Part 1. Accessing daily streamflow data from Water Survey of Canada’s HYDAT Database

CSHS hydRology Team

March 2021

One of the key opportunities for users of the R-package **CSHShydRology** is that users can tap into the computer functions that are used by other professionals across Canada. Many of the functions have been developed by practitioners and researchers and are shared so that the entire community can perform analysis and interpretation based on the current state of the art.

The current version of **CSHShydRology** can be obtained from GitHub at <https://github.com/CSHS-CWRA/CSHShydRology>. It is planned to release the package through CRAN (<https://cran.r-project.org/>) later this year.

## How can you access Canada’s hydrometric data using R?

This article, which is written using **Rmarkdown** (cran.r-project.org/web/packages/rmarkdown/vignettes/rmarkdown.html), shows how to obtain daily flow data from the primary source the HYDAT database. This database is updated quarterly which and that creates opportunities and risks.

There are two main approaches to doing obtaining the data:.

1. Access using the r-package **tidyhydat**.
2. Access data using ECDataExplorer (ECDE) by saving data files to a local directory.

You can obtain ECDataExplorer at: <https://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/explorer.html>

One always needs to be careful when using data from these types of sources. Sometimes, changes are made at the source and code developed for one version could be impacted by those changes.

You need to be aware that the HYDAT database is periodically updated and if your work or project depends on the same data being available you may need to approach this differently. HYDAT is actually several linked databases and many parts change over time (quarterly) and provide a ‘snapshot’ for the current version.

There are many options, so it is important to know who you are and what fits your needs. This is simply one important form of fitness for purpose. Is it important that you have the same data each time for your project?

Note that all **CSHShydRology** functions start with “ch\_”

## tidyhydat

**tidyhydat** has wide functionality and there are vignettes describing various ways to access data. **tidyhydat** can also be used to access real time streamflow data. It is possible to search quite widely using station names, numbers, and provinces. I use **tidyhydat** to access data for sites where I have an interest in something that I consider ‘one-off’ or short-term and I do not feel the need to have a static dataset.

One of the many useful features of **tidyhydat** is you can retrieve daily flow data for a single station or a list of stations, as shown by these examples.

## hy\_daily\_flows is a tidyhydat function to retrieve daily flows

m\_thydat <- hy\_daily\_flows(station\_number =c("05CK004"))
str(m\_thydat)

## tibble [21,641 x 5] (S3: hy/tbl\_df/tbl/data.frame)
## $ STATION\_NUMBER: chr [1:21641] "05CK004" "05CK004" "05CK004" "05CK004" ...
## $ Date : Date[1:21641], format: "1960-10-01" "1960-10-02" ...
## $ Parameter : chr [1:21641] "Flow" "Flow" "Flow" "Flow" ...
## $ Value : num [1:21641] 26.8 26.4 25.7 25.3 25.3 ...
## $ Symbol : chr [1:21641] NA NA NA NA ...
## - attr(\*, "missed\_stns")= chr(0)

stations <- c("05CK004","05CK003","05BB001")

m\_thydat\_list <- hy\_daily\_flows(station\_number=stations)
str(m\_thydat\_list)

## tibble [61,699 x 5] (S3: hy/tbl\_df/tbl/data.frame)
## $ STATION\_NUMBER: chr [1:61699] "05BB001" "05BB001" "05BB001" "05BB001" ...
## $ Date : Date[1:61699], format: "1909-05-01" "1909-05-02" ...
## $ Parameter : chr [1:61699] "Flow" "Flow" "Flow" "Flow" ...
## $ Value : num [1:61699] NA NA NA NA NA NA NA NA NA NA ...
## $ Symbol : chr [1:61699] NA NA NA NA ...
## - attr(\*, "missed\_stns")= chr(0)

What **tidyhydat** returns is a tibble. Tibbles are a modern take on dataframes and have some useful features, but since they have a different structure to a dataframe they need a bit of tweaking for use in some of the function in **CSHShydRology**.
**CSHShydRology**\_ has a function that uses **tidyhydat** to get the flow data but returns the data in the same format as from ECDE. If there are multiple stations, values are returned in a list of dataframes.

There are several changes made in the data structure. The variable names are changed as are some contents. In ECDE, PARAM value 1 is Flow and 2 is Level; in **tidyhydat** these are explicitly named. In ECDE, the data quality variable SYM can be "“,”A“,”B“,”D“, or”E" in **tidyhydat** these are changed to NA, “A”, “B”, “D”, or “E”.

The function ch\_tidyhydat\_ECDE makes the resulting dataframe from **tidyhydat** match those from ECDE.

th\_ECDE <-ch\_tidyhydat\_ECDE(m\_thydat)
str(th\_ECDE)

## 'data.frame': 21641 obs. of 5 variables:
## $ ID : chr "05CK004" "05CK004" "05CK004" "05CK004" ...
## $ PARAM: int 1 1 1 1 1 1 1 1 1 1 ...
## $ Date : Date, format: "1960-10-01" "1960-10-02" ...
## $ Flow : num 26.8 26.4 25.7 25.3 25.3 ...
## $ SYM : chr "" "" "" "" ...

mdata <- hy\_daily\_flows(station\_number=c("05CK004","08MF005","05BB001"))
m\_list <- ch\_tidyhydat\_ECDE(mdata)

## [1] "Original tibble contained 3 stations. A list of dataframes is returned"

 str(m\_list[[1]])

## 'data.frame': 39936 obs. of 5 variables:
## $ ID : chr "05BB001" "05BB001" "05BB001" "05BB001" ...
## $ PARAM: int 1 1 1 1 1 1 1 1 1 1 ...
## $ Date : Date, format: "1909-05-01" "1909-05-02" ...
## $ Flow : num NA NA NA NA NA NA NA NA NA NA ...
## $ SYM : chr "" "" "" "" ...

 str(m\_list[[2]])

## 'data.frame': 21641 obs. of 5 variables:
## $ ID : chr "05CK004" "05CK004" "05CK004" "05CK004" ...
## $ PARAM: int 1 1 1 1 1 1 1 1 1 1 ...
## $ Date : Date, format: "1960-10-01" "1960-10-02" ...
## $ Flow : num 26.8 26.4 25.7 25.3 25.3 ...
## $ SYM : chr "" "" "" "" ...

 str(m\_list[[3]])

## 'data.frame': 39387 obs. of 5 variables:
## $ ID : chr "08MF005" "08MF005" "08MF005" "08MF005" ...
## $ PARAM: int 1 1 1 1 1 1 1 1 1 1 ...
## $ Date : Date, format: "1912-03-01" "1912-03-02" ...
## $ Flow : num 538 538 538 538 538 538 479 479 459 459 ...
## $ SYM : chr "" "" "" "" ...

#'#note the order of the list is in increasing alphabetical station\_number order

 t04CK004 <-m\_list[[2]]

ECDataExplorer provides much of the same access as **tidyhydat** outside of the R environment. It too is subject to the quarterly updates but it has a graphical user interface and lets the user explore the database in many ways. As a researcher I use it to select stations that meet the criteria of my project. I can save the “tombstone” data (i.e. station metadata) for all the stations that I select, and I can also bulk download the hydrometric data in separate files. This is more cumbersome that doing it in **tidyhydat**, but since repeatability and reproducibility are important to my work I prefer having the station data files and the tombstone information for the day I made the selection. That way, I have files that are ‘fixed’ in time. I store these in a project directory and these are never changed.

In the following I access a single station file with daily flow data from ECDE from the working directory.

e05CK004<- ch\_read\_ECDE\_flows("05CK004\_Daily\_Flow\_ts.csv")

str(e05CK004)

## 'data.frame': 21275 obs. of 5 variables:
## $ ID : chr "05CK004" "05CK004" "05CK004" "05CK004" ...
## $ PARAM: int 1 1 1 1 1 1 1 1 1 1 ...
## $ Date : Date, format: "1960-10-01" "1960-10-02" ...
## $ Flow : num 26.8 26.4 25.7 25.3 25.3 ...
## $ SYM : chr "" "" "" "" ...

So, the data from **tidyhydat** [21641] is longer than from ECDE [21275] because they originated in different versions of HYDAT.

**CSHShydRology** has many functions that help us work with these data however you choose to access them.

In Part 2, we will show how to access the station metadata that we might want to include in plots.

P.H.Whitfield