Quality Control R Script for meteorological and physical sensor data

Version: 07.02.2023
Produced by: Nils Kreuter

The idea behind this quality control (QC) script is the transformation of raw logger data to “ready-to-be-uploaded” data for the SITES Data Portal. The QC itself is more or less rudimental and corrects for sensor failures or obvious measuring errors. The script produces several outputs, which allows to quickly check what the QC has done and helps to trouble shoot.

Install R and Rstudio:

To run the R script download the latest Version of R (or use your existing one) and the latest Version of RStudio (or use your existing one). You don’t need to download any packages for R as the script will check your downloaded packages and download the ones needed if not installed. Keep in mind that future R versions and future changes of R packages might lead to failures of the script.

R: https://cran.rstudio.com/

RStudio: https://posit.co/download/rstudio-desktop/

What is the QC script capable of?

- **Gapfill**:
  - Procedure that looks at an array of data containing dates that have an evenly spaced measurement interval that is predefined by the operator of the QC. Every row is checked to see if the expected interval of data matches the raw data. If not, missing rows containing time information are added and NAs for each parameter are added to the data. Rows that were added or deleted can be found in the “Gapfill_added_rows.csv” and “Gapfill_removed_rows.csv” files. If times deviate from the expected interval they are deleted and can be found in the respective output csv file.

- **DoubleTime**:
  - Procedure checking whether every timestamp is only occurring once in the data sheet. In case lines are found to have the same time, only the first one will be kept and the deleted lines stored in the “Gapfill_removed_rows.csv”. **In case this happens, you need to manually check why a date occurred twice. There might be a general problem with the logger or data.**

- **Max/Min:**
  - Procedure to check meteorological data and replace these data with NA if the values exceed a maximum or minimum threshold. The thresholds are specific values that can be manually defined in the “thresholds.R” file (More info below). Before manually defined values are checked, a short check will be run to replace all "-7999" values (Thresholds) with NA. For some parameters the minimum-threshold is set below 0 even though values might not be possible to reach below

0. This step was added to account for inaccuracies. In case where the value is below 0 but above the minimum threshold the value is set to 0.

- **Repeat**:
    - Procedure to check the data and replace these data with NA if the values repeat (exactly). The parameters that should be checked with this method are defined in the "parameters.R" file (More info below) and can be adjusted. Two consecutive equal values are allowed, but all other number of repeating values in the series are set to NA.

All of the following scripts and **the raw data** need to be in the same folder in order to run:

- **QC_SITES.R**
    - This is the main script that will execute the QC script and will take information from the other scripts mentioned below

- **parameters.R**
    - Contains the parameters used for the QC. Needs to be adjusted to individual use (More info below)
    - In addition, it can be specified which parameters and should be checked within the repeated sequences section

- **thresholds.R**
    - Thresholds for minimum and maximum values can be set in this file. This script needs to be revisited every time the script is used to adjust for changing environments and conditions. (More info below)

- **SITES_parameters_hourly.R**
    - Change parameter names to SITES guidelines to get an output for hourly data that can be uploaded to the SITES Data Portal. Different outputs can be individually defined.

Guide:

1. Download the R scripts from the GitHub repository: https://github.com/fieldSITES/scripts/tree/main/QC%20R%20script

2. Move the scripts into a desired separate folder on your computer

3. Copy your raw data into the folder with all the above-mentioned R scripts.

4. Open the "parameters.R" file with RStudio.
    a. Exchange the variables in the brackets ("EXAMPLE") with the name of the corresponding variable in your raw data. Not used parameters can be left empty (""). At the moment only the parameters shown in this file can be used in the QC script. **Be aware that R is case-sensitive!**

b. Scroll down in the file to adjust the parameters that should be used for the “Repeat” step. You can delete or add a “#” to the ones you do not want to check. Make sure that the first line does not contain the “rbind” command.
c. Save and close the “parameters.R” file.

5. Open the “thresholds.R” file with RStudio.
    a. In this file you can change certain thresholds for given parameters. As of now only the ones shown in the file are able to be controlled for. Keep in mind that all values above or below the thresholds you set will be deleted in the raw data later. Make sure to account for extreme events in the past and to revisit this step every time you use the script to adjust the thresholds.
    b. The thresholds are either in groups of 12 (For each month, starting from January) or one value that is general for the whole year.
    c. Save and close the “parameters.R” file.

6. Open the “SITES_parameters_hourly.R” file with RStudio.
    a. This is used to change the parameters from each station to the respective SITES guideline.
    b. Exchange the variables with the names of the corresponding variable in your raw data and the SITES Guideline

        Example: (Better formatted in RStudio)

        names(QC_data_final_SITES)[names(QC_data_final_SITES) == 'NAME OF YOUR PARAMETER in your datasheet'] <- 'SITES SPECIFIC NAME FOR PARAMETER'

    c. At the moment only the parameters shown in this file can be used in the QC script. **Be aware that R is case-sensitive!**
    d. This step includes an automated conversion of hPa pressure (Used at Erken) to Pa (Used by SITES). In case your station already measures pressure in Pa units you can delete the respective conversion line in the file.
    e. At the lower end of the script you can adjust what kind of SITES-specific output the script should produce. As of now this is mainly focused on Erken data and therefore needs adjustment for the specific station.
    f. Save and close the “SITES_parameters_hourly.R” file.

7. Open the “QC_SITES.R” file in RStudio
    a. Adjust the size of the main window to have a nicer formatting of the script. Once the box (drawn with the “#”) is aligned the formatting is easier to read.
    b. Every step within the script is explained in detail. However, if certain loops repeat only the first loop is explained
    c. Start the script by clicking into the main window and pressing “CTRL+SHIFT+S”
    d. R will start with checking for any packages that might need to be downloaded. (Installing packages only happens the first time and might take a bit of time)
    e. When everything is installed ten questions are shown that need your input in the console on the bottom left
        i. <u>Enter name of datafile including extension (.csv,.dat,.txt etc.):</u>

1. You need to enter your specific raw data file name (example_data.dat) without any brackets or “”. (**Be aware that R is case-sensitive!).** Once you entered your file name press “enter” on your keyboard.

ii. <u>In which row of the logger output are the parameters stated? (TIMESTAMP, RECORD etc.)(Number):</u>
    1. For this you need to look at the raw data file. Give the row-number in which the parameters are and press “enter” on your keyboard.

iii. <u>In which row of the logger does the raw data start?(Number):</u>
    1. For this you need to look at the raw data file. Give the row-number of your first data point and press “enter” on your keyboard.

iv. <u>In which column is the first measured parameter stated? (NOT TIMESTAMP or RECNBR)(Number):</u>
    1. For this you need to look at the raw data file. Give the column-number of your first data point. If your data looks like this for example: “TIMESTAMP”, “RECNBR”,”AIR_TEMP”; enter 3 and press “enter” on your keyboard

v. <u>At what interval is the data taken? (In minutes) (1-day = 1440):</u>
    1. Enter the interval as a number of minutes and press “enter” on your keyboard.

vi. <u>Enter name of parameter file including extension (.R):</u>
    1. Enter parameters.R and press “enter” on your keyboard.

vii. <u>Enter name of threshold file including extension (.R):</u>
    1. Enter thresholds.R and press “enter” on your keyboard.

viii. <u>Enter name of SITES parameter file including extension (.R):</u>
    1. Enter SITES_parameters_hourly.R and press “enter” on your keyboard.

ix. <u>Enter the path where the Output (Folder for SITES and QC) should be saved. (Leaving it empty will save all the files in the folder where the script was opened):</u>
    1. This will let you determine where all the Output (See below for more info) will be stored. You can either leave it empty (will save all the files in the folder where the script was opened) or enter a path like: C:\Users\YOURUSERNAME\Documents\Data\QC_Output

x. <u>Should a .csv file be created for each step of the QC (in addition to the final Output)? (y/n):</u>
    1. You can decide whether you want to have a csv file for every step of the QC. This might be helpful to trouble shoot the script, but as you will receive a final Output this might not be necessary. Enter “y” for yes or “n” for no (Do not write the “”) and press “enter” on your keyboard.

xi. The script will now show you a “[1] "QC started"” indicating it has started.

xii. While it is running do not open the other R scripts or enter something in the console. It might take some time for the script to finish.

xiii. When it finished it will show a “[1] "QC DONE"” and you can close RStudio.

Output:

- The QC script will create two folders in your desired location (See question ix above):
    - **QC**
        - Output_info_hourly.txt
            - Text file with all relevant information about the script. (When it was run, what data was deleted & the thresholds used, etc.)
        - Csv-files with only the deleted values for each step of the QC
        - Csv-file with the final complete data ("QC_data_final.csv")
        - Plot-Folder
            - Showing graphs for all parameters in the data
            - Deleted values are marked in red and the legend gives information how many data points were deleted and how many are NAs
    - **SITES**
        - Different csv-files for the previously defined data upload for SITES structured like the following:
            - QC_data_hourly_1_SITES.csv
            - QC_data_hourly_2_SITES.csv
            - …