

Filling data gaps using the GapFiller tool

Tool version 1.0

November 2016

Authors

Froukje de Boer

Peter Droogers

Wilco Terink

Report FutureWater: 158



FutureWater

Costerweg 1V
6702 AA Wageningen
The Netherlands

+31 (0)317 460050

w.terink@futurewater.nl

www.futurewater.nl

1 Purpose

Data series of environmental variables (e.g. precipitation, discharge, temperature) often contain gaps. Furthermore start and end dates of environmental data series are often inconsistent. The best way to get an overview of missing and available data, and to fill the data gaps, is to make graphs of the data series of different stations and compare them with each other. Since this process of data analysis and preparation of a complete dataset can take a lot of time, FutureWater has developed a GapFiller tool that partly automatizes and streamlines this process. The GapFiller tool works with data organized in Excel files (with the extension *.xls* or *.xlsx*). In this manual the use of the GapFiller tool is explained and demonstrated using an example dataset of discharge containing missing values.

2 Downloading the GapFiller tool

The GapFiller tool can be downloaded from:

ftp://95.97.194.183:22/PUBLIC_SHARED_DATA/Gapfiller

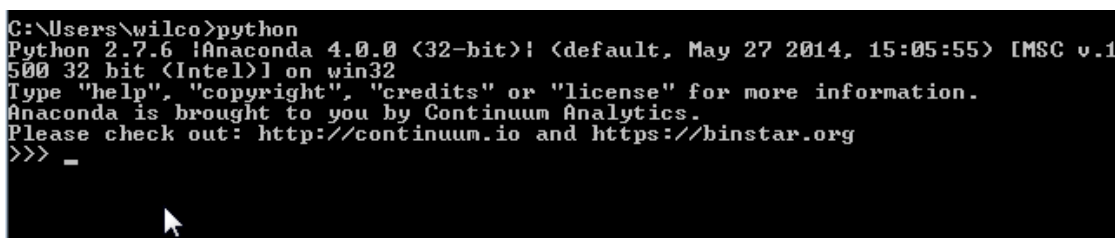
Username: guest01

Password: FW.guest012_

3 Installing Python

The GapFiller tool is a python script that must be run from the command window. Python 2.7.6 and a couple of python packages (Pandas and NumPy) should be installed on your computer for the tool to work correctly. Check if these are installed on your computer. If not, then the following steps should be taken before the GapFiller tool can be used:

- 1: Uninstall existing python packages
- 2: Download Anaconda from: <https://repo.continuum.io/archive/Anaconda2-4.2.0-Windows-x86.exe>
- 3: Run the downloaded Anaconda executable
- 4: After installation, open the command window and type *conda install python=2.7.6*
- 5: Check if python is installed correctly by typing python. You then should see a screen similar to the figure below



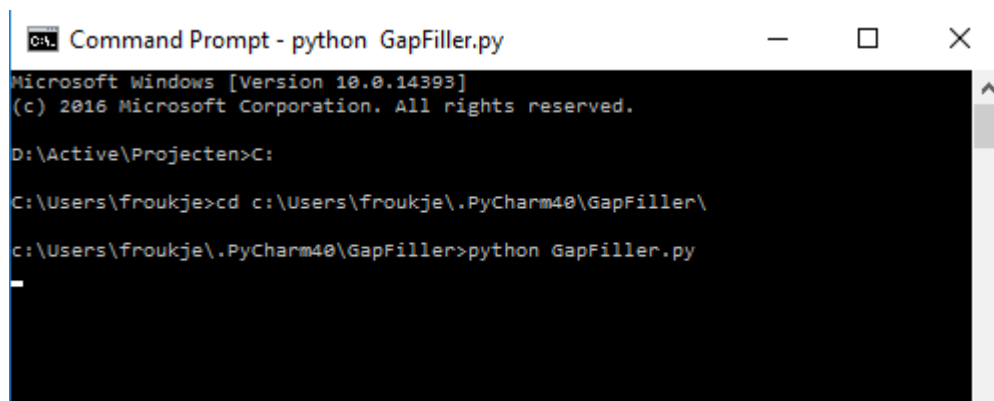
```
C:\Users\wilco>python
Python 2.7.6 |Anaconda 4.0.0 (32-bit)| (default, May 27 2014, 15:05:55) [MSC v.1
500 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io and https://binstar.org
>>> _
```

Figure 1: Successful installation of Anaconda Python.

4 Run GapFiller.py

- 1: Open the Command Prompt (Figure 2).
- 2: Change the directory to where the script GapFiller.py is located by typing the drive, e.g. C: and then the command *cd* (change directory) followed by the directory, e.g. *c:\Users\froukje\PyCharm40\GapFiller*
- 3: Type *python GapFiller.py*

After the third step the GapFiller interface will be opened; this can take a while (see Figure 3):



```
Command Prompt - python GapFiller.py
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

D:\Active\Projecten>C:

C:\Users\froukje>cd c:\Users\froukje\.PyCharm40\GapFiller\

c:\Users\froukje\.PyCharm40\GapFiller>python GapFiller.py
```

Figure 2. Command Prompt.

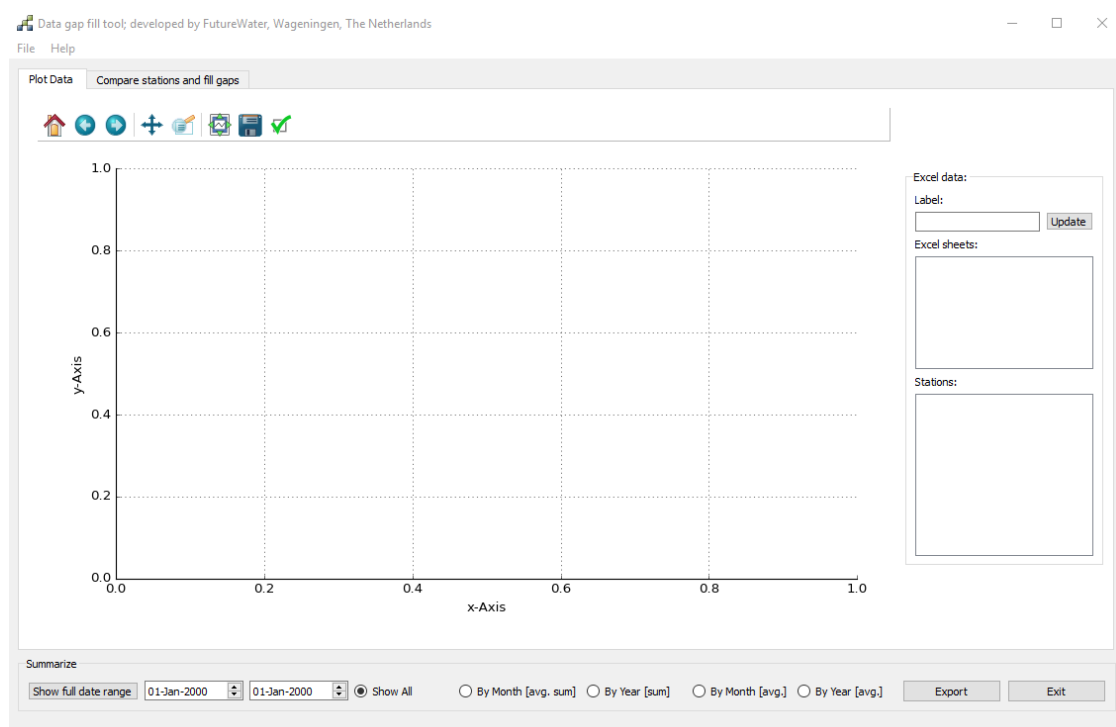


Figure 3: GapFiller interface.

5 Preparing the input data

To fill data gaps with the GapFiller tool your dataset should be saved in an Excel sheet. Furthermore the data should be in a certain format, so that the date and environmental data columns can be recognized. As an example we open the Excel sheet *MultipleDates.xlsx* (see Figure 4). This sheet contains five rows with random information and one row with headers. The Excel sheet that you want to import should contain at least one row similar to this example, with a column called 'Date' and a column next to it that contains the name of the discharge station. Then the GapFiller tool will automatically recognize the header. Missing or erroneous data, which is given as -9999, should be changed to blank cells. The exact date format (e.g. 1-Jan-68 or yyyyymmdd) is not important, as long as the format is defined in Excel as a 'Date' or 'Custom'

format. Missing days do not have to be added to the Data column because this will be handled by the tool.

	A	B	C	D	E	F	G	H
1				3	4	5		
2			dfsdf					
3	sdfdf		sdf		2-Jan-00			
4					1-Jan-00			
5	Daily	m3/s	Daily	m3/s	Daily	m3/s	Daily	m3/s
6	Date	Sebwe	Date	Rwimi	Date	Rukoki	Date	Mubuku
7	1-Jan-68	1.131	25-Apr-52	1.442	18-May-54	3.144	1-Jan-54	13.431
8	2-Jan-68	1.106	26-Apr-52	1.356	19-May-54	2.835	2-Jan-54	13.192
9	3-Jan-68	1.126	27-Apr-52	1.437	20-May-54	2.579	3-Jan-54	13.153
10	4-Jan-68	1.101	28-Apr-52	1.391	21-May-54	3.418	4-Jan-54	12.762
11	5-Jan-68	1.101	29-Apr-52	1.326	22-May-54	3.506	5-Jan-54	12.724
12	6-Jan-68	1.101	30-Apr-52	1.25	23-May-54	2.841	6-Jan-54	12.724
13	7-Jan-68	1.101	1-May-52	1.222	24-May-54	2.865	7-Jan-54	12.724
14	8-Jan-68	1.101	2-May-52	1.213	25-May-54	2.841	8-Jan-54	12.704
15	9-Jan-68	1.101	3-May-52	1.195	26-May-54	2.557	9-Jan-54	12.494
16	10-Jan-68	1.101	4-May-52	1.222	27-May-54	2.39	10-Jan-54	12.285
17	11-Jan-68	1.101	5-May-52	1.304	28-May-54	2.438	11-Jan-54	12.266
18	12-Jan-68	1.092	6-May-52	1.534	29-May-54	2.139	12-Jan-54	12.266
19	13-Jan-68	1.198	7-May-52	1.502	30-May-54	3.089	13-Jan-54	12.266
20	14-Jan-68	1.106	8-May-52	1.442	31-May-54	3.019	14-Jan-54	12.266
21	15-Jan-68	1.073	9-May-52	1.655	1-Jun-54	3.048	15-Jan-54	12.266
22	16-Jan-68	1.044	10-May-52	4.176	2-Jun-54	3.624	16-Jan-54	12.266
23	17-Jan-68	1.044	11-May-52	2.035	3-Jun-54	4.888	17-Jan-54	12.266
24	18-Jan-68	1.044	12-May-52	1.631	4-Jun-54	5.014	18-Jan-54	12.266
25	19-Jan-68	1.044	13-May-52	1.972	5-Jun-54	3.363	19-Jan-54	12.266
26	20-Jan-68	1.044	14-May-52	2.133	6-Jun-54	2.989	20-Jan-54	12.266
27	21-Jan-68	1.039	15-May-52	3.654	7-Jun-54	2.569	21-Jan-54	12.266
28	22-Jan-68	0.994	16-May-52	2.186	8-Jun-54	2.369	22-Jan-54	12.266
29	23-Jan-68	0.989	17-May-52	2.758	9-Jun-54	2.729	23-Jan-54	12.136
30	24-Jan-68	0.989	18-May-52	1.806	10-Jun-54	2.434	24-Jan-54	11.838
31	25-Jan-68	0.989	19-May-52	1.487	11-Jun-54	2.327	25-Jan-54	11.82
32	26-Jan-68	0.989	20-May-52	1.837	12-Jun-54	2.358	26-Jan-54	11.82
33	27-Jan-68	0.989	21-May-52	1.686	13-Jun-54	2.864	27-Jan-54	12.043
34	28-Jan-68	0.989	22-May-52	1.6	14-Jun-54	2.828	28-Jan-54	12.266
35	29-Jan-68	0.989	23-May-52	1.663	15-Jun-54	2.722	29-Jan-54	12.043
36	30-Jan-68	1.008	24-May-52	1.507	16-Jun-54	2.628	30-Jan-54	11.82

Figure 4: Example of environmental data series in an Excel sheet containing discharge data which can be opened by the GapFiller tool.

6 Plotting the data

To open an Excel sheet with the GapFiller tool go to *File* → *Open* and navigate to the directory where the environmental data is saved (in this example it is discharge data). As an example we open the Excel sheet *MultipleDates.xlsx* When this sheet is opened with the GapFiller tool it looks like this (Figure 5):

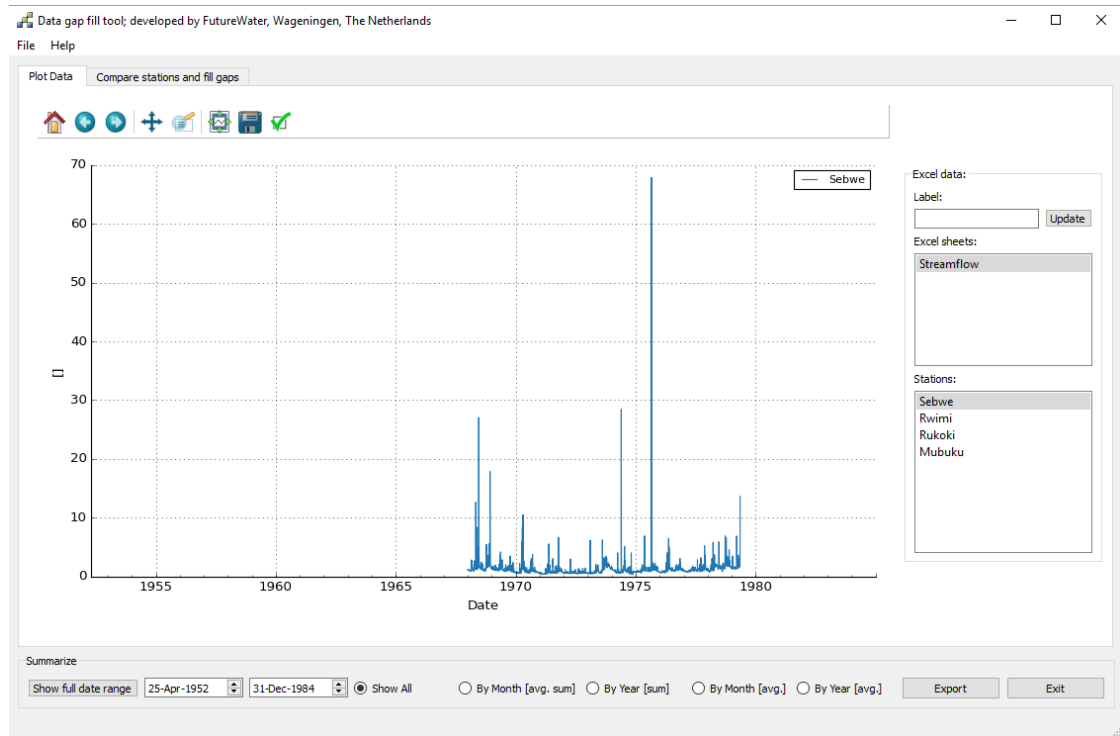


Figure 5. Example of an Excel sheet with discharge data that is opened with the GapFiller tool.

As can be seen at the right of the graph the Station *Sebwe* from the sheet *Streamflow* is selected and shown in the graph. Under *Label* the title for the y-axis can be changed, for example by typing *Discharge (m3/s)* in the box under label and then pressing *Update*.

More stations can be selected by clicking on them under *Stations*. By clicking on a selected station a second time the station will be deselected. The selected stations are marked by a light blue or grey background.

At the top of the graph are different buttons with options to change the view of the graph. With the 'Pan and zoom axes' button (Figure 6) the x-axis and y-axis can be shifted and zoomed independently. When 'Pan and zoom axes' is activated the left mouse button can be used to shift the x- or y-axis, whereas the right mouse button can be used to zoom in and out. Pressing the right mouse button and moving it up it will zoom in on the y-axis, moving it down will zoom out on the y-axis. Moving it right will zoom in on the x-axis and moving it left will zoom out on the x-axis.

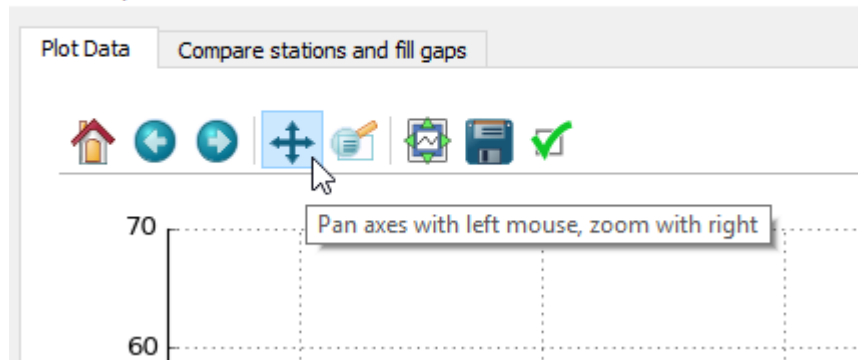


Figure 6. Pan and zoom axes.

The 'zoom to rectangle' button (Figure 7) can be used to zoom in to a certain area of the graph.

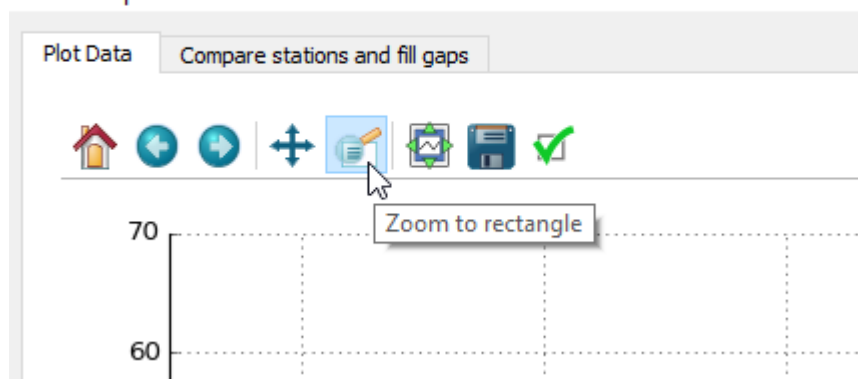


Figure 7. Zoom axes

With the home button (Figure 8) the graph can be reset to the original view. When another station is selected the graph will also be reset to the original view. The 'left' and 'right' buttons next to the home button can be used to navigate back to the previous view or forth to the next view.

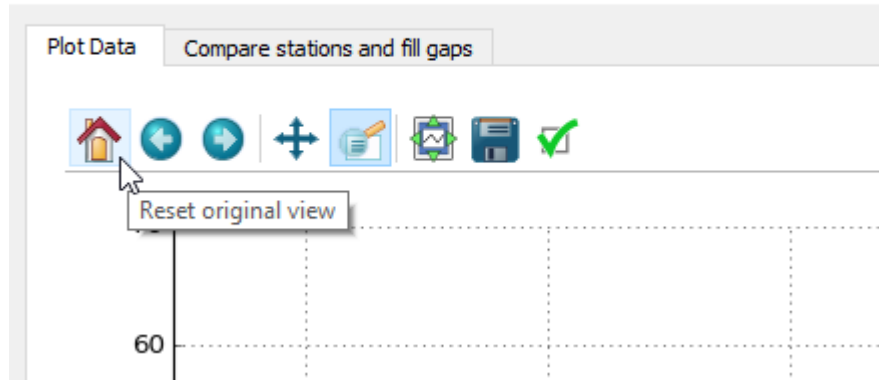


Figure 8. Reset original view

The graph can be saved as different file types (*.png, *.jpeg, etc.) by the 'Save figure' button (Figure 9).

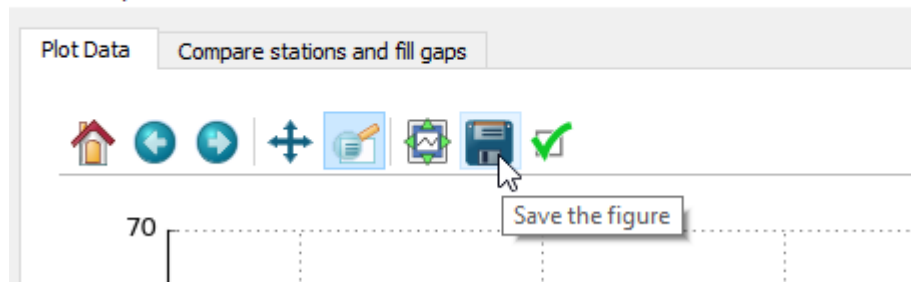


Figure 9. Save figure

At the bottom of the graph are different options for displaying the data. The start and end date can be changed with the date boxes at the bottom left. When the 'Show full data range' button at the most bottom left is pressed, the whole data range is shown again. With the four tick boxes at the bottom right the data can be shown by month or by year as bar plots.

By Month [avg. sum] gives the average monthly sum of the considered environmental variable over the number of years.

By Year [sum] gives the sum of all daily environmental values for each year.

By Month [avg.] gives the average of all daily environmental values for a certain month.

By Year [avg.] the average of all daily environmental values for each year is calculated.

For editing these graphs the layout buttons at the top can be used also.

The 'Export' button at the bottom right of the screen can be used to export the data. Since the data gaps have not been filled yet, the warning 'Gap filled data cannot be exported because gaps are NOT filled' will be shown. However, an Excel sheet called 'Tool_Results.xlsx' will be exported to the location where the GapFiller.py script is located. This Excel sheet contains a tab called 'Unfilled'. Exporting the data without filling the gaps can already be useful because the exported data will be synchronized. This means the stations will have the same start and end date and missing dates will be added and filled with blank values.

7 Compare stations and fill gaps

For comparing the stations and filling the data gaps the tab next to 'PlotData' should be selected (Figure 10).

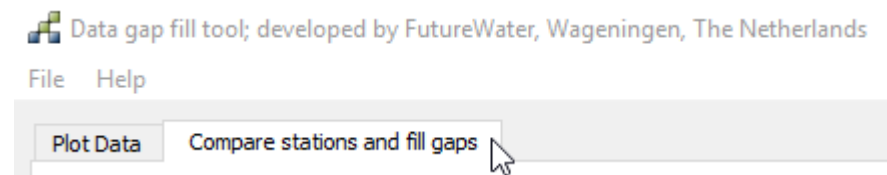


Figure 10. Compare stations and fill gaps tab.

When this tab is selected a scatter plot with the station data is shown at the left of the window (Figure 11). The station data that is shown at the x-axis can be selected in the left box under 'Station 1 (x-axis)'. The stations that are shown on the y-axis can be selected in the right box under 'Stations with gaps to be filled (y-axis)'. Selection of the stations works in the same way as under the 'Plot Data' tab. Also the options for changing the layout of the graph are the same as under the 'Plot Data' tab. For changing the y-axis label the 'Label' option under the 'Plot Data' tab should be used. By checking the 'Cumulative plot' box, the scatter plot becomes cumulative.

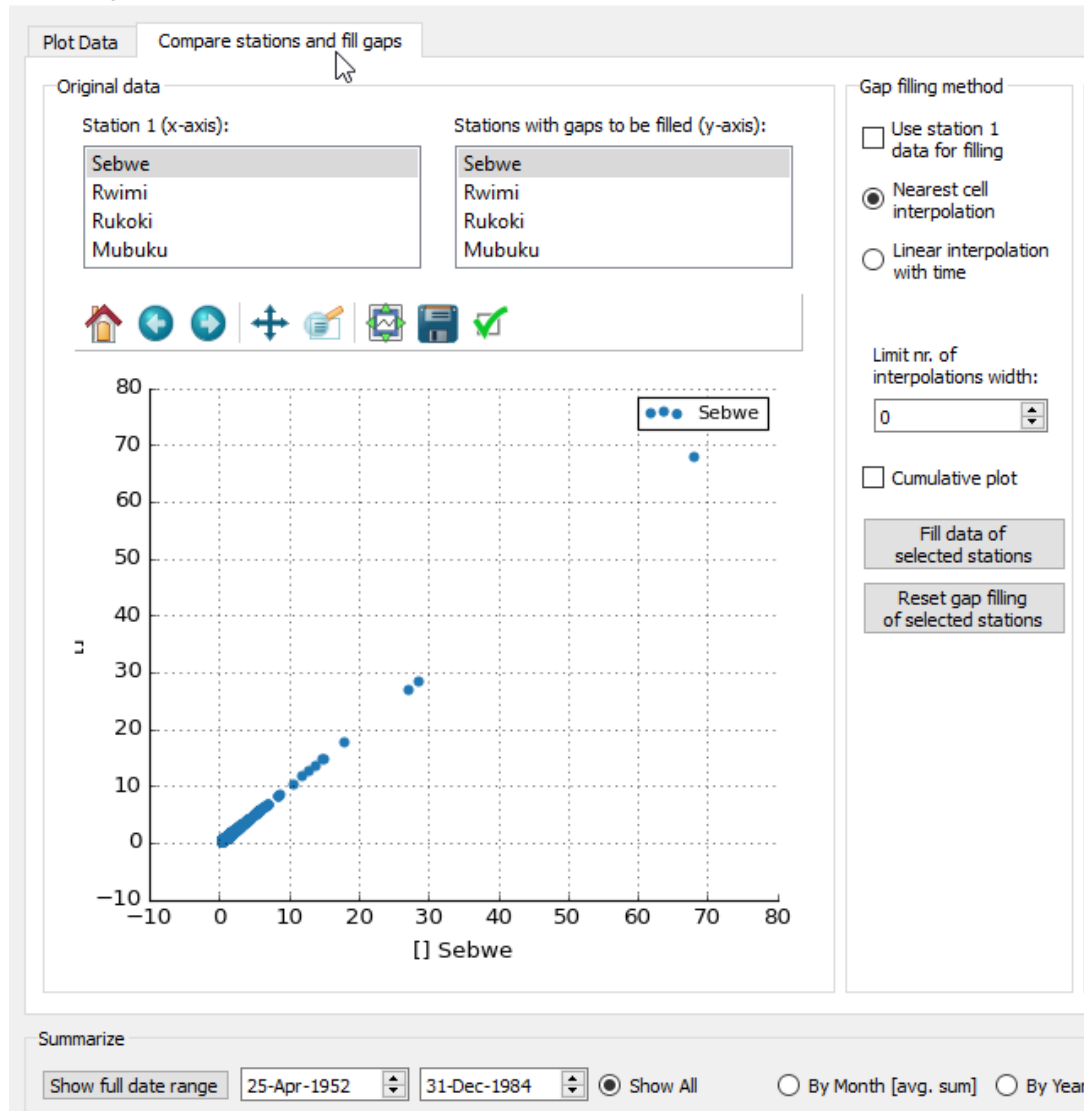


Figure 11. Left part of the 'Compare stations and fill gaps' window.

Under 'Gap filling method', three options for filling the data gaps are shown (Figure 11):

- Use station 1 data for filling
- Nearest cell interpolation
- Linear interpolation with time

By selecting 'Use station 1 data for filling' the station selected in the left box under 'Station 1 (x-axis)' will be used to fill the gaps of the stations that are selected in the right box under 'Stations with gaps to be filled (y-axis)'. If there is a data gap on a certain day this gap will be filled by the data from Station 1 on the corresponding day. However, it is also possible that both stations have data gaps on the same days. Therefore one of the two other options should be chosen as well. When the options 'Nearest cell interpolation' is checked data gaps will be filled with data that is available on the nearest date. For example, if there is a data gap from 2 to 5 March for both stations, the data gap on 2 and 3 March will be filled by the data from 1 March and the data gap on 4 and 5 March will be filled by the data of 6 March. If both dates are equally close the data from the previous day will be used. If 'linear interpolation with time' is chosen as

gap-filling method, the data gaps will be filled by the linear interpolation between the last previous day with data and the first next day with data. If the data series of Station 1 is longer, this data will be used to extend the data series of the other station. If only a data gap on a certain period should be filled this period can be selected at the bottom of the graph.

It is also possible not to fill the data gaps with station data (deselect 'Use station 1 data for filling'), but only by nearest cell or linear interpolation of the data of the station itself.

By default the option to limit the number of interpolations width is zero. This means that for all data gaps all days without data will be filled. However, the number of interpolations can be limited. For example, if there is a data gap of thirty consecutive days and 10 is chosen, the ten first days without data will be filled and the last ten days without data will be filled.

After the right options for filling the data gaps have been selected, data gaps will be filled by pressing 'Fill data of selected stations'. Only stations that are selected in the 'Stations with gaps to be filled (y-axis)' window will be gap filled. In the lower left corner of the GapFiller window the message 'START filling gaps' will appear. When the tool has finished filling the data gaps the message 'END filling gaps' will be shown. By pressing 'Reset gap filling of selected stations' the dataset will be returned to its original state.

GapFilled data can be plotted and analyzed at the right part of the window under 'Gap-filled data' (Figure 12).

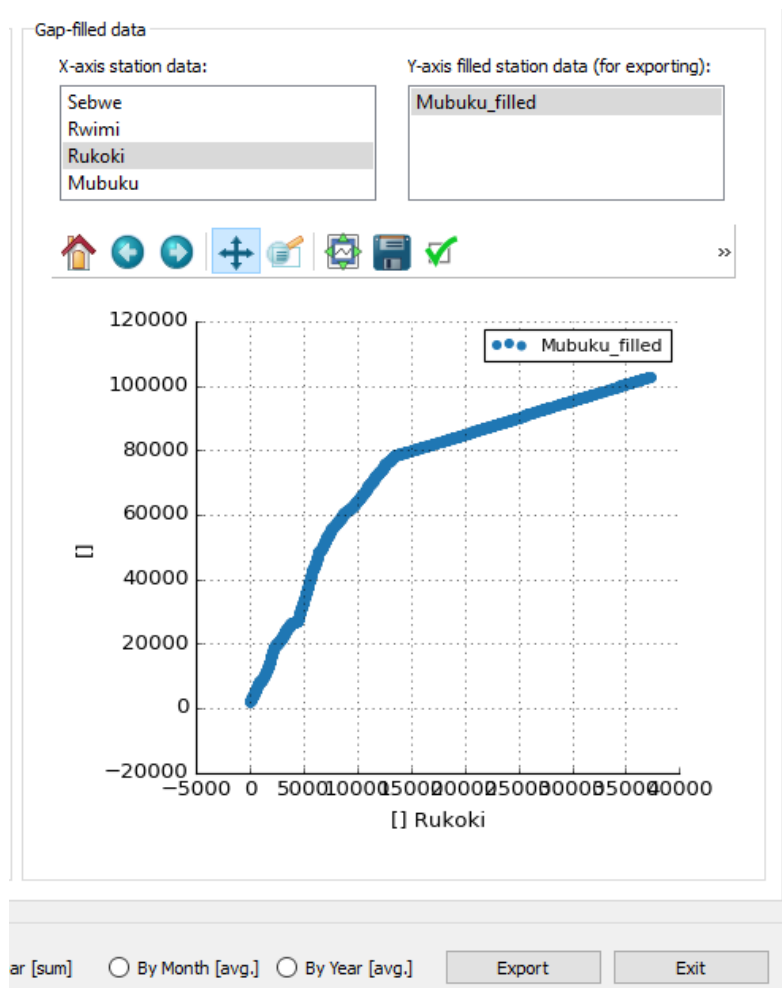


Figure 12. Right part of the 'Compare stations and fill gaps' window.

8 Exporting the results

When the data gaps are filled the results can be exported by pressing the 'Export' button at the lower right of the window. An Excel file named 'Tool_Results' will be saved to the folder where the GapFiller.py script is located. This file contains two tabs: 'Unfilled' and 'Filled'. It was already explained at the end of Chapter 6 that the 'Unfilled' tab contains the 'synchronized' data for all stations. The 'Filled' tab contains the synchronized and gap-filled data for the gap-filled stations. Since previous results in 'Tool_Results' are overwritten every time you press the Export button, you should save this file under another name if you want to keep the previous results.