Manual for downloading NWIS data

using R-scripts

Updated: 2024-05-02 J. Davis

# Description:

This R Project and its associated R scripts are intended to facilitate model calibration by streamlining the downloading/processing of average daily flow and instantaneous (e.g., 15-min sensor data) monitoring data from the National Water Information System (NWIS). Please note that these scripts do \*\*not\*\* query or download any water quality data based on grab samples, nor are they validated for parameters that are not currently listed in the data download lookup table.

This R Project leverages functions available in the USGS ‘dataRetrieval’ package (De Cicco et al. 2018) and the ‘tidyverse’ suite of packages (Wickham 2017), and interfaces them with an Excel lookup table. This allows the user to select USGS gages and water quality parameters to download without directly editing R scripts. The scripts use ‘readNWISdv’ to download average daily flow data, ‘readNWISuv’ to download instantaneous water quality, flow, and gage height data (e.g., 15-min intervals), and ‘readNWISsite’ to download the associated location data.

Raw data are processed and formatted to aid in the calibration of water quality models. When possible, the scripts use ‘do.Saturation’ (Moulton 2018) to back-calculate DO percent saturation when it is not directly measured by NWIS. During processing, timestamps for daily average flow values are changed from midnight to noon (12:00 pm). Instantaneous data are subset to ≤4 records per day and formatted for import into the Water Resources Database (WRDB). The original (i.e., non-subsetted) instantaneous water quality day are used to calculate daily statistics (e.g., daily difference, average, etc.). A variety of time series plots are generated from the measured and daily aggregated datasets.

Prior to running the calibration scripts, the user will need

* RStudio, R Statistical Software (R v4.4.0), and all required R packages to be pre-installed
* List of USGS gage numbers and USGS parameter codes

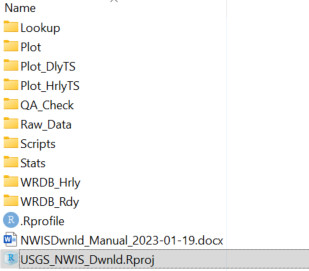
Once completed, the scripts will generate a variety of output, including the following:

* WRDB formatted results and associated support tables (e.g., Station table, Pcode table, etc.)
  + For flow, outputs daily average results table (timestamp reported as noon for each day)
  + For water quality parameters, outputs three separate results tables: full instantaneous dataset w/ duplicate records removed, hourly average, and instantaneous filtered to 4 records per day (6:00, 12:00, 18:00, and 24:00). All times are reported based on user-selected time zone.
* Time series plots for all flow, gage height, and water quality parameters, by USGS gage
* For instantaneous parameters, additional daily summary and hourly average plots
* Summary tables reporting summary statistics for each parameter
* Raw, unformatted data files that are directly from USGS NWIS
* Various QA/QC files

# Set-up required prior to running R scripts

## Unzipping R Project folder

1. Unzip the R Project calibration file and save the entire folder. This can be anywhere because the R Project automatically knows its location. When unzipped, the following folders and files should be available.



# Create R Project

## Learning how to create and work with R projects:

Implementing these scripts as a Project in R Studio eliminates the need to edit any of the underlying R scripts when processing data. It also allows the entire R Project folder to be copied/pasted to new file directory when a new dataset needs to be downloaded/processed. The project is designed so the user only needs to update an Excel lookup table and then run the ‘master script’ via R Studio. Unless a critical error is encountered, the user should not need to open or edit any of the subscripts.

Although the NWIS download scripts are designed to allow a user to run the scripts with minimal understanding of R syntax, a working knowledge of the R language will facilitate the process and help troubleshoot. The implementation of the R calibration scripts as an R Project also helps to organize the various input and output files. The R Project working directory will default to the location of the R Project \*.RProj file, which allows the user to apply the scripts to new projects with minimal to no changes to the lookup table.

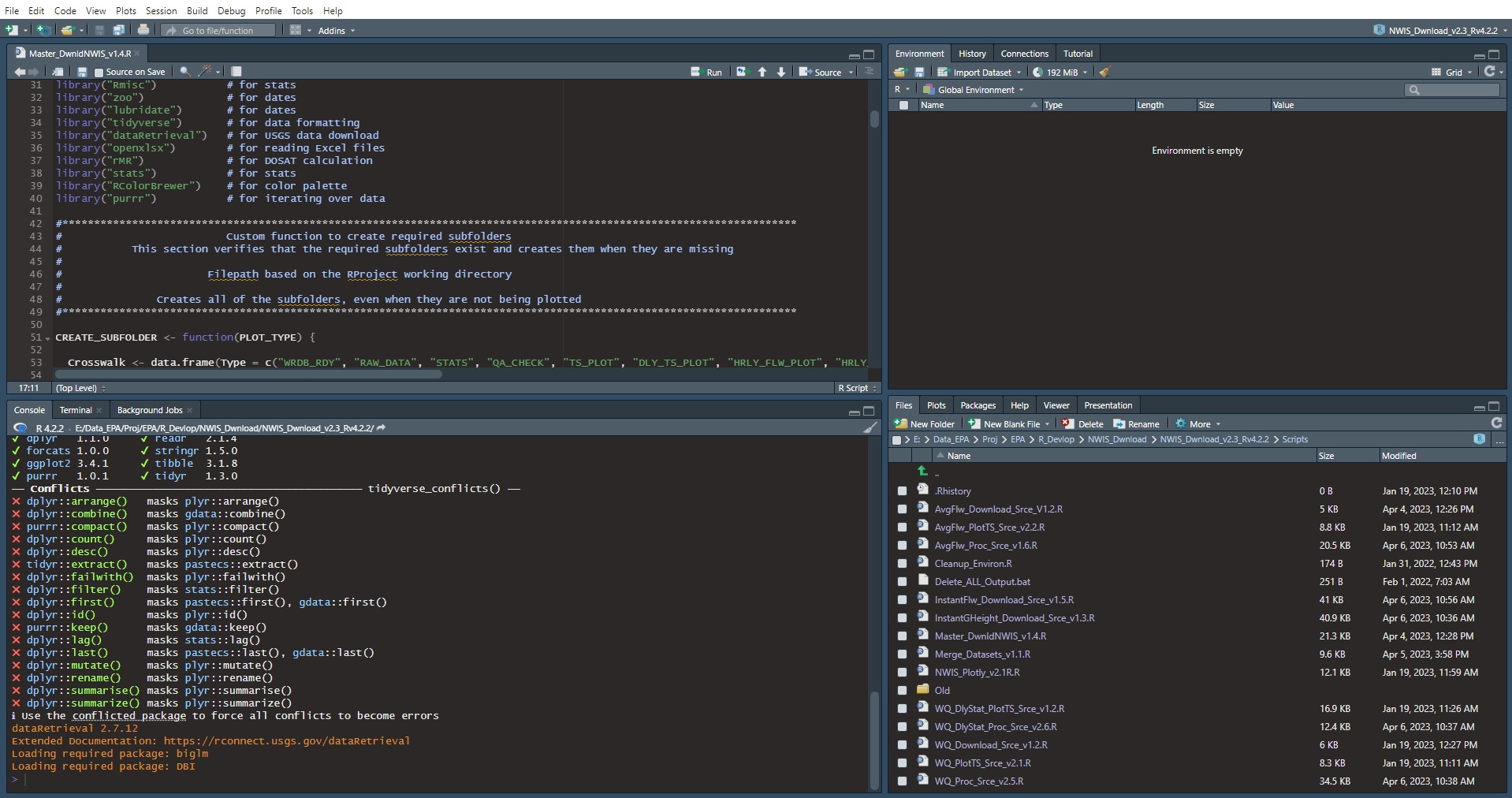
A basic knowledge of creating and working with R Projects in RStudio will be beneficial. A variety of resources and tutorials (e.g., <https://stackoverflow.com>) are available online for learning rudimentary R functions/syntax and working with RStudio. For an overview of creating R projects in Rstudio see <https://support.rstudio.com/hc/en-us/articles/200526207-Using-Projects>.

It is recommended to periodically update your version of RStudio, R Statistical Software, and the R packages to make sure you are running the most current versions.

## Overview of R project pane

The easiest way to run the associated calibration scripts is to run them interactively as an R Project through RStudio. The R project can be opened by double clicking the R Project file (USGS\_NWIS\_Dwnld.Rproj) that is located in the parent directory of the calibration package.

This will open RStudio, which will display a variety of panels that can be customized (Tools 🡪 Global Options 🡪 Pane Layout). The format displayed below shows the Source panel with the currently active script (upper left panel), the list of objects that were created in the current R environment (upper right panel; no objects will be available until the scripts are executed in the R console), the R console which represents the actual R Statistical Program where the scripts are executed (bottom left panel), and the folders contained within the working directory of the R Project (bottom right).



## Pre-installing R packages

The scripts included in this R Project require a variety of R packages that are listed in the table below and will need to be downloaded into your local R library. This will save the various packages locally and allow the scripts to call the required functions as necessary.

If not prompted by R Studio, the following line of code can be copied and pasted into the R console (after the “>” symbol) and pressing ‘Enter.’ This will tell R to download the required packages and install them to your local R package library.

install.packages(c("gdata", "Rmisc", "zoo ", "lubridate", "tidyverse", "ggplot2", "dataRetrieval", "openxlsx", "rMR", "RColorBrewer", "plotly", “purr”))))

|  |  |
| --- | --- |
| **R Package** | **Reason** |
| install.packages(‘gdata) | # for data processing |
| install.packages(‘Rmisc’) | # for stats |
| install.packages(‘zoo’) | # for working with dates |
| install.packages(‘lubridate’) | # for working with dates |
| install.packages(‘tidyverse’) | # for data reformatting |
| install.packages(‘ggplot2’) | # for plotting functions |
| install.packages(‘dataRetrieval’) | # for NWIS download functions |
| install.packages(‘openxlsx’) | # for reading/writing Excel files |
| install.packages(‘rMR’) | # for DOSAT calculations |
| install.packages(‘RColorBrewer’) | # for color palette |
| install.packages(‘purrr’) | # for iterating over data |
| install.packages(‘plotly’) | # for optional plotly interactive plots |

## Creating folder hierarchy for R project

To help organize the large number of files required and generated by this project, the scripts require a strict file tree. If a subfolder is missing, the scripts will automatically create it in the current working directory. When unzipped, the following files and folders should have been created. They cannot be renamed nor moved:

* **Lookup:** Where the Excel \*.xlsx lookup file is located. This is where the user enters the gages and parameters to download.
* **Plot:** Where time series figures are saved (i.e., daily average flow, instantaneous WQ)
* **Plot\_DlyTS**: Where daily summary time series figures are saved
* **Plot\_HrlyTS**: Where hourly average summary time series figures are saved
* **QA\_Check:** Where QA/QC output are saved
* **Stats**: Where summary statistic tables are saved
* **Raw\_Data:** Where the raw unprocessed downloaded datafiles are saved
* **Scripts:** Location of the data download scripts that will be sourced by R
* **WRDB\_Hrly:** Where the WRDB formatted hourly average results are saved
* **WRDB\_Rdy:** Where the WRDB formatted results and support tables are saved
* **.RProfile:** Text file that automatically loads the necessary packages when the R Project is opened
* **USGS\_NWIS\_Dwnld.Rproj:** Rstudio project file that can be used to run the scripts interactively

# Set-up script options in Excel data download lookup table:

The R scripts require a variety of settings that can be defined within the Excel download lookup table (i.e., USGS\_NWIS\_Lookup.xlsx). This allows the user to set the necessary arguments without having to directly edit the scripts. NOTE: The name of this file or worksheet tab \*\*cannot\*\* be changed nor can the location of the ‘Lookup’ subfolder be changed from the location of the .RProj file.

The Gage\_Lookup tab is where the user has the ability to set a variety of options that will customize the download process. This tab is broken into several sections that are discussed separately.

## USGS gage selection:



* **Gage\_Number:**  This is a list of USGS gage numbers that will be used to query NWIS. The gage number provided must match a known USGS gage identifier and needs to be prepended with ‘USGS-.’ This helps avoid Excel’s stripping leading zeros from numbers.
  + NOTE: All gages listed in this column will be queried for the selected parameters.
* **Gage\_Elev\_m:** If known, the elevation of the gage in meters can be entered in this field. This elevation will be used to back-calculate DO saturation when temperature and DO are known. Blank cells will be ignored.
* **Gage\_Salin\_ppt:** If known, an average salinity in parts per thousand (ppt) can be entered in this field and used to back-calculate DO saturation (when DO mg/L and temperature are available). Blank cells will be ignored.
  + NOTE: If measured salinity is available and downloaded for a gage and record, the script will use the measured data rather than the salinity value reported here.

## Parameter selection

* Two separate matrices are used to select the water quality parameters that will be queried from NWIS. The USGS\_parmCd has to match known USGS parameter codes. Average daily flows are queried using ‘readNWISdv’ and instantaneous data are queried using ‘readNWISuv.’





* **Selected\_WQ:** Any cell with an ‘X’ in this field will query the NWIS database for instantaneous measurements for the analyte indicated in the corresponding ‘USGS\_parmCd’ field.
* **USGS\_parmCd:** Five digit USGS defined parameter code. Values reported in this field have to correspond to a known USGS parameter code (see https://nwis.waterdata.usgs.gov/nwis).
  + NOTE**:** Excel will strip leading zeros from these codes, but the R scripts will automatically add them back in during data processing.
* **Name\_WQ:** Name of the analyte indicated in the USGS\_parmCd field. This name will not be used within the scripts and only included for user reference.
* **Pcode\_WQ:** User defined parameter code that will be used to abbreviate the water quality parameter. This code must be unique for each parameter and will be used in all data output and figures. NOTE: The code cannot include any spaces.
* **Selected\_Flw/Flow (Including Gage Height):** If an ‘X’ is entered in this field, the scripts will query NWIS to download the average daily flows and/or instantaneous flows all gages listed in the Gage\_Number column. For gage height, only instantaneous gage heights will be queried and downloaded. NOTE: That both average daily and instantaneous can be activated, but that the downloading of instantaneous flows can be time intensive.

## Query date selection





* **Start\_Date:** Inclusive start date of the data query.
* **End\_Date:** Inclusive end date of the data query.
* **Selected\_TZ/Gage\_TimeZone:** The field with an ‘X’ will use the corresponding time zone listed in the Gage\_TimeZone. This allows the user to query and report out data in the correct time zone.

# Running the data download scripts as an R Project

1. After updating the data download lookup table with the required information, double-click the R Project (USGS\_NWIS\_Dwnld.Rproj).
2. The R Project will open in Rstudio and automatically set the R working directory to the location of this \*.RProj file. All output will be saved to relative filepaths associated with the working directory.
3. Within R Studio, open the ‘Master\_DwnldNWIS.R’ script either by navigating to the ‘Files’ tab on the appropriate pane (i.e., bottom right pane in the example) or via File 🡪 Open File then navigating to the location of the ‘Scripts’ subfolder. This will only open the Master script in the Source window pane (e.g., upper left pane in the example).
   1. NOTE: Unless the scripts encounter a critical error, the user should not need to open or edit any of the additional scripts located in this folder. The user only needs to run the Master\_Dwnld script, as it will source the other sub-scripts as needed.
4. To execute the script and send it to the R Console, the user can highlight all text in the SOURCE panel and either press ‘Ctrl + Enter’ or by click the Run icon. A third option is to simply click the ‘Source’ icon.



1. If the scripts run without errors, the R Console will display various messages to indicate data processing progress.
2. When the R Console is actively running a particular script, the R Console will have a stop sign displayed in the upper right corner (if necessary, the script can be terminated by clicking the stop sign).
3. Once completed, the plots and output will be written to the default filepath (i.e., the location of the R Project). See Appendix A for an overview of the output and where it is located.

# Overview of major data processing steps

## Average daily flow

* Average daily flow is downloaded using “readNWISdv” based on parameter code “00060” and statistical code “00003”
* Before any processing of raw datafile, it is saved as a \*.csv file in ‘Raw\_Data’ folder.
* Time stamp of daily average flow changed from 00:00 (midnight) to 12:00 (noon)
* Flow in cubic meters per second calculated from flow in cubic feet per second, with a separate parameter code assigned to each (FLOW\_CMS and FLOW\_CFS, respectively).
  + Calculated values designated with a Ccode = “CAL”
* WRDB result and supporting tables created and saved in ‘WRDB\_Rdy’ folder
  + For data tracking, a placeholder LCode with the query date is added to the dataset
* Time series plots generated and saved in ‘Plot’ folder
* Summary statistics calculated and saved in ‘Stats’ folder

## Instantaneous water quality / flow / gage height parameters

* Instantaneous water quality data downloaded using ‘readNWISuv’ based on their USGS parmCd and statistical code “00000”
* Before any processing of raw datafile, it is saved as a \*.csv file in ‘Raw\_Data’ folder.
* If DO, water temperature, and salinity are measured, DO saturation (%) is back-calculated. Uses default salinity values in lookup table, if no instantaneous salinity values available. Calculations use ‘do.Saturation’ function from rMR package (De Cicco et al. 2018). If DO saturation was directly measured and available in NWIS, DO saturation only back-calculated with ‘do.Saturation’ for those records when measured DO saturation is missing.
* Calculated values designated with a Ccode = “CAL
* Replaces sensor errors (i.e., value = -999999) with a blank record
* Parameter codes defined in lookup table used to remap USGS parmCd to user defined Pcode.
* Full dataset subset to 4 or fewer records per day (6am, noon, 6pm, and midnight; or closest record to those times without duplication).
* Data subset saved as WRDB result and supporting tables. Saved in ‘WRDB\_Rdy’ folder.
  + For data tracking, a placeholder LCode with the query date is added to dataset
* The scripts also processed the full dataset by mapping USGS parmCd to user defined Pcode, then writing out a WRDB formatted results table that contains all instantaneous records
* Using the full dataset, the scripts will also generate average hourly values for all water quality parameters and write out a WRDB formatted results table that contains hourly averages.
  + If any of the records used to calculate a particular hourly average have a ‘Provisional’ Rcode (i.e, P), the scripts will assign a provisional Rcode to that hourly average, even if all other Rcodes used for that particular hourly average have an ‘Accepted’ Rcode (i.e, A).
* Summary statistics calculated across entire dataset and saved in ‘Stats’ folder
* Based on the original instantaneous data that were not subset to 4 records per day, daily summary statistics are calculated (daily min, max, percentiles, etc). The scripts calculate the values from the original instantaneous data and do not query NWIS based on daily min or max statistical codes.
* Time series plots of 4 records per day data subset and daily statistics generated. Saved in ‘Plot’ folder
* Time series plots of hourly averages are also exported and saved in ‘Plot\_HrlyTS’

## Instantaneous flow / gage height

* Similar workflow as the instantaneous WQ data processing
* Instantaneous flow data are downloaded using ‘readNWISuv’ based on their USGS parmCd “00060” and statistical code “00000,” while gage height is downloaded based on USGS parmCd “00065” and statistical code “00000”
* Before any processing of raw datafile, it is saved as a \*.csv file in ‘Raw\_Data’ folder.
* Applies unit conversion to convert flow in cubic feet per second (CFS) into cubic meter per second (CMS), and assigns a new Pcode (FLOW\_CMS).
* For gage height, a conversion from feet to meters is applied and both parameters are exported as separate time series (GH\_INST\_M & GH\_INST\_FT)
* Removes any true duplicate records based on Station and Date/Time
* Full dataset subset to 4 or fewer records per day (6am, noon, 6pm, and midnight; or closest record to those times without duplication).
* Calculates daily average, daily min, daily max, and hourly flow. These are based on the raw instantaneous values and are not based on a separate NWIS query based on daily min or daily max statistical codes.
* All calculated flows assigned CCode = CAL
* Data are exported as instantaneous data subset to four records per day, hourly averages, as well as daily averages.
* Plots are also generated for the various data types and saved in their respective folders.

Appendix A:

Data output

# Script output: Datafiles

## Raw Data:

* The scripts will create separate .csv files for raw average daily flow (USGS\_FlwRaw\_XXXXXX.csv) and instantaneous water quality parameters (USGS\_WGRaw\_XXXXXX.csv). Associated station metadata are also reported out. No formatting or processing have been applied to these files. Filenames are appended with the date of the query and saved in the ‘Raw\_Data’ subfolder.

## WRDB ready datafiles:

* The scripts will generate results and associated support tables (Rcode, Pcode, Station, and Ccode tables) that can be imported into a Water Resources Database (WRDB) project. Three different versions of the water quality data are exported.
  + Average daily flow data are saved separately from instantaneous parameters. Filenames are appended with the query date and saved in the ‘WRDB\_Rdy’ subfolder.
  + Instantaneous water quality measurements that have been subset to 4 records per day (~6:00, 12:00, 18:00, 24:00). Filename appended with ‘Subset.’
  + Full dataset containing all instantaneous water quality measurements, with their parameters mapped to user defined Pcodes and duplicate records removed. Filename appended with ‘AllData.’
  + Hourly average values for water quality parameters based on full dataset. Filename appended with ‘HrlyAvg.’

## Daily summary files

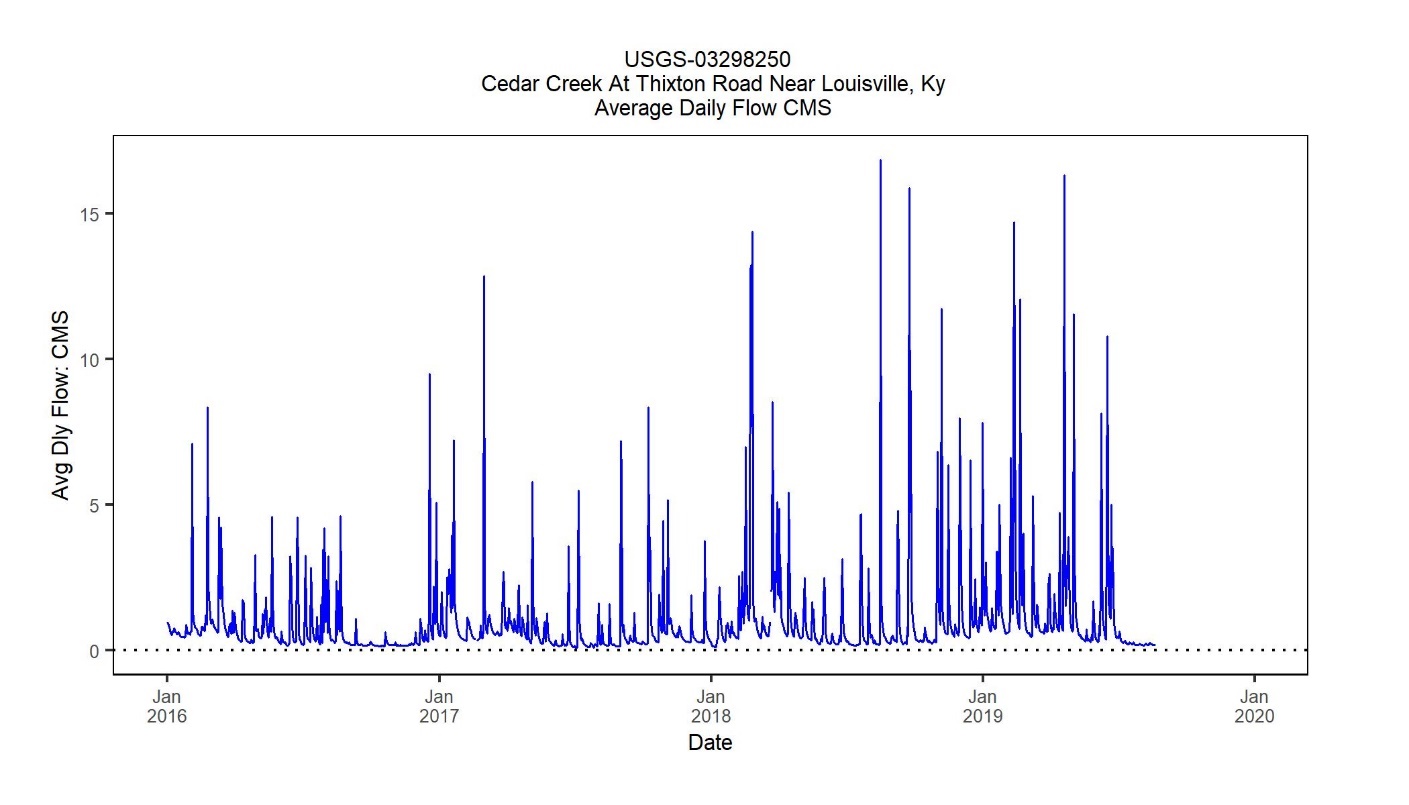
* Up to three files are saved in the ‘Stats’ subfolder, helping to identify outliers or errors.
  + USGS\_Flw\_Sumry\_Stats\_XXXXXX.csv: Provides data inventory, global statistics, and annual summary statistics for each gage. Annual statistics are reported separately for flow in CMS and flow in CFS. Full summary statistics presented in long format.
  + USGS\_WQ\_Sumry\_Stats\_XXXXXX.csv: For each gage and parameter, provides a range of summary statistics based on the entire dataset, including a basic data inventory. Parameters or gages without data are indicated by blank cells. Also, includes a tab reporting daily statistics (e.g., daily min and daily max) and daily difference (e.g., Dly\_Diff = Daily Max – Daily Min). NOTE: This file contains more statistics than what is plotted (i.e., daily 20th and 50th percentiles).

## QA/QC files

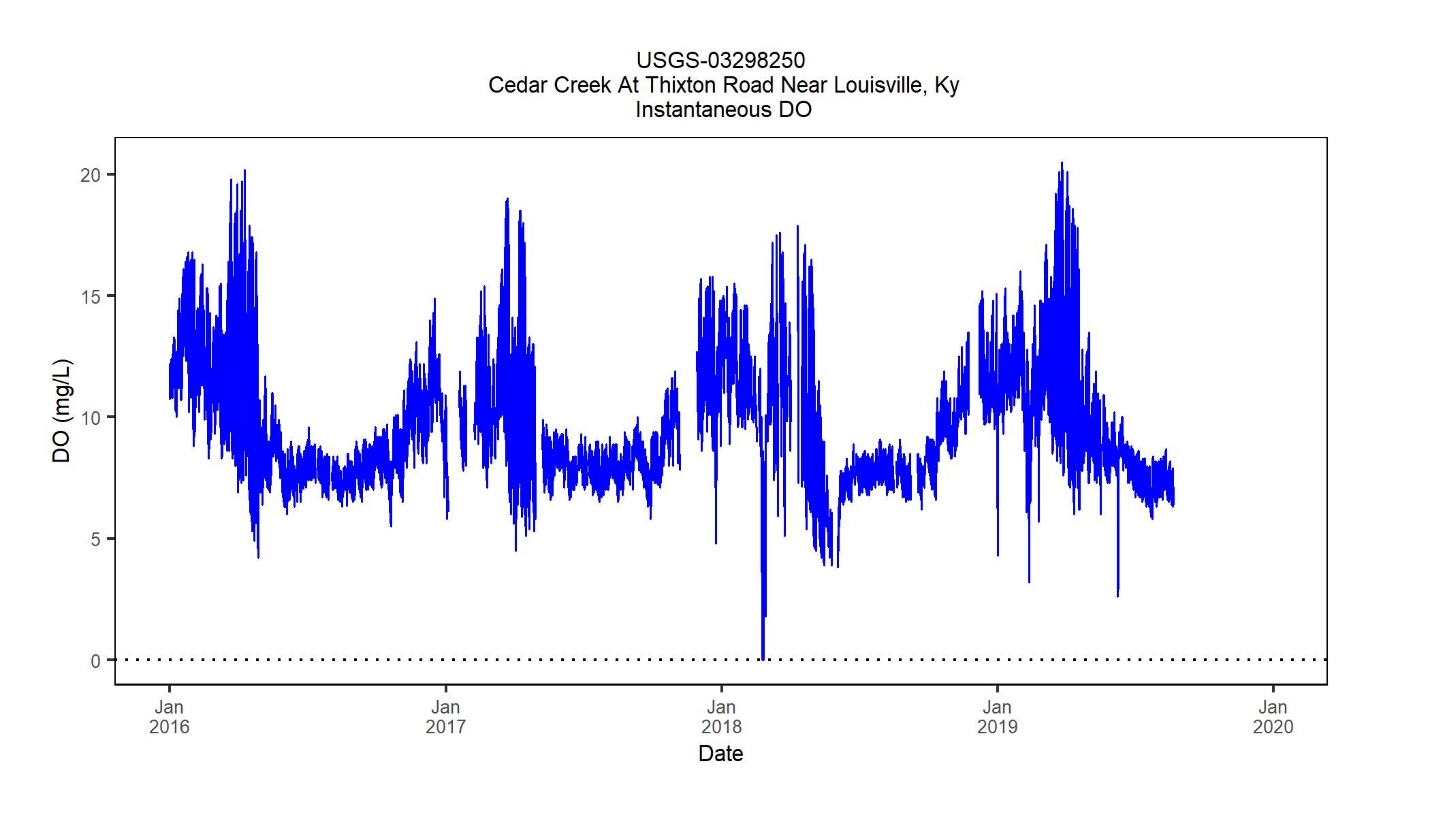
* For data archiving and tracking, a variety of echo and query settings files are saved in the QA\_Check subfolder. These files can be compared with the processed data to verify that raw data were correctly queried and processed. All files are appended with the date of the query.
  + FlwQuery\_Echo\_XXXXXX.csv: Summarizes which gages were queried and possessed data, in addition to other query settings.
  + Pcode\_Inventory\_XXXXXX.csv: Indicates which parameters had data at each gage. Empty cells indicate that gage did not have data for that parameter.
  + WQQuery\_Echo\_XXXXXX.csv: Summarizes the settings that were provided in Excel lookup table and used for the data query.
  + WQQuery\_Settings\_XXXXXX.csv: Summarizes additional settings that were provided in the lookup table.

# Script output: Time series of measured data

## Average daily flow



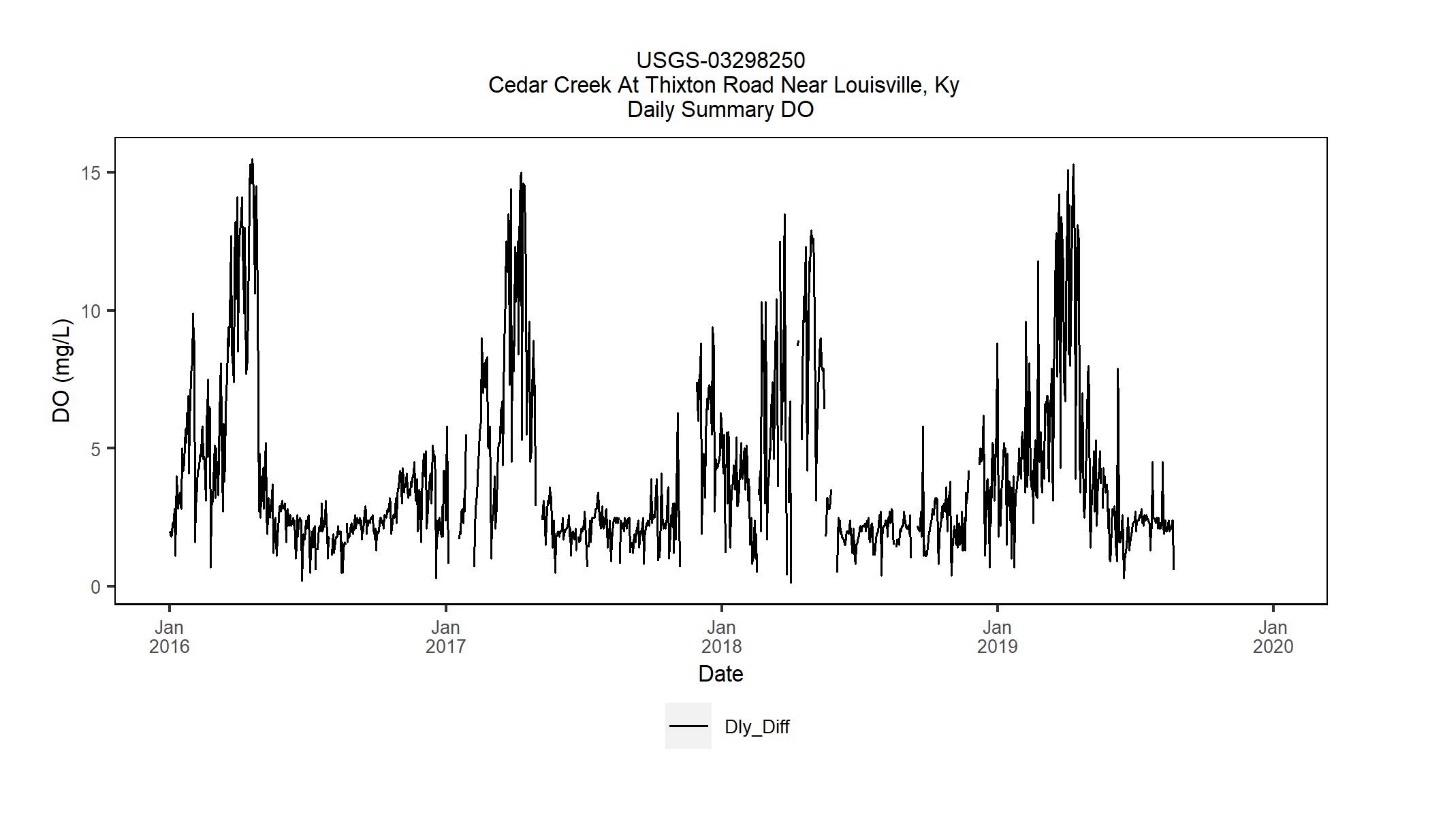
## Instantaneous water quality



The scripts will generate a separate time series plot for each parameter and gage with available data. Flow will be plotted as average daily flow and the water quality parameters will be plotted based on the ≤4 X’s a day instantaneous measurements (6am, noon, 6pm, and midnight, or the next closest record to those times without duplication). Days with missing data will be indicated by gaps. Files are named based on parameter and gage number.

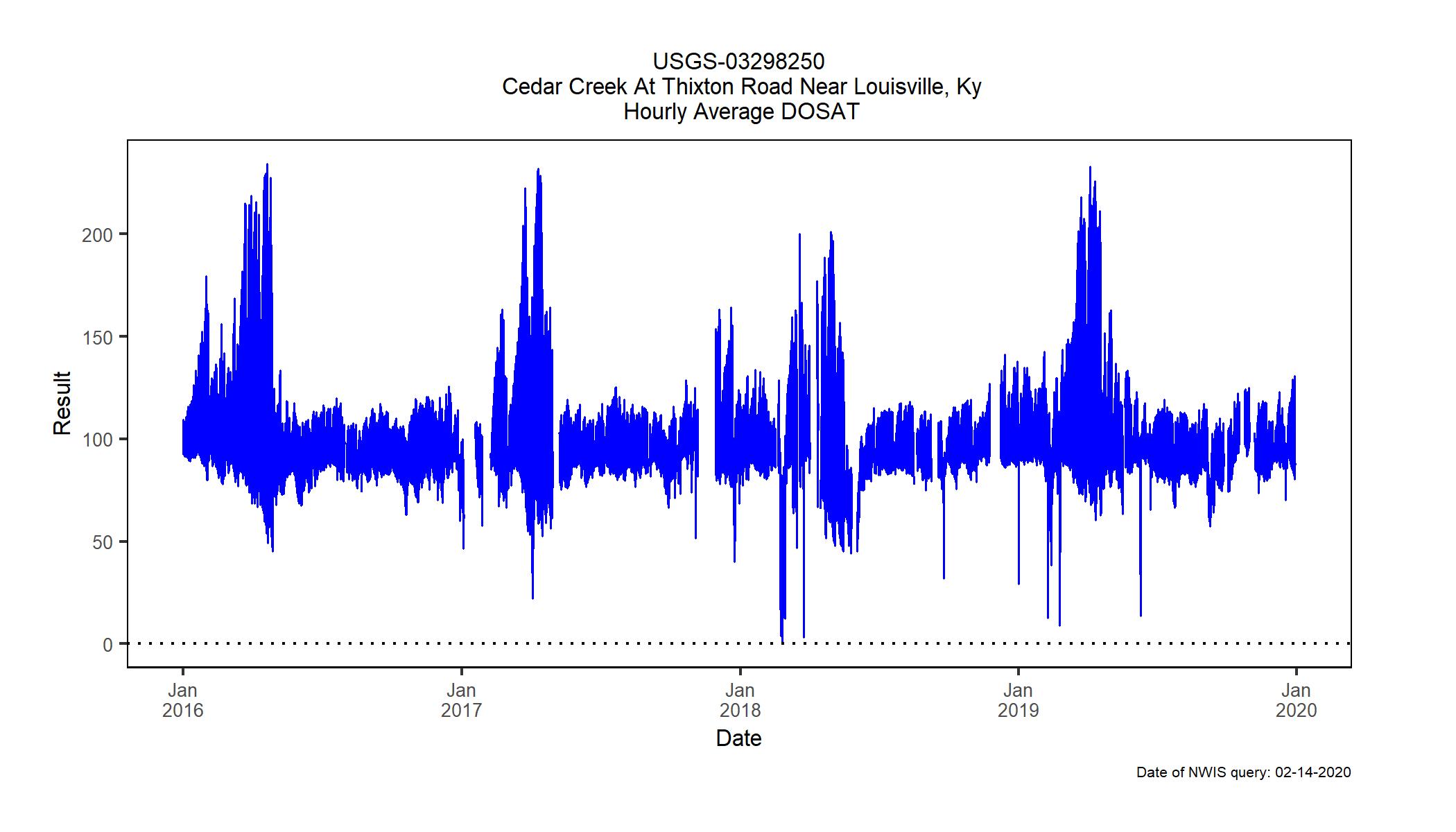
# Script output: Time series of daily statistics

## Daily difference time series



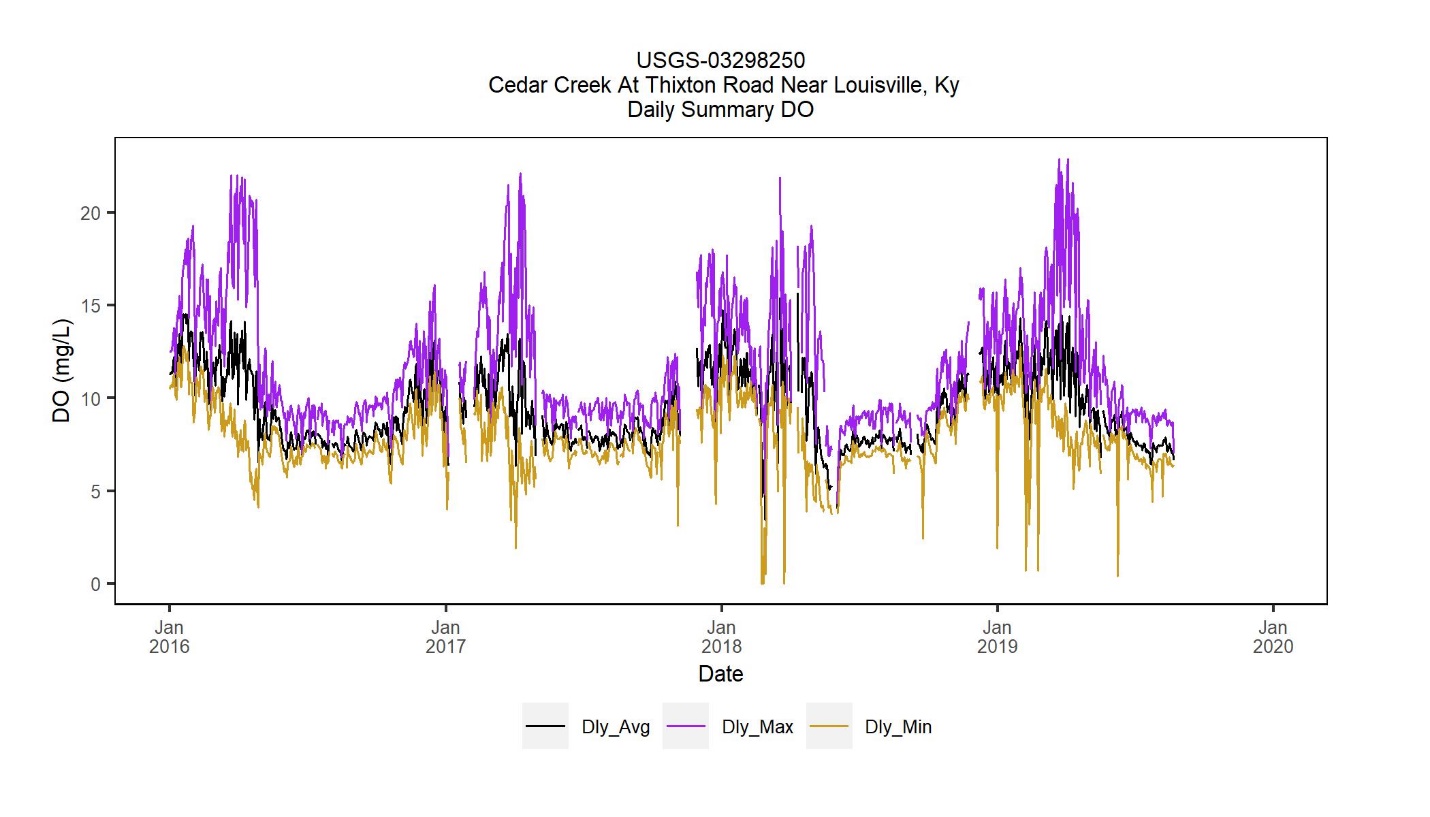
The plot above indicates the daily difference (Daily Max – Daily Min) based on the original raw data (not subset to 4 X’s a day). Days with missing data are represented by data gaps.

## Hourly average time series



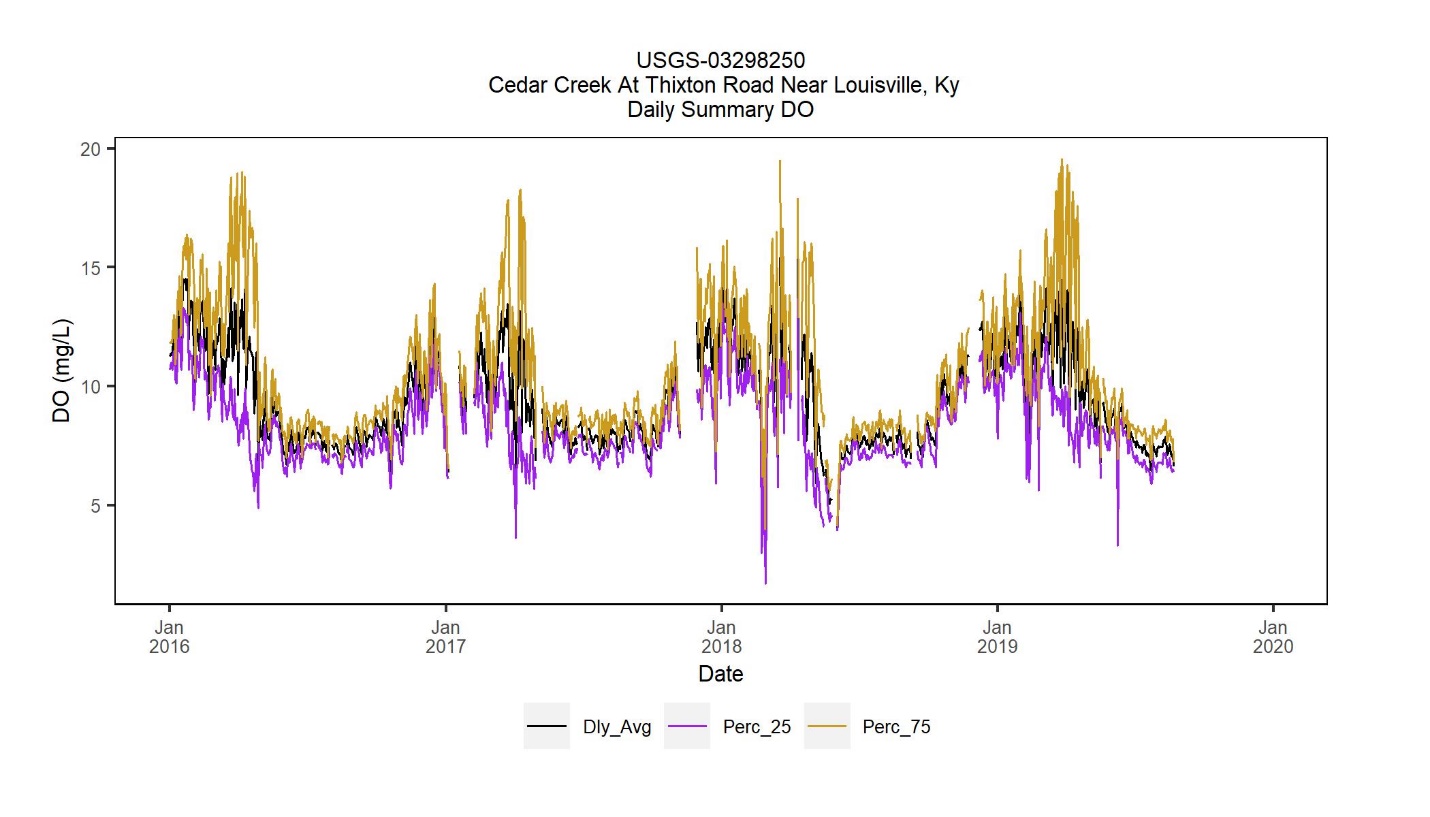
Based on the full dataset (i.e., not subset to 4 records per day), this plot indicates the hourly average for each parameter and gage. Days with missing data are represented by data gaps.

## Daily average/min/max time series



The plot above indicates the Daily Average, Daily Max, and Daily Min based on the original raw data (i.e., not subset to 4 records per day). Days with missing data are represented by data gaps.

## Daily average/percentiles time series



The plot above indicates the Daily Average, Daily 25th Percentile, and Daily 75th Percentile based on the original raw data (i.e., not subset to 4 records per day). Days with missing data are represented by data gaps. Additional daily percentiles are reported in the WQ\_Dly\_Stats\_XXXXXX.csv file.

# Literature Cited

De Cicco, L.A., Hirsch, R.M., Lorenz, D., Watkins, W.D., 2018, dataRetrieval: R packages for discovering and retrieving water data available from Federal hydrologic web services,<[doi:10.5066/P9X4L3GE](https://doi.org/10.5066/P9X4L3GE)>

Moulton, T. L. 2018. rMR: Importing data from loligo systems software, calculating metabolic rates and critical tensions.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Wickham, H. (2017). tidyverse: Easily install and load the 'Tidyverse'. https://CRAN.R-project.org/package=tidyverse