and marked "to be deleted" **(c)**.

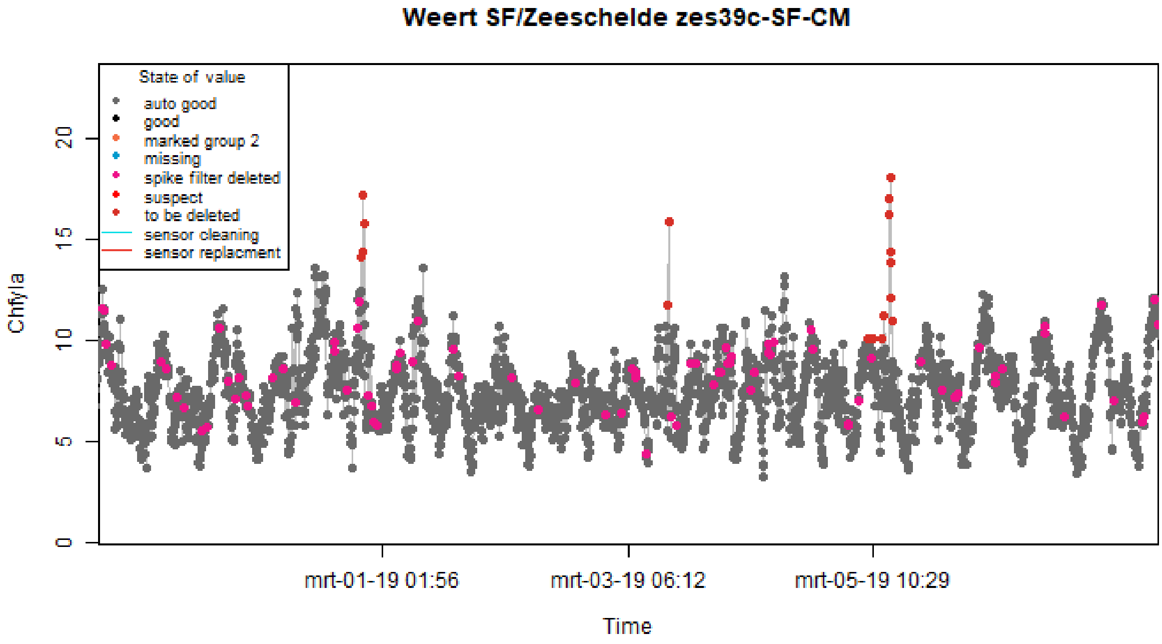


Zooming in on the next suspicious area you can see that these high levels of chlorophyl stop exactly at the sensor change. The likely hood of this being naturally occurring is rather slim and this could easily be explained by biofouling. The marked group 3 will be set as "suspect" with the "reclassify" button **(f)** and the remaining points that weren't marked before will be highlighted and marked "as suspect" **(c)**.



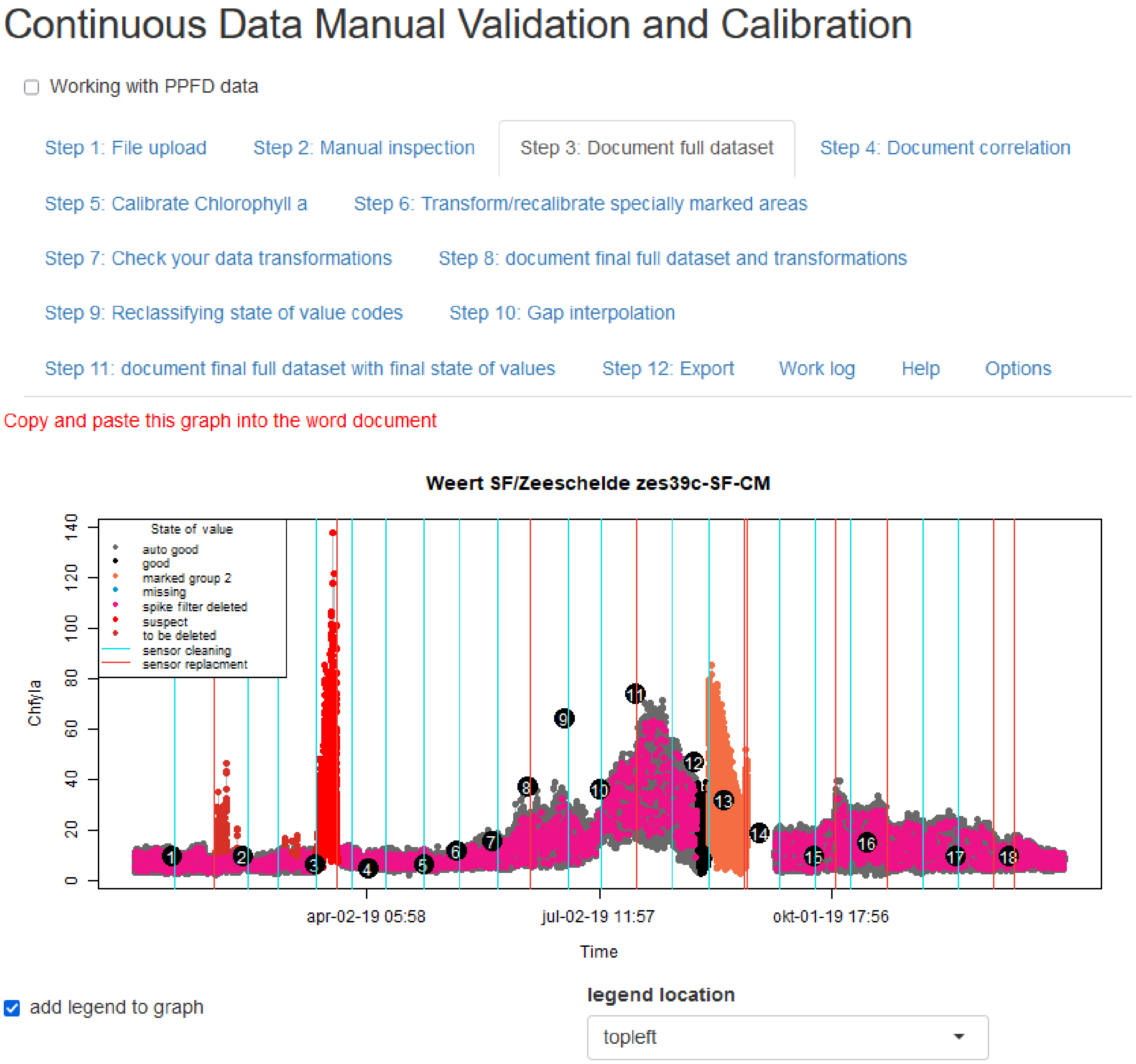
Zooming in on the last suspicious area of data we can see that exactly at the sensor cleaning there is a shift in the magnitude of the data. This is also highly unlikely to be natural and the reference data doesn't agree with it. Probably they made a mistake with the sensor settings and this data simply can be recalibrated. It will be left as "Marked group 2" so that that group can be recalibrated later. I accidentally marked some of the points before the sensor change as well and these can be highlighted and marked "as good" **(c).**

Step 3:



You should still look at each month up close to see if there are any spikes you have missed. Here in March there were a few more spikes which I marked "to be deleted" **(c)**.

49. Move on to the “Step 3: Document full dataset” tab and document the full dataset. You can copy this graph into a word document. Adjust the legend location to better see the data.



Continuous Data Manual Validation and Calibration

☐ Working with PPFD data

Step 1: File upload | Step 2: Manual inspection | Step 3: Document full dataset | Step 4: Document correlation

Step 5: Calibrate Chlorophyll a | Step 6: Transform/recalibrate specially marked areas

Step 7: Check your data transformations | Step 8: document final full dataset and transformations

Step 9: Reclassifying state of value codes | Step 10: Gap interpolation

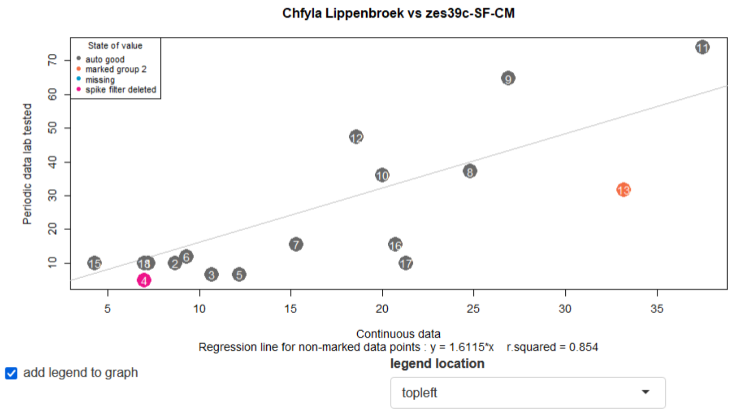
Step 11: document final full dataset with final state of values | Step 12: Export | Work log | Help | Options

Copy and paste this graph into the word document

☑ add legend to graph

legend location: topleft

50. Move to the “Step 4: Document correlation” tab and copy the correlation plot into the word document for records. Adjust the legend location to better see the data. If there is no reference data then there will be no plot. Note the regression equation below the graph; this is your calibration curve for your chlorophyl data.

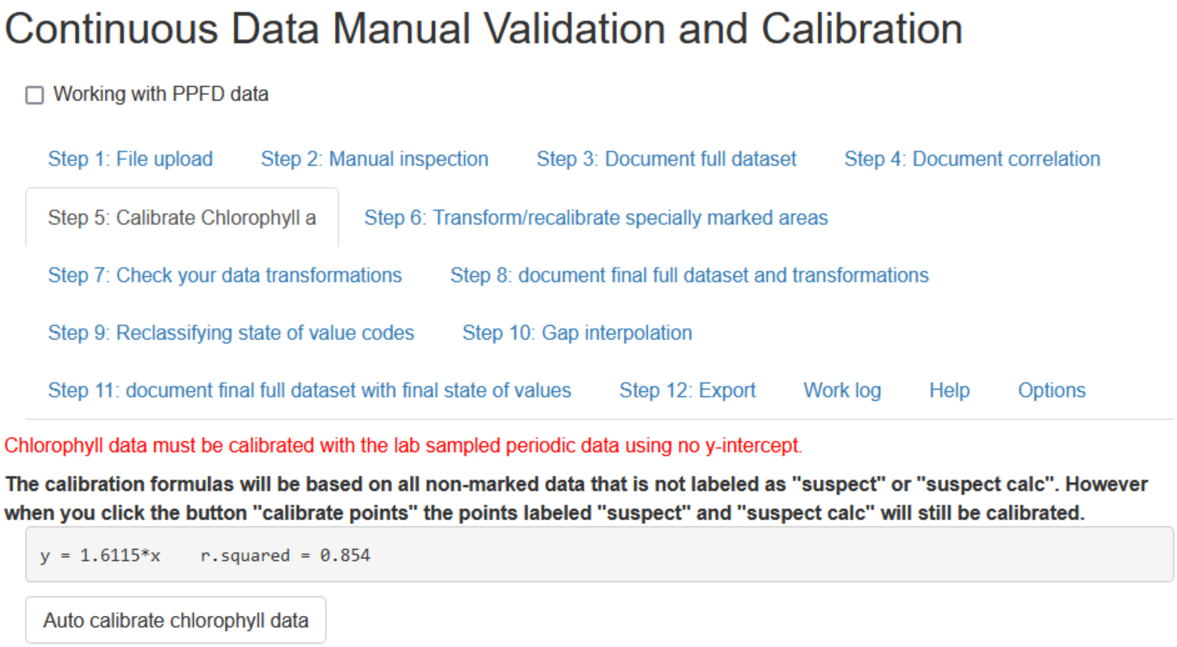


Regression line for non-marked data points : y = 1.6115*x r.squared = 0.854

☑ add legend to graph

legend location: topleft

51. The next tab “Step 5: Calibrate Chlorophyll a” will automaticaly calculate the calibration curve for the chlorophyll a data for that year based on the non marked and non suspect data and your reference data. Click the “Auto calibrate chlorophyll data” button to calibrate the data. If the data isn’t chlorophyll then this tab will be empty.



Continuous Data Manual Validation and Calibration

☐ Working with PPFD data

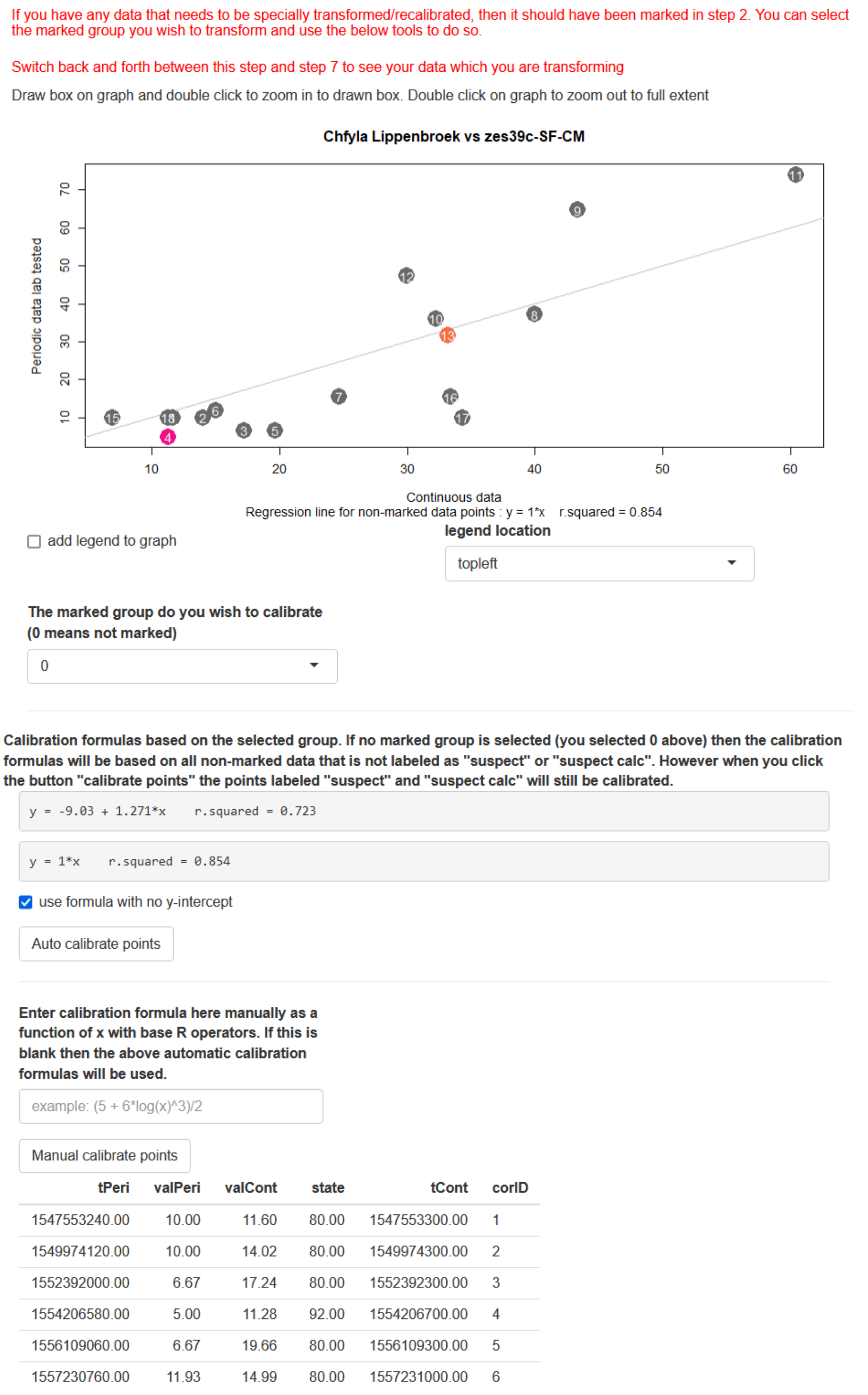
Step 1: File upload
Step 2: Manual inspection
Step 3: Document full dataset
Step 4: Document correlation
Step 5: Calibrate Chlorophyll a
Step 6: Transform/recalibrate specially marked areas
Step 7: Check your data transformations
Step 8: document final full dataset and transformations
Step 9: Reclassifying state of value codes
Step 10: Gap interpolation
Step 11: document final full dataset with final state of values
Step 12: Export
Work log
Help
Options

Chlorophyll data must be calibrated with the lab sampled periodic data using no y-intercept.

The calibration formulas will be based on all non-marked data that is not labeled as "suspect" or "suspect calc". However when you click the button "calibrate points" the points labeled "suspect" and "suspect calc" will still be calibrated.

```
y = 1.6115*x    r.squared = 0.854
```

52. Now go to the “Step 6: Transform/recalibrate specially marked areas” tab. Here you can perform mathematical transformations on your data.



If you have any data that needs to be specially transformed/recalibrated, then it should have been marked in step 2. You can select the marked group you wish to transform and use the below tools to do so.

Switch back and forth between this step and step 7 to see your data which you are transforming

Draw box on graph and double click to zoom in to drawn box. Double click on graph to zoom out to full extent

a

Chfyla Lippenbroek vs zes39c-SF-CM
Periodic data lab tested
10
20
30
40
50
60
70
10
20
30
40
50
60
Continuous data
Regression line for non-marked data points : y = 1*x r.squared = 0.854
15
18
4
2
6
3
5
7
12
10
13
16
17
8
9
11

□ add legend to graph

legend location

topleft ▾

b **The marked group do you wish to calibrate (0 means not marked)**

0 ▾

Calibration formulas based on the selected group. If no marked group is selected (you selected 0 above) then the calibration formulas will be based on all non-marked data that is not labeled as "suspect" or "suspect calc". However when you click the button "calibrate points" the points labeled "suspect" and "suspect calc" will still be calibrated.

c
```
y = -9.03 + 1.271*x    r.squared = 0.723
```

d
```
y = 1*x    r.squared = 0.854
```

e ☑ use formula with no y-intercept

f Auto calibrate points

Enter calibration formula here manually as a function of x with base R operators. If this is blank then the above automatic calibration formulas will be used.

g example: (5 + 6*log(x)^3)/2

h Manual calibrate points

i

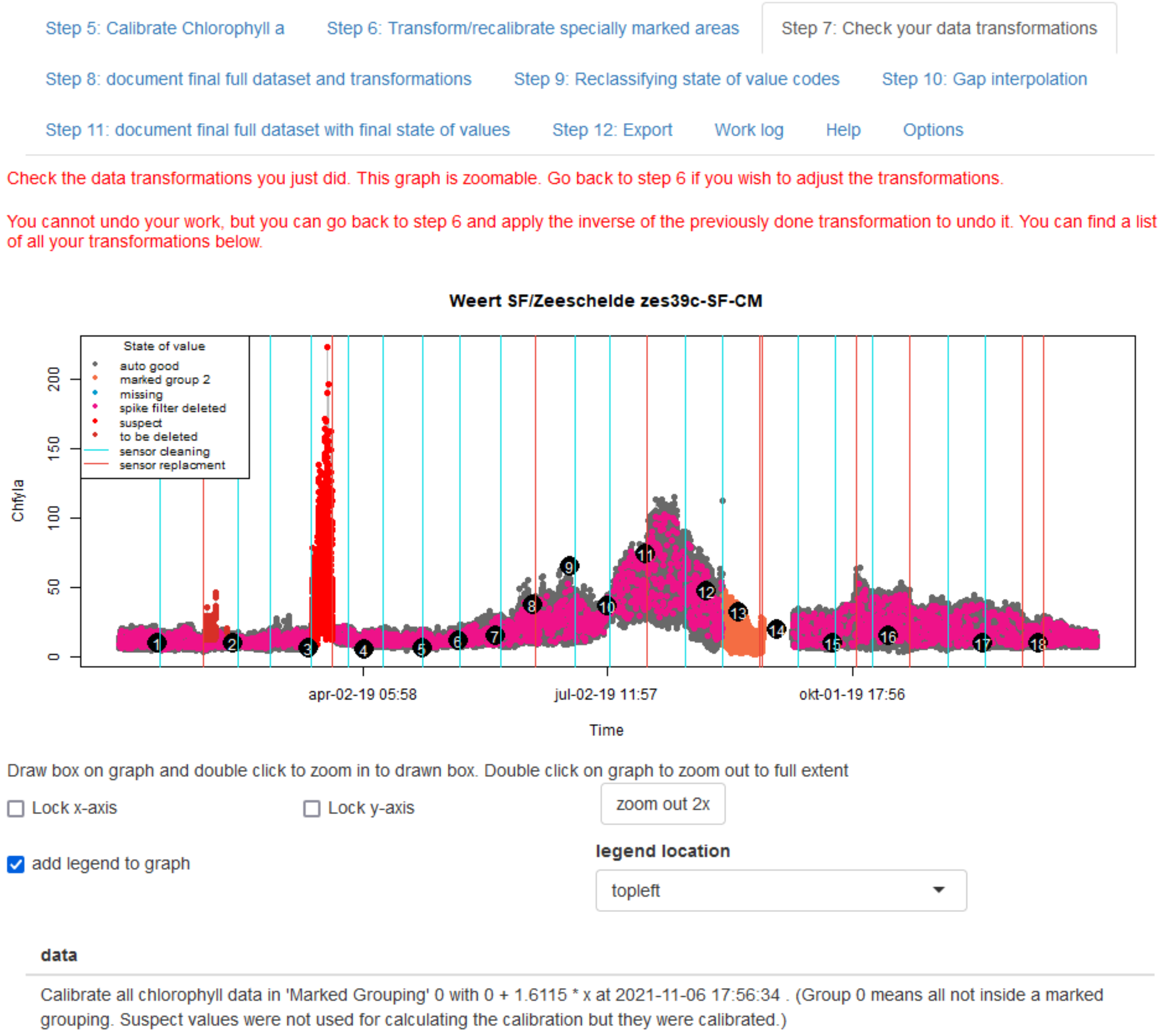
tPeri	valPeri	valCont	state	tCont	corID
1547553240.00	10.00	11.60	80.00	1547553300.00	1
1549974120.00	10.00	14.02	80.00	1549974300.00	2
1552392000.00	6.67	17.24	80.00	1552392300.00	3
1554206580.00	5.00	11.28	92.00	1554206700.00	4
1556109060.00	6.67	19.66	80.00	1556109300.00	5
1557230760.00	11.93	14.99	80.00	1557231000.00	6

List of tools on the tab:

a. If you have reference data, then a correlation graph and linear regression equations will be displayed. The linear regression equations and the trend line on the graph are based on the marked grouping that you have selected in the "The marked group do you wish to calibrate" dropdown menu **(b)** and with or without a y intercept based on your selection of checkbox **e**.
b. The marked grouping that you have selected in the "The marked group do you wish to calibrate" dropdown menu is also the grouping that the mathematical transformations will be done on. So even if you have no reference data, it is still important to select the correct grouping. If you have group 0 selected (which stands for not marked), then any data point that is marked will not be transformed, but all other data will be. You will have to do each group separately if you want to do the same transformation on all your data.
c. This is the linear regression line with y intercept for the selected marked grouping as compared to the reference data. If you selected marked grouping 0 then the regression is based on all data that is not marked and not suspect.
d. This is the linear regression line without y intercept for the selected marked grouping as compared to the reference data. If you selected marked grouping 0 then the regression is based on all data that is not marked and not suspect.
e. Check this box if you don't want a y intercept for the auto calibration or on the formula displayed on the plot.
f. The "Auto calibrate points" button will use the linear regression equation to calibrate the points. With the check box "use formula with no y-intercept" **(e)** you can chose which linear regression you want to use.
g. Type your custom formula in here. Enter the formula with base R arguments and as a function of x. e.g. (5+6*log(x)^3)/2
h. Use this button to transform your data in the selected marked group using your custom formula. With functionality works even when there is no reference data.
i. If you have reference data, the reference-data-point next to the nearest in time continuous-data-point is shown in a table at the bottom of the page.

53. You have two tools to calibrate with. An automatic tool which uses linear regression based on your reference data and a manual tool. You need to select the correct marked group on which to perform the transformation **(b).** You can use the correlation plot to see you're your data compares to the reference data **(a)**. If you want to use an automatic calibration then select if you want to use a y intercept **(e)** and click the button "auto calibrate points" **(f)**. If you want to use a manual transformation then write your formula into the field **(g)** as a function of x in base R code e.g. (5+6*log(x)^3)/2 and then click the button **(h)**.
54. To check your data after the transformation (or before if you need to estimate a transformation) you can move to tab "Step 7: Check your data transformations". You also have a table of your reference data next to the nearest in time continuous data point at the bottom of the page **(i)** to check and estimate transformations.
55. If you need to undo a transformation then you should use the inverse of that transformation. If you don't remember the transformation you did, then you can find all transformations in a little table at the bottom of tab "Step 7".

56. In the next tab “Step 7: Check your data transformation” you have a zoomable plot that you can use to verify that your transformation was ok before moving on. At the bottom of the page you have a list of all the transformations you did to the data.



Step 5: Calibrate Chlorophyll a
Step 6: Transform/recalibrate specially marked areas
Step 7: Check your data transformations
Step 8: document final full dataset and transformations
Step 9: Reclassifying state of value codes
Step 10: Gap interpolation
Step 11: document final full dataset with final state of values
Step 12: Export
Work log
Help
Options

Check the data transformations you just did. This graph is zoomable. Go back to step 6 if you wish to adjust the transformations.

You cannot undo your work, but you can go back to step 6 and apply the inverse of the previously done transformation to undo it. You can find a list of all your transformations below.

Weert SF/Zeeschelde zes39c-SF-CM
State of value
auto good
marked group 2
missing
spike filter deleted
suspect
to be deleted
sensor cleaning
sensor replacment
Chfyla
0
50
100
150
200
apr-02-19 05:58
jul-02-19 11:57
okt-01-19 17:56
Time
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

Draw box on graph and double click to zoom in to drawn box. Double click on graph to zoom out to full extent

☐ Lock x-axis

☐ Lock y-axis

zoom out 2x

☑ add legend to graph

legend location: topleft

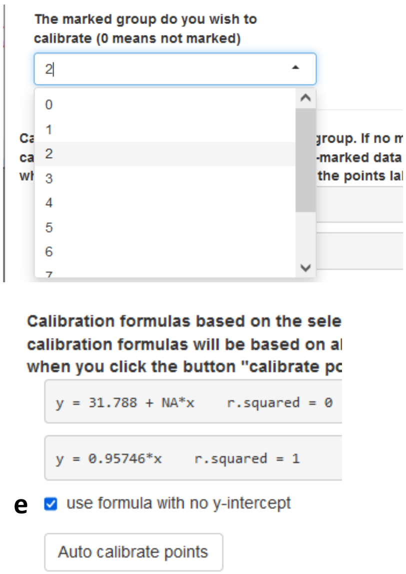
data

Calibrate all chlorophyll data in 'Marked Grouping' 0 with 0 + 1.6115 * x at 2021-11-06 17:56:34 . (Group 0 means all not inside a marked grouping. Suspect values were not used for calculating the calibration but they were calibrated.)

Example of recalibrating data:

I have the data in marked group 2 which I want to recalibrate to make it fit with the rest of the data. In the “Step 6: Transform/recalibrate specially marked areas” tab I will select group 2 **(b)** from the drop down menu.

I will first try to do an auto calibrate on the data. Looking at the regression formulas you can see that only the formula with no y intercept is valid because we only had one reference point during this period thus make sure check box **e** is checked. I will click “auto calibrate points”. And then move to tab ”Step 7” to check the transformation.



The marked group do you wish to calibrate (0 means not marked)

2

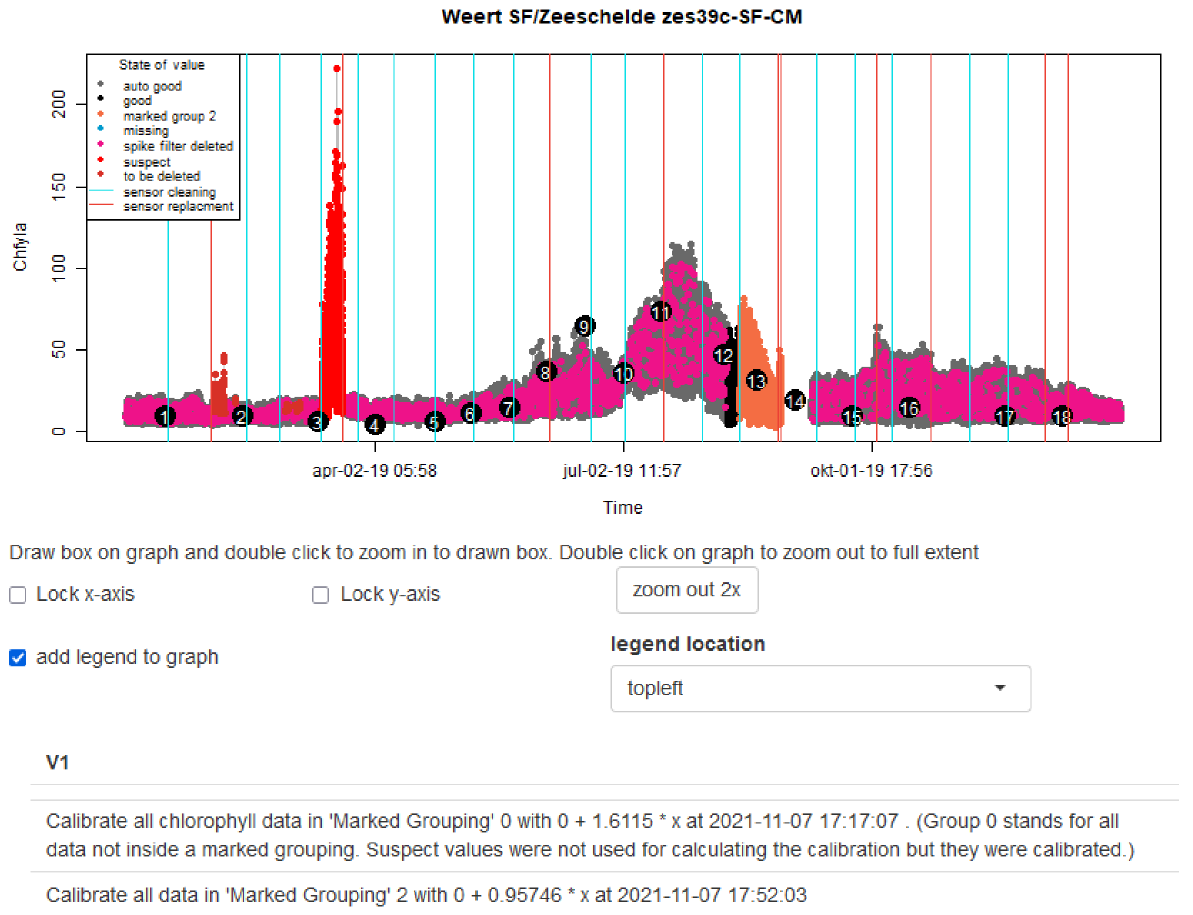
Calibration formulas based on the sele
calibration formulas will be based on al
when you click the button "calibrate pc

```
y = 31.788 + NA*x    r.squared = 0
```

```
y = 0.95746*x    r.squared = 1
```

e ☑ use formula with no y-intercept

Auto calibrate points



Draw box on graph and double click to zoom in to drawn box. Double click on graph to zoom out to full extent

☐ Lock x-axis ☐ Lock y-axis

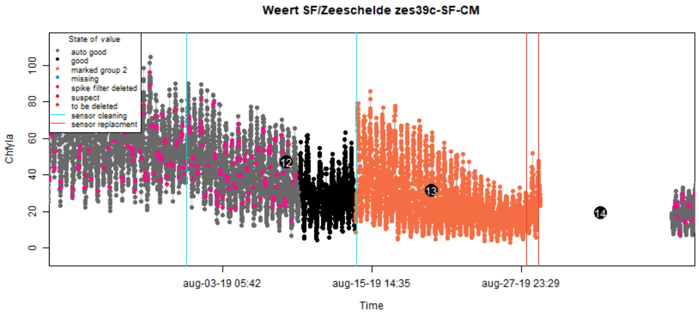
☑ add legend to graph

legend location: topleft

V1

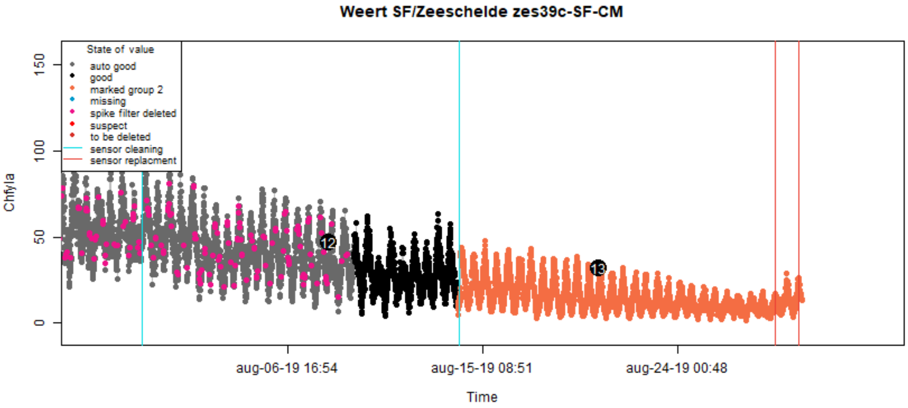
V1
Calibrate all chlorophyll data in 'Marked Grouping' 0 with 0 + 1.6115 * x at 2021-11-07 17:17:07 . (Group 0 stands for all data not inside a marked grouping. Suspect values were not used for calculating the calibration but they were calibrated.)
Calibrate all data in 'Marked Grouping' 2 with 0 + 0.95746 * x at 2021-11-07 17:52:03

On the tab ”Step 7: check your data transformations” you can see that the auto transformation wasn’t successful. We will go back to tab “Step 6” and in the manual calibrate points field **(g)** enter in inverse of the regression formula that was just applied to the data. You can find all the transformations applied to the data at the bottom of tab “step 7”. Thus on tab “step 6” we will enter into field **g** the equation x/0.95746 to go back to the original data and press the “manual calibrate points” button **(h)**.

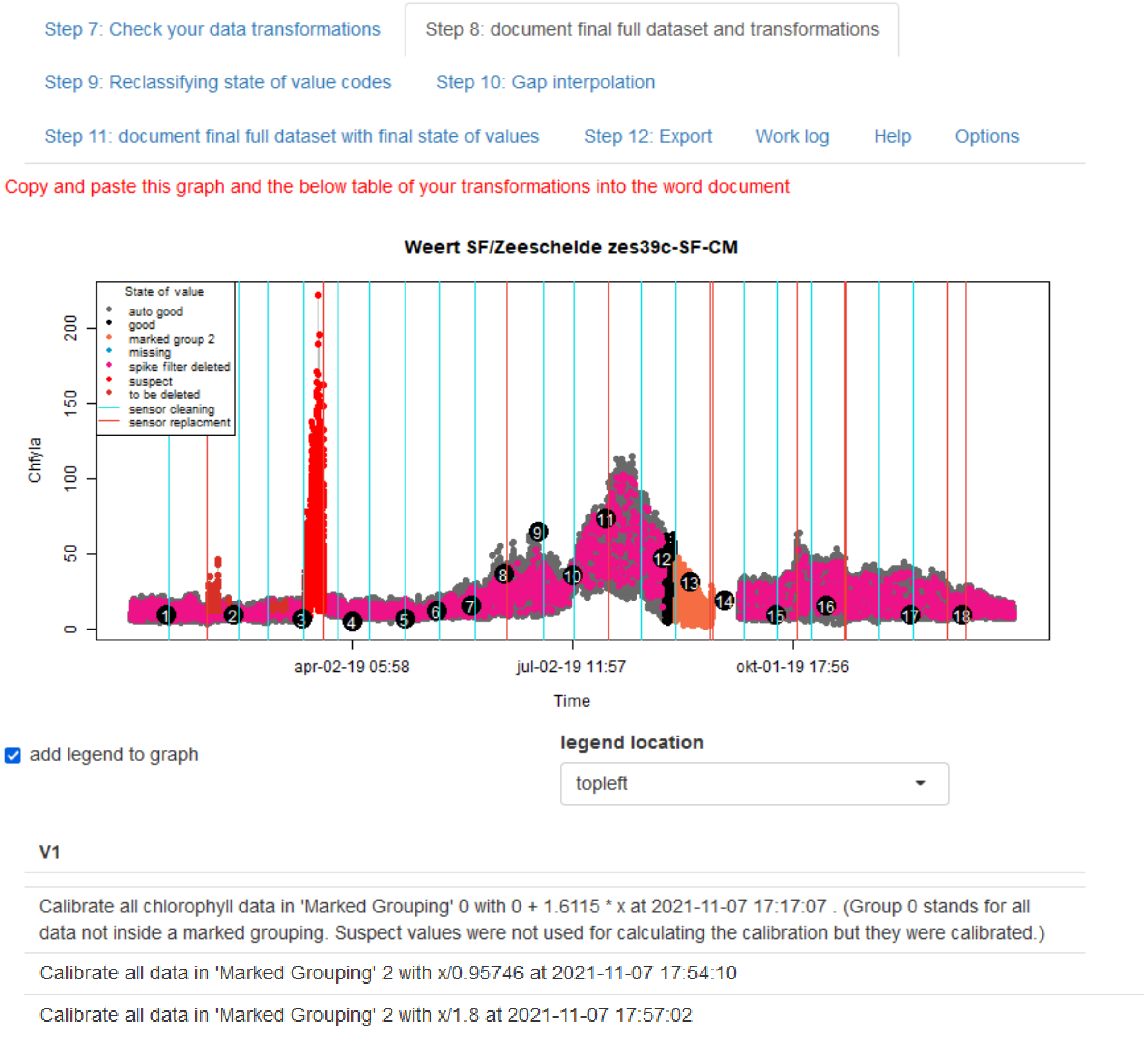


Now on the “step 7” tab again I can zoom in on the data in question and estimate what the transformation should be. I think it is x/1.8. Now I will go back to tab “step 6” and will enter into field **g** the equation x/1.8 and press the “manual calibrate points” button **(h)**.

Looking at the data on tab “step 7” the data transformation seems quite good. You will have to play around probably before you get a good estimate of what the data should look like. This points will get the state of value code “estimate”.



57. In the next tab “Step 8: document final full dataset and transformations”, copy the plot of the full dataset and the list of transformations at the bottom of the page into the word document for your records. Adjust the legend location to better see the data.


Step 7: Check your data transformations
Step 8: document final full dataset and transformations
Step 9: Reclassifying state of value codes
Step 10: Gap interpolation
Step 11: document final full dataset with final state of values
Step 12: Export
Work log
Help
Options
Weert SF/Zeeschelde zes39c-SF-CM
State of value
auto good
good
marked group 2
missing
spike filter deleted
suspect
to be deleted
sensor cleaning
sensor replacment
Chfyla
0
50
100
150
200
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
apr-02-19 05:58
jul-02-19 11:57
okt-01-19 17:56
Time

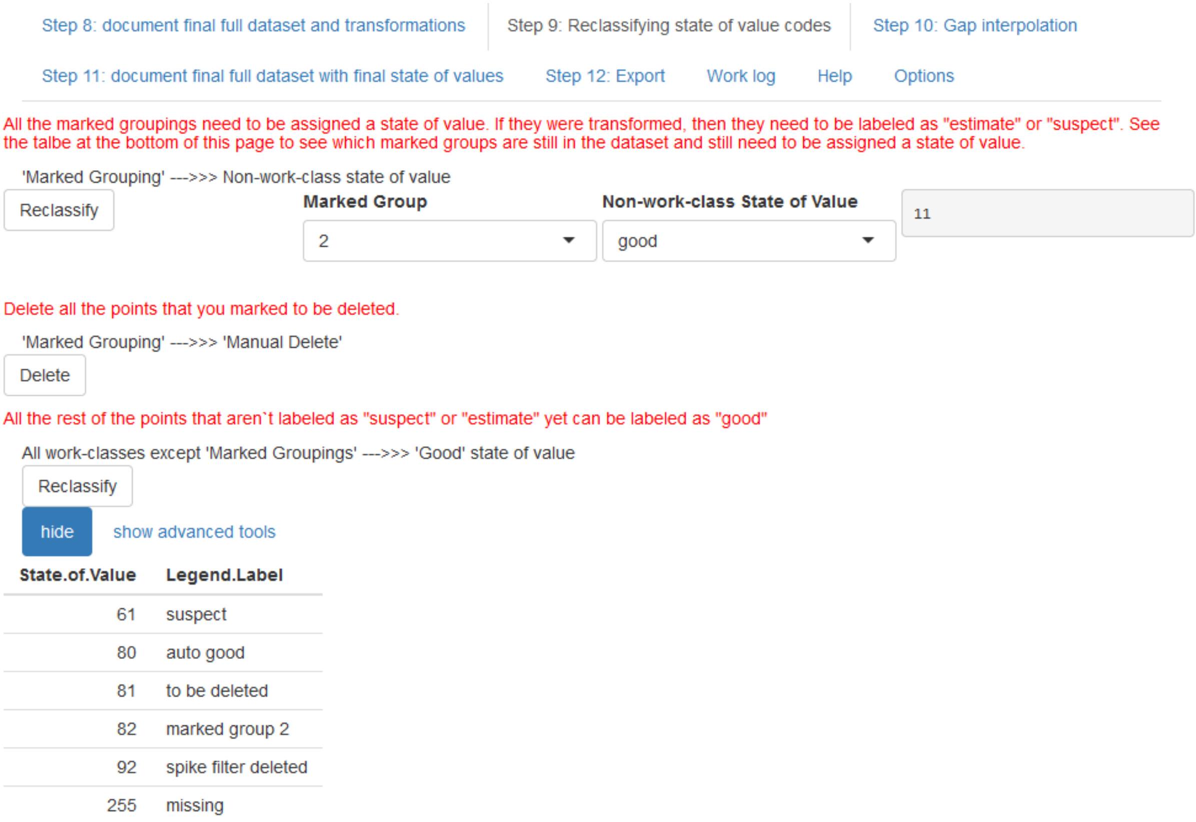
Copy and paste this graph and the below table of your transformations into the word document

☑ add legend to graph

legend location: topleft

V1
Calibrate all chlorophyll data in 'Marked Grouping' 0 with 0 + 1.6115 * x at 2021-11-07 17:17:07 . (Group 0 stands for all data not inside a marked grouping. Suspect values were not used for calculating the calibration but they were calibrated.)
Calibrate all data in 'Marked Grouping' 2 with x/0.95746 at 2021-11-07 17:54:10
Calibrate all data in 'Marked Grouping' 2 with x/1.8 at 2021-11-07 17:57:02

58. Go to the next tab “Step 9: Reclassifying state of value codes”. In this step you will first define which state of value your marked groupings will get with the first ”reclassify” button **(a)** and the drop down menus. If you transformed them then you need to mark them as “estimate” or “suspect”. You cannot mark a transformed datapoint as “good”.
59. Next you will delete the points that you marked to be deleted with the “Delete” button **(b).**
60. All remaining data points will be marked as “good” with the last “Reclassify” button **(c).**
61. At the bottom of the page you can see I little table **(d)** of the state of value codes that you have in your data right now. Check that there are no “marked group” state of values left in the list before you move on.



Step 8: document final full dataset and transformations
Step 9: Reclassifying state of value codes
Step 10: Gap interpolation
Step 11: document final full dataset with final state of values
Step 12: Export
Work log
Help
Options

All the marked groupings need to be assigned a state of value. If they were transformed, then they need to be labeled as "estimate" or "suspect". See the talbe at the bottom of this page to see which marked groups are still in the dataset and still need to be assigned a state of value.

a

'Marked Grouping' --->>> Non-work-class state of value

Reclassify

Marked Group: 2

Non-work-class State of Value: good

11

Delete all the points that you marked to be deleted.

b

'Marked Grouping' --->>> 'Manual Delete'

Delete

All the rest of the points that aren`t labeled as "suspect" or "estimate" yet can be labeled as "good"

c

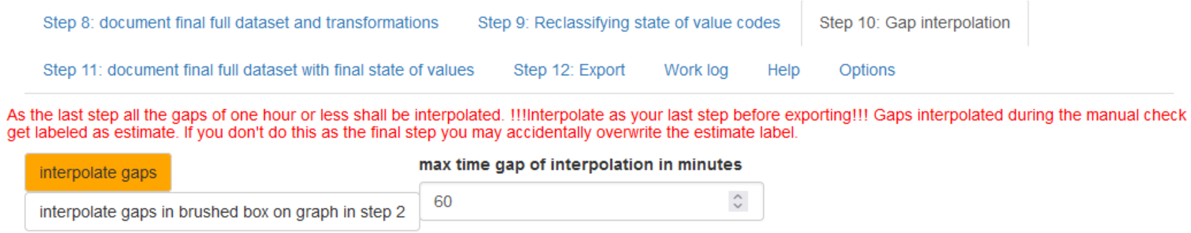
All work-classes except 'Marked Groupings' --->>> 'Good' state of value

Reclassify

d

State.of.Value	Legend.Label
61	suspect
80	auto good
81	to be deleted
82	marked group 2
92	spike filter deleted
255	missing

62. In the next step “Step 10: Gap interpolation” you will interpolate all the gaps in the data that are no more than 1 hour which you created during the manual cleaning. Just click the button “interpolate gaps”. These values will automatically get the state of value “estimate”. This needs to be the last transformation you do to the data to not accidentally overwrite these automatic state of value assignments.



Step 8: document final full dataset and transformations
Step 9: Reclassifying state of value codes
Step 10: Gap interpolation
Step 11: document final full dataset with final state of values
Step 12: Export
Work log
Help
Options

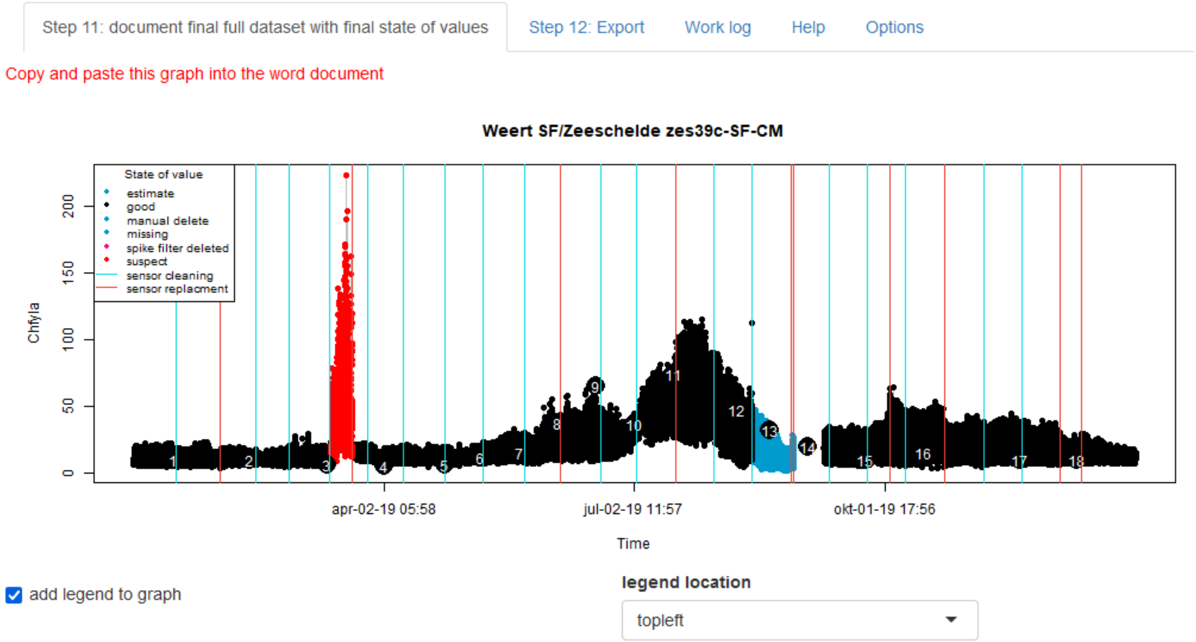
As the last step all the gaps of one hour or less shall be interpolated. !!!Interpolate as your last step before exporting!!! Gaps interpolated during the manual check get labeled as estimate. If you don't do this as the final step you may accidentally overwrite the estimate label.

interpolate gaps

interpolate gaps in brushed box on graph in step 2

max time gap of interpolation in minutes: 60

63. In the next tab “Step 11: Document final full dataset with final state of values” you will be presented with the final cleaned dataset. Copy this plot into your word document for your records. Adjust the legend location to better see the data.



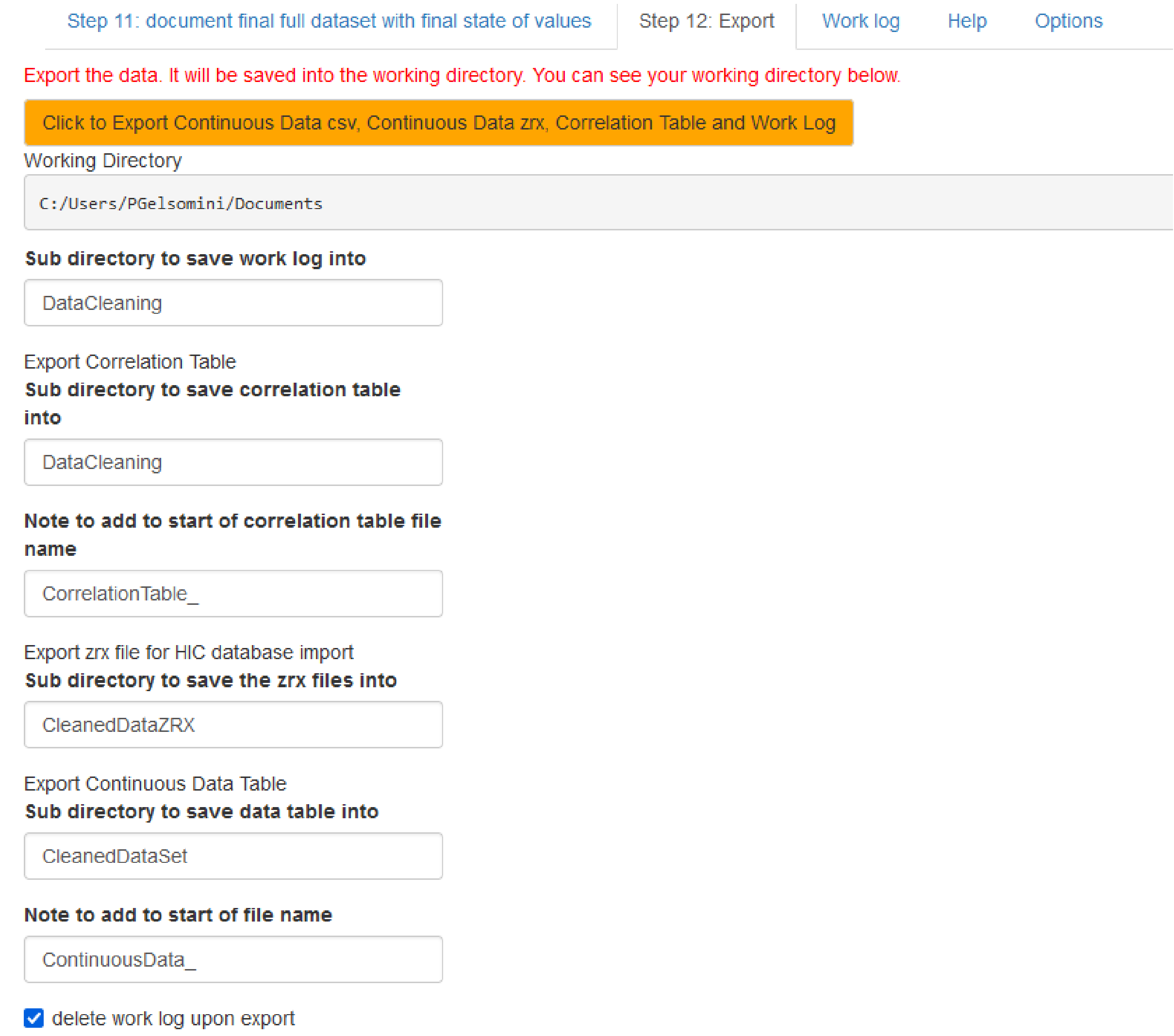
Copy and paste this graph into the word document

☑ add legend to graph

legend location

topleft

64. The last step “Step 11: Export”. When you click the orange button at the bottom of the page, the current version of the continuous data will be exported to a csv file, a zrx file will be exported for import back into the HIC database with the state of value codes all formatted to the requirements of the HIC, the correlation table from the “correlation and calibration” tab will be exported and the work log where you will find a detailed account of every change you made to the file for later reference and reproducibility will be exported as well. The final data will all be saved into your working directory in the subdirectory folders that you gave on this page. You can see you working directory at the top of this page.


Step 11: document final full dataset with final state of values
Step 12: Export
Work log
Help
Options
Click to Export Continuous Data csv, Continuous Data zrx, Correlation Table and Work Log

Export the data. It will be saved into the working directory. You can see your working directory below.

Working Directory

C:/Users/PGelsomini/Documents

Sub directory to save work log into

DataCleaning

Export Correlation Table

Sub directory to save correlation table into

DataCleaning

Note to add to start of correlation table file name

CorrelationTable_

Export zrx file for HIC database import

Sub directory to save the zrx files into

CleanedDataZRX

Export Continuous Data Table

Sub directory to save data table into

CleanedDataSet

Note to add to start of file name

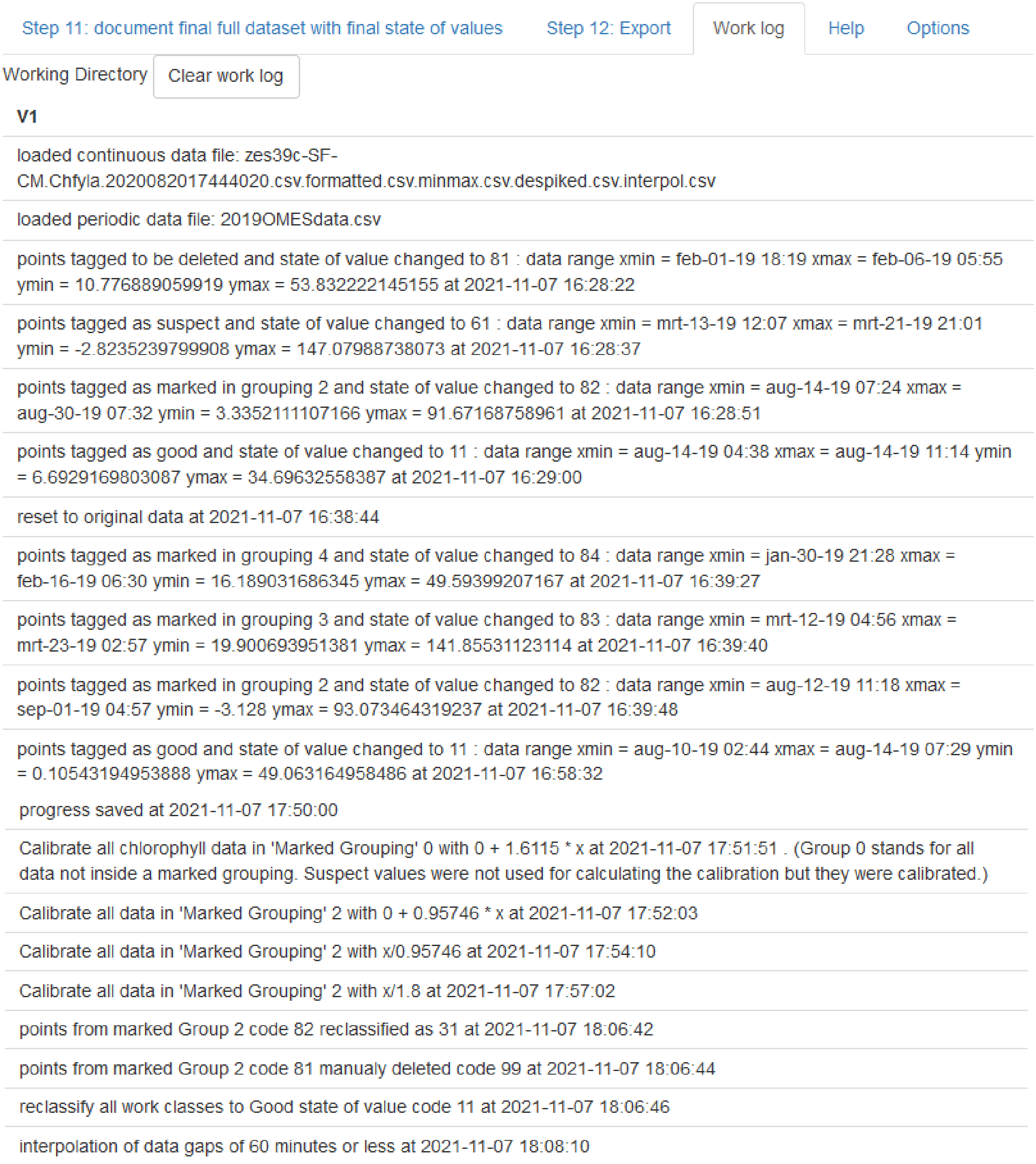
ContinuousData_

☑ delete work log upon export

NOTE:The station parameter codes that are needed for import back into the HIC database which are used in the zrx files are stored inside this package. Type ***zrxFileStationCodes*** *into the R console and press enter to see the list of codes. Please contact us if you need to make changes to this file.*

65. Now go back to the tab “Step 1: File upload” (step 38 in this document) and load the next dataset and repeat the process.

66. In the tab “Work log” you can see an account of everything you did to your data. This can be very useful especially when you need to know what calibration you used on your data. This work log will be deleted from the app once you export the data, but it will be saved as a text file.



V1
loaded continuous data file: zes39c-SF-CM.Chfyla.2020082017444020.csv.formatted.csv.minmax.csv.despiked.csv.interpol.csv
loaded periodic data file: 2019OMESdata.csv
points tagged to be deleted and state of value changed to 81 : data range xmin = feb-01-19 18:19 xmax = feb-06-19 05:55 ymin = 10.776889059919 ymax = 53.832222145155 at 2021-11-07 16:28:22
points tagged as suspect and state of value changed to 61 : data range xmin = mrt-13-19 12:07 xmax = mrt-21-19 21:01 ymin = -2.8235239799908 ymax = 147.07988738073 at 2021-11-07 16:28:37
points tagged as marked in grouping 2 and state of value changed to 82 : data range xmin = aug-14-19 07:24 xmax = aug-30-19 07:32 ymin = 3.3352111107166 ymax = 91.67168758961 at 2021-11-07 16:28:51
points tagged as good and state of value changed to 11 : data range xmin = aug-14-19 04:38 xmax = aug-14-19 11:14 ymin = 6.6929169803087 ymax = 34.69632558387 at 2021-11-07 16:29:00
reset to original data at 2021-11-07 16:38:44
points tagged as marked in grouping 4 and state of value changed to 84 : data range xmin = jan-30-19 21:28 xmax = feb-16-19 06:30 ymin = 16.189031686345 ymax = 49.59399207167 at 2021-11-07 16:39:27
points tagged as marked in grouping 3 and state of value changed to 83 : data range xmin = mrt-12-19 04:56 xmax = mrt-23-19 02:57 ymin = 19.900693951381 ymax = 141.85531123114 at 2021-11-07 16:39:40
points tagged as marked in grouping 2 and state of value changed to 82 : data range xmin = aug-12-19 11:18 xmax = sep-01-19 04:57 ymin = -3.128 ymax = 93.073464319237 at 2021-11-07 16:39:48
points tagged as good and state of value changed to 11 : data range xmin = aug-10-19 02:44 xmax = aug-14-19 07:29 ymin = 0.10543194953888 ymax = 49.063164958486 at 2021-11-07 16:58:32
progress saved at 2021-11-07 17:50:00
Calibrate all chlorophyll data in 'Marked Grouping' 0 with 0 + 1.6115 * x at 2021-11-07 17:51:51 . (Group 0 stands for all data not inside a marked grouping. Suspect values were not used for calculating the calibration but they were calibrated.)
Calibrate all data in 'Marked Grouping' 2 with 0 + 0.95746 * x at 2021-11-07 17:52:03
Calibrate all data in 'Marked Grouping' 2 with x/0.95746 at 2021-11-07 17:54:10
Calibrate all data in 'Marked Grouping' 2 with x/1.8 at 2021-11-07 17:57:02
points from marked Group 2 code 82 reclassified as 31 at 2021-11-07 18:06:42
points from marked Group 2 code 81 manualy deleted code 99 at 2021-11-07 18:06:44
reclassify all work classes to Good state of value code 11 at 2021-11-07 18:06:46
interpolation of data gaps of 60 minutes or less at 2021-11-07 18:08:10

67. When you are done, you must close the app window before continuing work in R or opening another R Shiny app.

R package help files

Function for downloading all continuous biological data from the HIC server from a given year

Description

Downloads all the HIC data for Chfyla, DO, pH and PPFD for a given year. Additional time series groups can also be added to the download list. The package HICwebservices is required.

Usage

```
HICwebservicesBioDownload(
  year,
  credentials = NULL,
  other.group.ids = NULL,
  output.dir = NULL
)
```

Arguments

`year`
: The year that you wish to download

`credentials`
: This is the credentials input variable from the HICwebservices package function get_token(). See documentation from HICwebservices package.

`other.group.ids`
: Any other timeseries group IDs that you wish to also download. Take note that the entire group will be downloaded.

`output.dir`
: The folder name where you wish all the downloaded files to be saved to. If left as NULL it will by default save to 'DownloadedHICDataYYYYMMDDHHMMSS'

Value

A folder with all the downloaded time series as csv files and two additional csv files (DownloadSummaryBinary.csv and DownloadSummaryID.csv) that give the summary of which files were downloaded and site coordinates in Belgian Lambert 72. The metadata columns DateTimeUnix, Parameter.Name, Station.Name, Station.Number, Parameter.Unit are added to the downloaded table.

HIC.maint {HICbioclean} R Documentation

Batch process folder for converting maintenance Excel files to csv files

Description

This function was specifically made to convert the excel sensor maintenance files into csv files that can be easily read into R.

Usage

```
HIC.maint(input.directory, output.directory)
```

Arguments

`input.directory`
folder directory of Excel files. No back slashes(\), only use forward slashes(/) or double back slashes(\\).

`output.directory`
folder directory where you would like to save the csv files. No back slashes(\), only use forward slashes(/) or double back slashes(\\).

Details

It batch processes all the files inside a given folder.

If you don't provide a full path to the output directory then it will be placed in your working directory.

Value

A folder of csv files with the same name as the input excel files.

Examples

```
HIC.maint("C:/Rdata/MaintenanceFiles", "MaintenanceFilesCSV")
```

dspk.DespikingWorkflow.CSVfileBatchProcess {HICbioclean} R Documentation

Batch process: Despike and autovalidate continuous data

Description

A full work flow for auto validation of continuous data. It may be run as a batch process or on individual R objects. It performs min max filtering, despiking and linear gap interpolation. You may run all these steps or a select few. Files are all outputted as csv files to your working directory at each step.

Usage

```
dspk.DespikingWorkflow.CSVfileBatchProcess(
  steps = c(1, 2, 3),
  input.directory = NULL,
  sep = ",",
  dec = ".",
  header = T,
  Data = NULL,
  Value,
  val.NAvalue = NULL,
  unchecked.state.of.value.code = 110,
  NA.state.of.value.code = 255,
  add.original.data = T,
  DateTime = NULL,
  datetime.format = NULL,
  datetime.timezone = "GMT",
  ConditionalMinMaxColumn = NULL,
  ConditionalMinMaxValues = NULL,
  ConditionalMin = NULL,
  ConditionalMax = NULL,
  Min = (-Inf),
  Max = Inf,
  minmax.state.of.value.code = 91,
  sampling.interval = NULL,
  despiked.state.of.value.code = 92,
  good.state.of.value.code = 80,
  despike.threshold = 3,
  despike.Method = "median",
  precision = NULL,
  max.gap = Inf
)
```

Arguments

`steps`

Numeric vector containing the values 1, 2 and/or 3 corresponding to step 1 min max filter, step 2 despiking and step 3 gap interpolation. For example ‘steps=c(1,3)’ will run a min max filter and then interpolate the gaps. Will run all steps by default.

`input.directory`

Character string of the path to the folder containing all the csv files that you wish to batch process. This argument may be omitted if you are entering vectors directly into the ‘Value’ and ‘DateTime’ arguments.

`sep`

Arguments indicating the formatting of the input csv files. It is the field separator character. Values are separated by this character. By default it is comma ",".

`dec`

Arguments indicating the formatting of the input csv files. It the character used for decimal points. By default if is a period ".".

`header`

Arguments indicating the formatting of the input csv files. It is a logical value indicating if the first line is the column titles. By default it is TRUE.

`Data`

A dataframe object. If you only wish to process one data frame, then it can be entered directly from the R environment with this argument. If you enter in an input.directory then 'Data = ' will be ignored and the files from the input directory will be processed.

`Value`

If the data is from a csv file or a dataframe, it is a quoted character string indicating the column name or an integer indicating the column number of the column containing the data values that you wish to despike. Data may also be entered as a single vector object (unquoted) such as ‘Value = mydata$values’ or ‘Value = values’

`val.NAvalue`

The value indicating an NA value in your input data. If this value is NA, then this argument can be omitted.

`unchecked.state.of.value.code`

Number indicating that a given value is unchecked. By default 110.

`NA.state.of.value.code`

State of value code given to missing data.

`add.original.data`

A logical value indicating if the original input data should be included in the output tables. Note that if you input a csv file, then every column in that file will be kept. TRUE by default.

`DateTime`

If the data is from a csv file, it is a quoted character string indicating the column name or an integer indicating the column number of the column containing the datetime values of the samples. Data may also be entered as a single vector object (unquoted) such as ‘DateTime = mydata$time’ or ‘DateTime = time’

`datetime.format`

Character string giving the datetime format. See the strptime() help file for additional help.

`datetime.timezone`

Character string giving the time zone of the datetime. By default “GMT”. Use OlsonNames() for a list of all time zones.

`ConditionalMinMaxColumn`

The column name in quotes or column number or vector object that contains the factor variable to base your conditional min max filter on

`ConditionalMinMaxValues`

A vector containing the factor values to base the conditional min max filter on

`ConditionalMin`

A vector containing the condition minimums that correspond to the respective values in ConditionalMinMaxValues

`ConditionalMax`

A vector containing the condition maximums that correspond to the respective values in ConditionalMinMaxValues

`Min`

Number giving the minimum reasonable value. All values below this will be deleted.

`Max`

Number giving the maximum reasonable value. All values above this will be deleted.

`minmax.state.of.value.code`

Number indicating that the value has been deleted during the min max filter. By default 91.

`sampling.interval`

As numeric, the time between samples. If you enter NULL then it will calculate it for you. By default NULL.

`despiked.state.of.value.code`

Number indicating that a given value was deleted during the despiking. By default 92.

`good.state.of.value.code`

Number indicating that a given value has been check and deemed not a spike during the despiking. By default 80.

`despike.threshold`

Number indicating the threshold for defining a spike. By default it is 3, which corresponds to 3 median absolute deviations or 3 standard deviations.

`despike.Method`

Character string "median" or “mean” indicating the method to use for the despiking. By default “median”.

`precision`

A number indicating the precision of the input values. Interpolated values will be rounded to this precision. If left as NULL then the numbers will be rounded to the largest decimal length found in the data.

`max.gap`

As numeric, the time span of the maximum data gap you wish to interpolate

Details

Each csv file will be process separately and saved into new csv files at each step in the despiking process (pre-process: formatting, step 1: min/max filter, step 2: despiking, step 3: gap interpolation). The final data will be found in the folder “autodespikeYYYYMMDDHHMMSS” within the subfolder “step3Interpol.FinalData”. Use the function getwd() to find your working directory. State of value codes are added to the data to keep track of how each value was handled during the auto-validation process (110 unchecked, 255 missing, 80 auto good value, 91 deleted during min/max filter, 92 deleted during despiking). The original data will still be in the newly generated csv files, with the processed data saved in new columns.

Please see the FunctionLogFile.txt that was generated to see any error messages and details about the selected preferences and calculated preferences.

Quick Start

Place all your data you wish to auto-despike into one folder as csv files. The data values must be numeric. Date and time should be both in the same column with no time zone corrections (e.g. 13:20 +2 The +2 is a time zone correction). Datetime may also be numeric. If there are no interruptions in the sampling causing data gaps, then a datetime is not needed.

The default state of values codes

110 Unchecked
255 Missing value
80 Auto good
91 Auto delete min max filter
92 Auto delete Spike

Algorithm overview and workflow

Pre-processing: Formatting and compatibility check: If an ‘input.directory’ containing multiple csv files was provided, then each file will be processed and saved separately. The ‘Value’ data is checked that it is numeric and are then saved into a new column "dspk.Values" to not overwrite old data. NA codes ‘val.NAvalue’ will be replaced with the value NA. The ‘DateTime’ data will be checked if it is numeric and saved into a new column "dspk.DateTimeNum" to not overwrite old data. If it is a character string, then it will be converted to numeric using the provided ‘datetime.format’ and ‘datetime.timezone’. If no datetime is provided then the samples will be numbered consecutively and saved as the datetime. A new column "dspk.StateOfValue" will be generated with all values equal to the ‘unchecked.state.of.value.code’ (default 110). NA values will be given the 'NA.state.of.value.code' (default 255). If the entered csv data table already has a "dspk.StateOfValue" column, then the original state of values from that column will be used. If ‘add.original.data’ is equal to TRUE (this is the default) then the original data will be included in the formatted data table. The formatted data table will be saved to a new csv file in the directory “autodespikeYYYYMMDDHHMMSS/preprocFormat” within your working directory.

Step 1: Min/Max filter: Each csv file generated from the previous step will be processed and saved separately. All data points that are above the entered ‘Max’ or below entered ‘Min’ will be deleted. The "dspk.StateOfValue" of the deleted values will be set to ‘minmax.state.of.value.code’ (default 91). The data will be saved in a new csv file in the directory “autodespikeYYYYMMDDHHMMSS/step1MinMax” within your working directory.

Step 2: Despiking: Each csv file generated from the previous step will be processed and saved separately. With the default “despike.Method” median and the default “despike.threshold” 3: all data points that are more than 3 median absolute deviations away from the median of the 10 surrounding data points (5 before and 5 after) will be deleted. At least 5 surrounding data points is required for the sample to be evaluated. The algorithm will not look farther than 5 sampling intervals before and after the data point, for handling data gaps. If a “sampling.interval” is not provided then it will be calculated as the mode of the interval between samples. The "dspk.StateOfValue" of the deleted values will be set to “despiked.state.of.value.code” (default 92). The "dspk.StateOfValue" of the values that passed the despike test will be set to “good.state.of.value.code” (default 80). The data will be saved in a new csv file in the directory “autodespikeYYYYMMDDHHMMSS/step2Despike” within your working directory.'

Step 3: Data gap interpolation: Each csv file generated from the previous step will be processed and saved separately. All data gaps will be linear interpolated unless a ‘max.gap’ length for interpolation is given. If a ‘precision’ is given, then the interpolated values will be rounded to that precision. The state of values will not be changed to know what the original state of the value was. If the value has a state of value deleted or missing but there is a value then it can be assumed that it was interpolated. The data will be saved in a new csv file in the directory “autodespikeYYYYMMDDHHMMSS/step3Interpol.FinalData” within your working directory.

Details

Make sure that when you enter the path name for the directory it has forward-slashes(/) or double-back-slashes(\\) and not back-slashes(\). If you copy the directory path from windows, it will have back-slashes(\) and these need to be changed to forward-slashes(/) or double-back-slashes(\\).

Interpolated the values will be rounded to the same decimal places as the original data. If you wish you may enter a custom precision. Examples: 0.566 would be ‘precision = 0.001’ 1200 would be ‘precision = 100’ Measurements with steps of 5 would be ‘precision = 5’ Measurements to the nearest half unit would be ‘precision = 0.5’

You are not required to supply ‘DateTime’ for this function. This is only necessary for handling data gaps while despiking and interpolating the data. If you omit this argument, then it will generate a datetime column which contains the samples numbered consecutively so you can still indicate with ‘max.gap’ the maximum data gap you wish to interpolate by filling in the number of missing samples into ‘max.gap’. The ‘DateTime’ data can be numeric values or as a datetime character strings (e.g. “2018-04-23 15:32:18”). If the datetime data are character strings, then a ‘datetime.format’ must be provided (e.g. datetime.format = " function documentation for help on syntax. If you enter a date time as a character string, then it will be converted into UNIX seconds. The default time zone is GMT but it can be changed to GMT+1 with 'datetime.timezone = "Etc/GMT-1"' Use the OlsonNames() function for a list of all time zones. If you enter a datetime, then the date and time must be in the same cell, as in they cannot be in separate columns. Often a time conversion to GMT is supplied with the time (e.g. 16:32 +2). This function uses the as.POSIXct function and cannot handle these conversions (like the "+2" in the example). They need to be removed and dealt with prior to analysis.

When the data is being formatted the columns "dspk.Values", "dspk.DateTimeNum" and "dspk.StateOfValue" will be generated. If these columns already exist in the data table, then they will be overwritten. This may or may not be desired. If the column "dspk.StateOfValue" is in the original data, then that data will be used for the 'state of values', otherwise the 'unchecked.state.of.value.code' (default 110) will be used for all data signifying 'unchecked' status.

The spike removal algorithm is by default (despike.Method = "median") all sample points that are more than the threshold 3 median absolute deviations (the scale factor 1.4826 is used assuming normal distribution) from the median of the 10 surrounding data points (5 before and 5 after) are automatically deleted. The threshold of 3 can be changed with "despike.threshold =" You can set despike.Method = "mean" to use standard deviations and mean for the algorithm instead of the default median absolute deviations and median. However median is a much more robust statistic for handling outliers.

If you have data gaps and you have the sampling times and you don”t want the despiking algorithm to look past the data gaps, then make sure to to supply “DateTime”. The time interval between samples "sampling.interval = " will be calculated as the mode of the difference between consecutive samples. If there are irregularities in the sampling interval that will prevent this calculation then you can enter in the sampling interval in "sampling.interval = " in the numeric-datetime unit. If you entered in a character datatime column then it is POSIX time converted to numeric so the unit is in UNIX seconds.

The default is that it will do linear interpolation of all data gaps. You can restrict the size of the data gap with "max.gap.interpolate". This needs to be in the unit of your numeric datetime. If you entered in a datetime column containing character strings then it is POSIX time converted to numeric so the unit is in UNIX seconds. If you did not entered a time column, then the unit is in samples, for example "max.gap.interpolate = 5" will only interpolate gaps of up to 5 samples long.

Value

Your final data can be found in “autodespikeYYYYMMDDHHMMSS/step3Interpol.FinalData” within your working directory as comma separated csv files. The cleaned values will be in column "dspk.Values", the state of values in column "dspk.StateOfValue", and the numeric datetime will be in column "dspk.DateTimeNum".

Examples

```
#HIC data cleaning and validation protocol
#Batch process HIC database files that were formatted
#with the HIC.Continuous.Data.Import.Format() function.
#With default despiking algorithm (threshold of 3 MAD from the median).
dspk.DespikingWorkflow.CSVfileBatchProcess(
   input.directory = 'data/HIC.data', #Load csv files from the folder 'HIC.data'
   sep = ',', dec = '.', header = T,  #The csv files are separated by commas with
   #point decimals and the first line is the column names.
   Value = 3, val.NAvalue = -777, #The values are found in the third column.
   #-777 stands for no-value
   DateTime = "DateTimeUnix",
   #The numeric datetime column name.
   #Because it is numeric no datetime formatting info is needed
   ConditionalMinMaxColumn = 'Parameter.Name',
   ConditionalMinMaxValues = c('DO','pH','chfyla','PPFD1','PPFD'),
   ConditionalMin = c(0,0,0,0,0), ConditionalMax = c(30,15,1000,2000,2000)
   #conditional min max filter based on parameter with minimum reasonable value
   #for oxygen, pH, chlorophyll a, and PPFD being 0 and the maximum
   #reasonable value for oxygen and pH being 15 and chlorophyll 1000 and PPFD being 2000
   max.gap = 3600)
   #The maximum data gap that should be interpolated is one hour or 3600 seconds.

#Example: Running full despiking work flow with batch process from a folder of
#csv files, with default despiking algorithm (threshold of 3 MAD from the median)
```

```
dspk.DespikingWorkflow.CSVfileBatchProcess(
   input.directory = 'data/PPFD.data', #Load csv files from the folder 'PPFD.data'
   sep = ',', dec = '.', header = T,  #The csv files are separated by commas with
   #point decimals and the first line is the column names.
   Value = 3, val.NAvalue = -777, #The values are found in the third column.
   #-777 stands for no-value
   DateTime = "DateTime",
   #The datetime column name is "DateTime".
   datetime.format = "%Y-%m-%d %H:%M:%S",
   #The datetimes are character strings with this format "2018-04-23 15:32:18".
   datetime.timezone = 'Etc/GMT-1', #The time zone is UTC+1.
   Min=(-50), Max=1700, #The minimum reasonable value is -50 and the maximum  is 1700
   sampling.interval = NULL,
   #Sampling interval is regular and it will be calculated from the provided data
   precision = NULL, #The data precision will be calculated from the data.
   max.gap = 3600)
   #The maximum data gap that should be interpolated is on hour or 3600 seconds.

#Example: Max filter on a vector with interpolation of resulting data gaps of up to 2 records
example.data = c(2,2,4,16,-4,2,0,96,8,12,26,66,2)
dspk.DespikingWorkflow.CSVfileBatchProcess(
   steps = c(1,3), #run step 1 min/max filter and step 3 data gap interpolation
   Value = example.data, #The values are found in the vector 'example.data'
   Max=10, #Filter out all values above 10
   sampling.interval = 60,
   #This is extraneous information since no time data was given, it will be ignored
   precision = 2, #The data are all even so precision is set to 2.
   max.gap = 2)  #The maximum data gap that should be interpolated is 2 missing values.
>Output CSV file:
>"Value","dspk.Values","dspk.DateTimeNum","dspk.StateOfValue"
>  2,         2,                1,                110
>  2,         2,                2,                110
>  4,         4,                3,                110
>  16,        0,                4,                91
>  -4,       -4,                5,                110
>  2,         2,                6,                110
>  0,         0,                7,                110
>  96,        4,                8,                91
>  8,         8,                9,                110
>  12,        NA,               10,               91
>  26,        NA,               11,               91
>  66,        NA,               12,               91
>  2,         2,                13,               110

#Example: Full despiking work flow on an R dataframe with no data gaps.
#999 is the NA value. Despiking is done with a threshold of 4 using the
#standard deviations from the mean.
dspk.DespikingWorkflow.CSVfileBatchProcess(
     Value = datatable$values, val.NAvalue = 999,
     #The values are found in datatable$values. 999 stands for no-value
     Min=(0), Max=200, #The minimum resonable value is 0 and the maximum  is 200
     despike.threshold = 4, despike.Method = "mean",
     #The despiking algorithm is all data points more than 4 standard deviations
     #from the mean of the surrounding 10 data points
     precision = 0.01, #The data has a precision to the hundredth decimal place.
     max.gap = 10) #the maximum data gap that should be interpolated is 10 samples.

#Example: The same example as above but now entering in the dataframe in
#the Data =' argument.
dspk.DespikingWorkflow.CSVfileBatchProcess(
     Data = datatable, Value = 'values', val.NAvalue = 999,
     #The values are found in datatable$values. 999 stands for no-value
     Min=(0), Max=200, #The minimum reasonable value is 0 and the maximum is 200
     despike.threshold = 4, despike.Method = "mean",
     #The despiking algorithm is all data points more than 4 standard deviations
     #from the mean of the surrounding 10 data points
     max.gap = 10)  #the maximum data gap that should be interpolated is 10 samples.

#Example: Full despiking work flow with a conditional min max filter. No Min or Max
#was given so if the conditions are not met, then the min and max will be set to
#infinity.
dspk.DespikingWorkflow.CSVfileBatchProcess(
     Data = datatable, Value = "values", #The values are found in datatable$values.
     ConditionalMinMaxColumn = "Parameter.Name",
     #The factors for basing the conditional min max are in column "Parameter.Name"
     ConditionalMinMaxValues = c('PercentO2','Temp'),
     ConditionalMin = c(0,-30), ConditionalMax = c(100,150)
     #conditional min max filter based on parameter with minimum reasonable value
     #for percent oxygen being 0% and for temperature being -30C and maximum
     #reasonable value for percent oxygen being 100% and for temperature being 150C
     max.gap = 10) #The maximum data gap that should be interpolated is 10 samples.

#Example: Full despiking work flow with a conditional min max filter. A Min and a
```

```
#Max was now given so if the conditions are not met, then the min and max will be
#set to 10 and 100. If you give a Min and a Max in addition to the conditional
#values, then if the conditions are not met the min and the max will be set to
#those values given.
dspk.DespikingWorkflow.CSVfileBatchProcess(
     Data = datatable, Value = "values", #The values are found in datatable$values.
     Min = 10, Max = 100, #if the bellow conditional min max values are not met,
     #then it will take these values as the min and max
     ConditionalMinMaxColumn = "Parameter.Name",
     #The factors for basing the conditional min max are in column "Parameter.Name"
     ConditionalMinMaxValues = c('PercentO2','Temp'),
     ConditionalMin = c(0,-30),
     ConditionalMax = c(100,150)
     #conditional min max filter based on parameter with minimum reasonable value
     #for percent oxygen being 0% and for temperature being -30C and maximum
     #reasonable value for percent oxygen being 100% and for temperature being 150C
     max.gap = 10) #The maximum data gap that should be interpolated is 10 samples.
```

Graphical applications for formatting, auto-validation, manual-validation and final export of HIC database data

Description

Three graphical user interfaces (GUI) for the processing of the Flemish Hydrological Information Center (HIC) data. These graphical apps will walk you through the entire work flow of data import, validation/calibration and data export, in an intuitive visual manner without the need for coding. This is the only method of manual validation and calibration and final export.

Usage

```
HIC.App.manual.StepByStep()
```

Details

The formatting step and the auto validation step can be done in R code using the functions HIC.Continuous.Data.Import.Format(), spk.DespikingWorkflow.CSVfileBatchProcess(), and HIC.PPFDAutoValidation.CSVfileBatchProcess(). The manual validation, calibration and final export must be done using this graphical app.

Recomended workflow

First: format the csv files that are exported from the HIC database - HIC.App.format()

Second: auto-validate the formatted files - HIC.App.auto()

Third: manual-validate/calibrate the files and export the data - HIC.App.manual.StepByStep()

Files can be batch processed at each step so there is no need to run through all the steps for each individual file.

Manual Tutorial

https://github.com/pgelsomini/HICbioclean/blob/master/MANUAL-TUTORIAL-HICbioclean.-Rpackage.pdf

Value

Three shiny apps

Examples

```
HIC.App.format() # to format the HIC database output csv files
HIC.App.auto() # to auto-validate the formatted data
HIC.App.manual.StepByStep() # to manually check the auto-validated data
```

[Package *HICbioclean* version 0.1.1]