Packaging your research problem: the case of *CityWaterBalance* for R

Laura Erban

EPA R User Group Workshop

August 12, 2019

Introduction

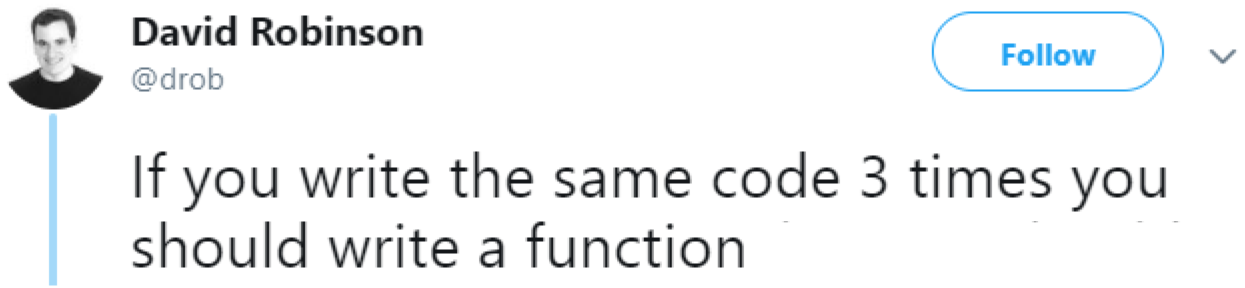
Speaker: hydrogeologist, earth systems scientist, postdoc

Location: ORD/NHEERL/AED (Narragansett, RI)

AspiRation: saving my future self

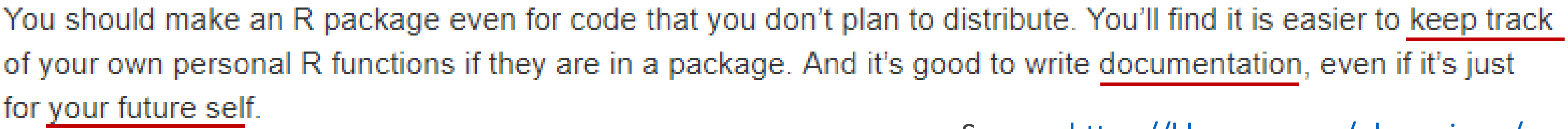
Disclaimer: all opinions are my own

Heuristics



David Robinson
@drob

If you write the same code 3 times you should write a function



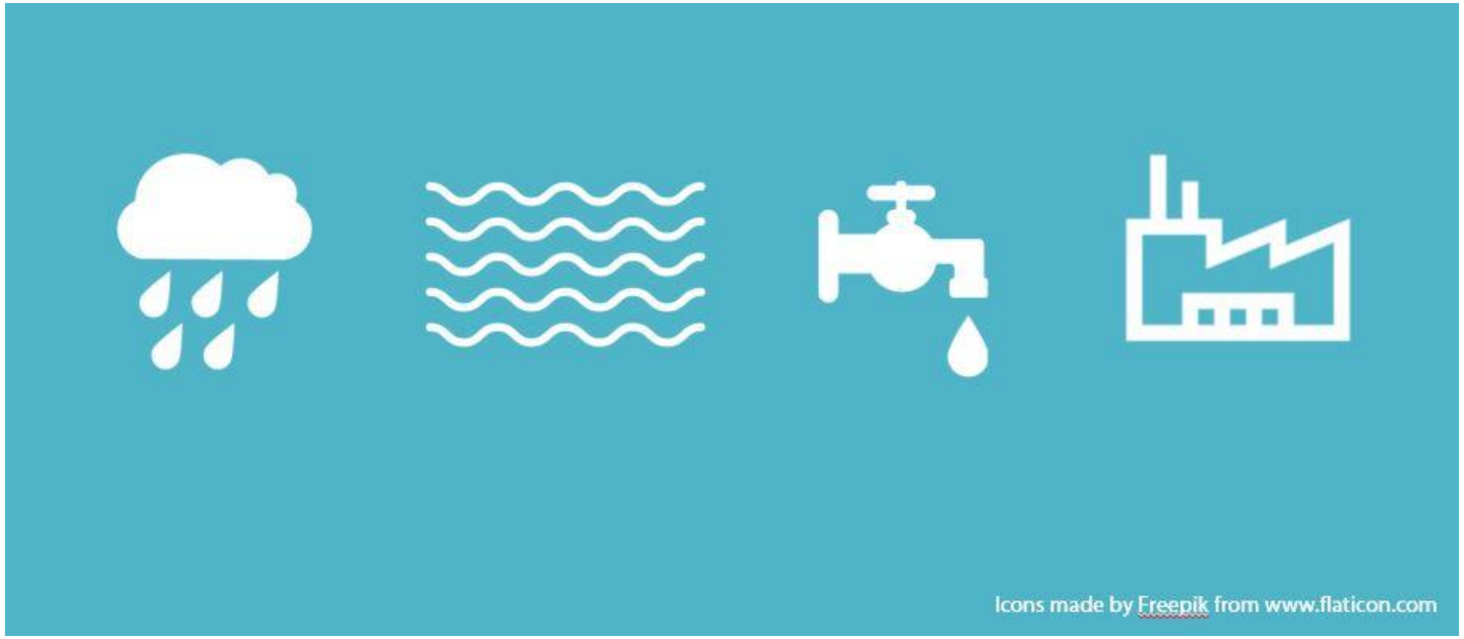
You should make an R package even for code that you don't plan to distribute. You'll find it is easier to keep track of your own personal R functions if they are in a package. And it's good to write documentation, even if it's just for your future self.

Source: https://kbroman.org/pkg_primer/

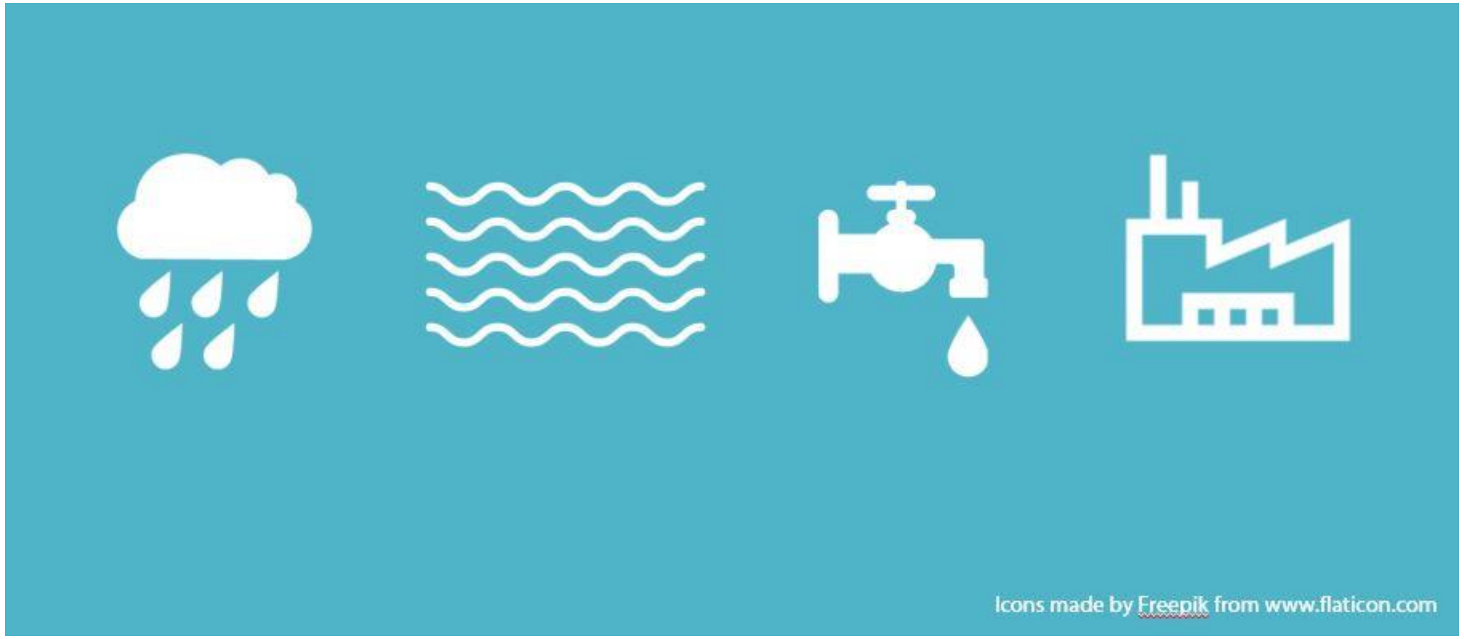
Research problems

- complexity (of subject, analysis, collaborators)
- reproducibility
- quantity / task-switching

How much water flows through a city?



How much water flows through a city?



Icons made by Freepik from www.flaticon.com

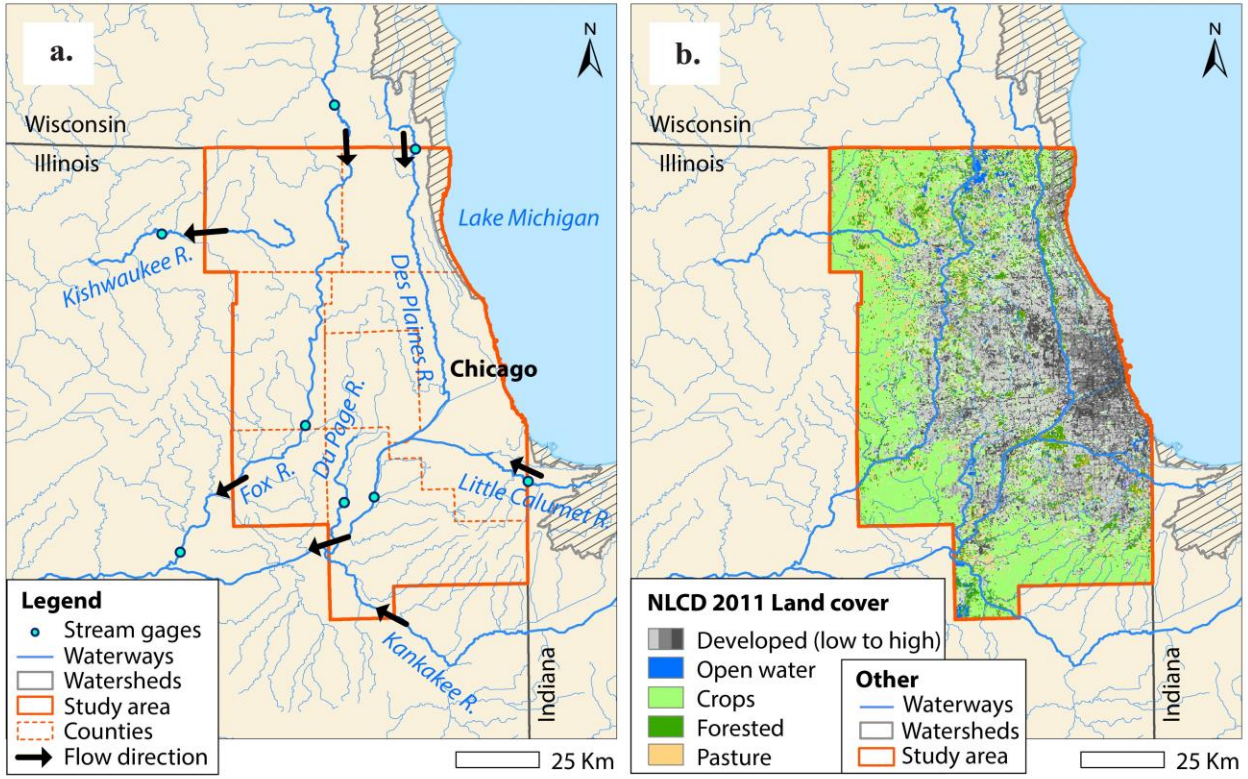
- diverse sources / uses
- internal recycling / reuse
- scattered data
- unmeasured flows

Water flows in greater Chicago*

*seven counties under regional planning agency: Chicago Metropolitan Agency for Planning (CMAP)

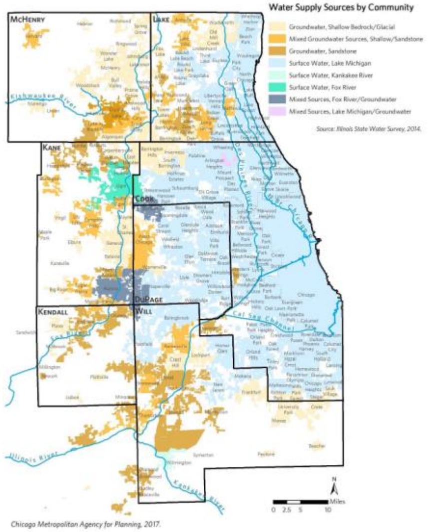
Study area

Developed land



Erban et al., 2018 *Open Water Journal*

Lake Michigan users



Water Supply Sources by Community
Groundwater, Shallow Bedrock/Glacial
Mixed Groundwater Sources, Shallow/Sandstone
Groundwater, Sandstone
Surface Water, Lake Michigan
Surface Water, Kankakee River
Surface Water, Fox River
Mixed Sources, Fox River/Groundwater
Mixed Sources, Lake Michigan/Groundwater
Source: Illinois State Water Survey, 2014.
McHENRY
LAKE
KANE
COOK
DUPAGE
KENDALL
WILL
Chicago Metropolitan Agency for Planning, 2017.

CMAP, 2017 *On to 2050: Water Resources*

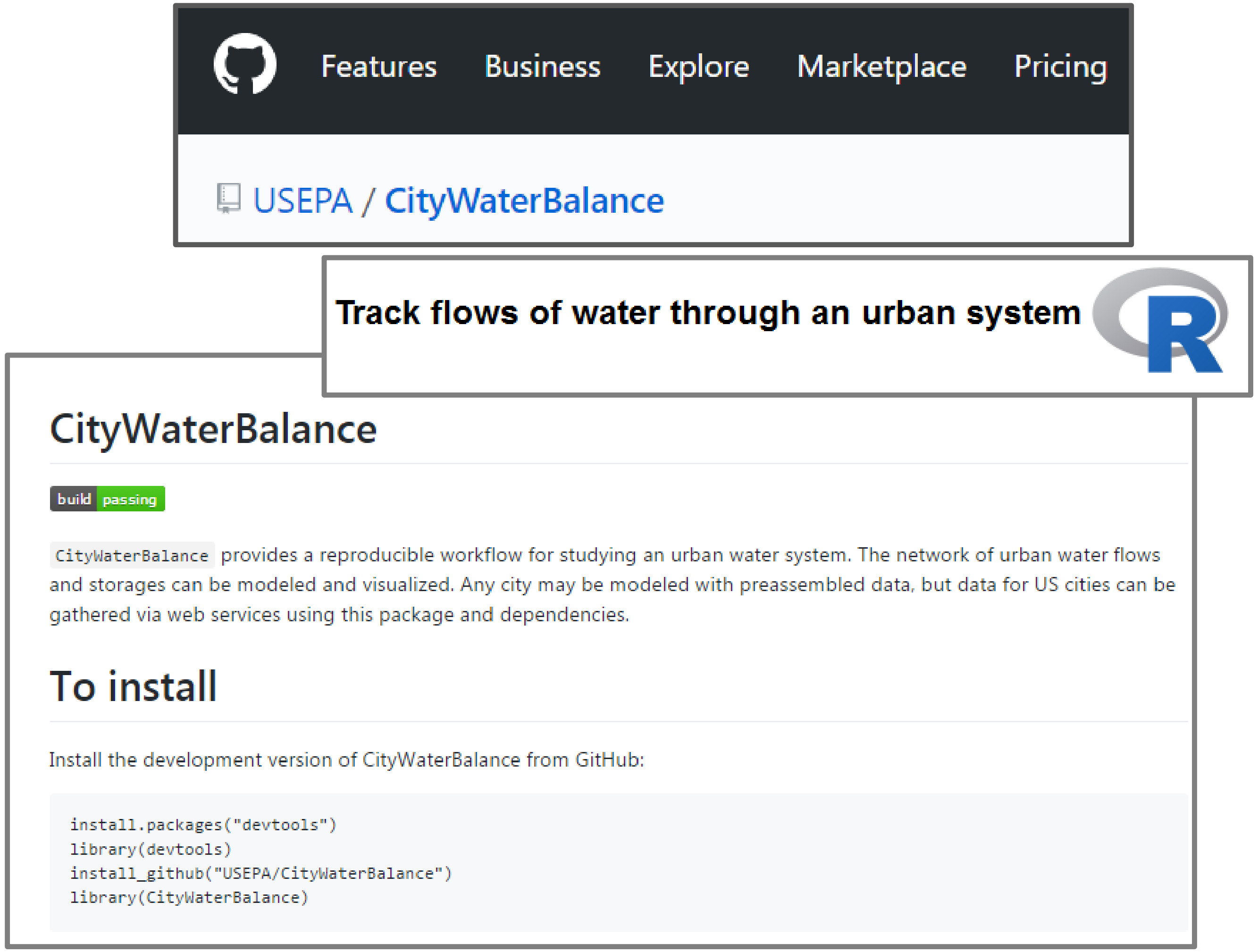
Packaging the workflow

CityWaterBalance for R

(available on CRAN and USEPA GitHub)

Design principles:

- open source, open access
- generalizable to any urban area
- automate data retrieval
- estimate unmeasured flows



Track flows of water through an urban system

CityWaterBalance

`CityWaterBalance` provides a reproducible workflow for studying an urban water system. The network of urban water flows and storages can be modeled and visualized. Any city may be modeled with preassembled data, but data for US cities can be gathered via web services using this package and dependencies.

To install

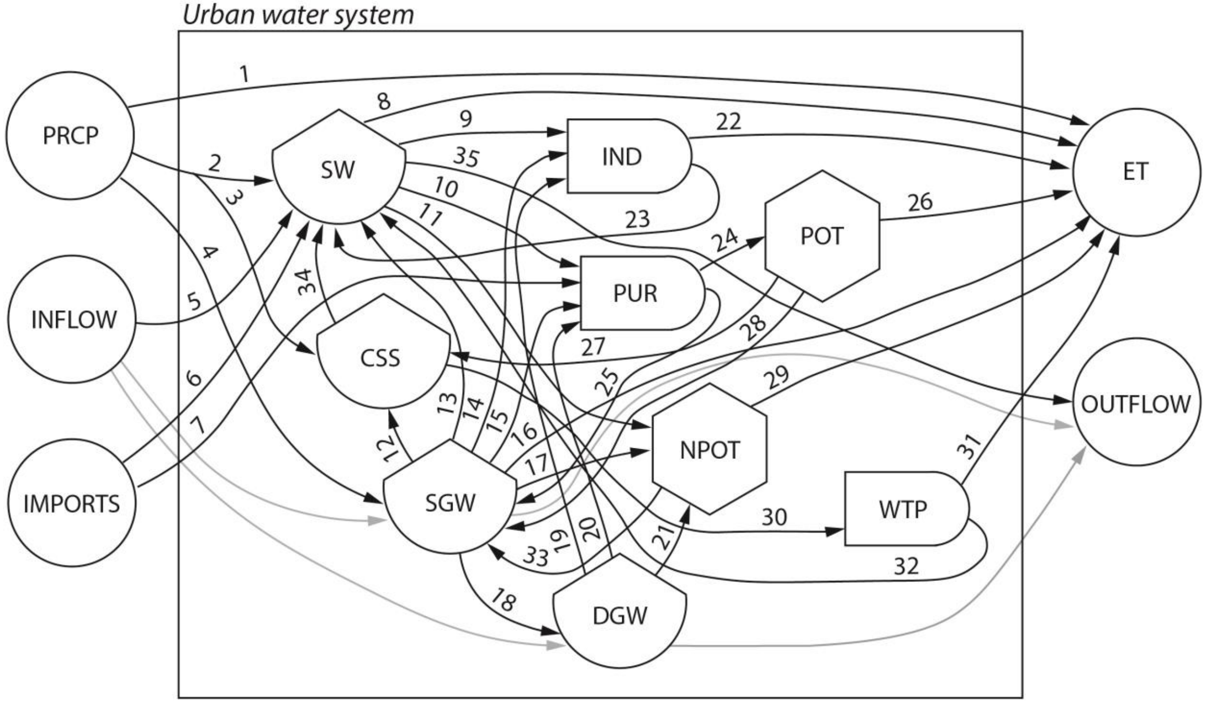
Install the development version of CityWaterBalance from GitHub:

```
install.packages("devtools")
library(devtools)
install_github("USEPA/CityWaterBalance")
library(CityWaterBalance)
```

Inside *CityWaterBalance*

Networked model:

- accounts for inflows, outflows, changes in storage
- based on visual math (right), coded in the core package function



Key

○ Global flows: precipitation (PRCP), streamflow in (INFLOW), imports (IMPORTS), evapotranspiration (ET), streamflow out (OUTFLOW)

⬠ Storages: surface water (SW), combined sewer system (CSS), shallow groundwater (SGW), deep groundwater (DGW)

▭ Producers: purification plants (PUR), industrial facilities (IND), wastewater treatment plants (WTP)

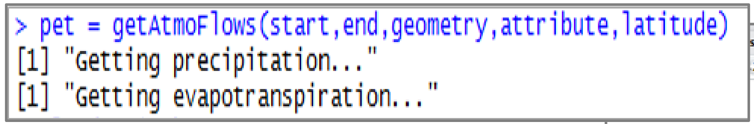
⬡ Consumers: potable use (POT), non-potable use (NPOT)

Erban et al., 2018 *Open Water Journal*

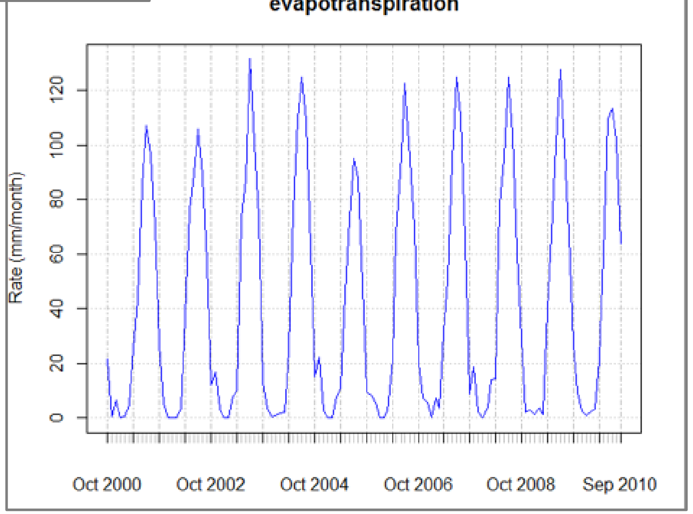
Inside *CityWaterBalance*

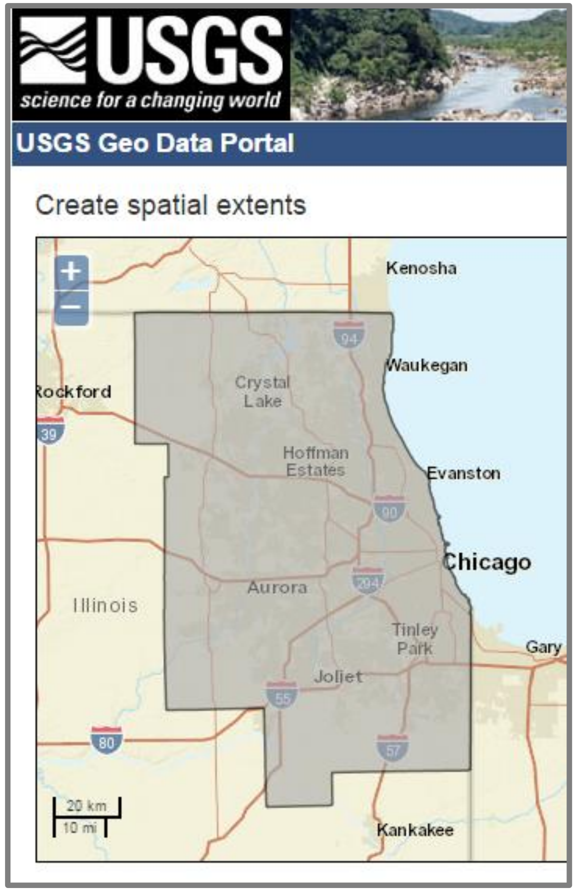
Usage:

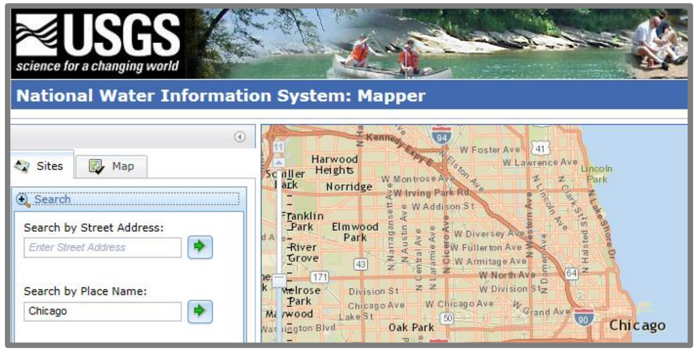
- functions retrieve data from web services
- help from USGS R packages *geoknife* and *dataRetrieval*



```
> pet = getAtmoFlows(start,end,geometry,attribute,latitude)
[1] "Getting precipitation..."
[1] "Getting evapotranspiration..."
```





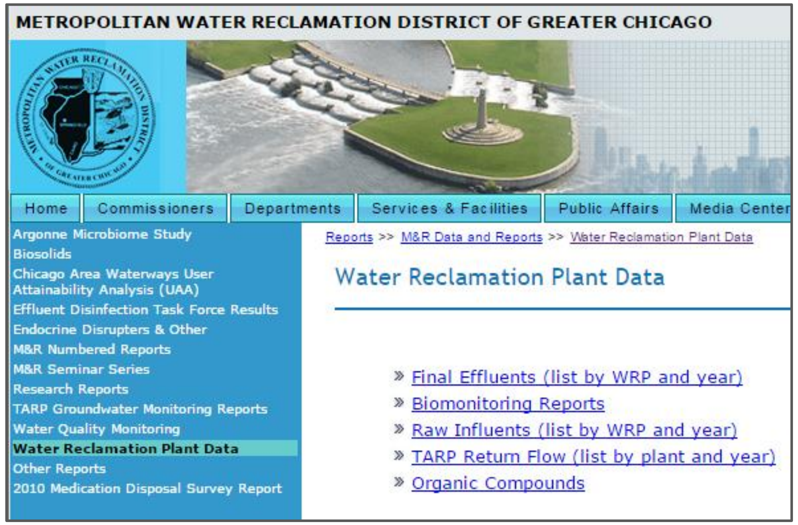


Inside *CityWaterBalance*

Usage:

- add data that is not federated, or not served online

- package as placeholder for future accessible data



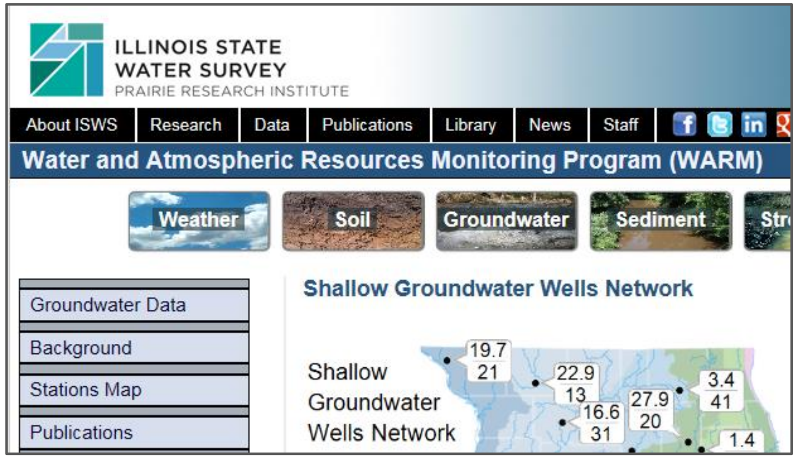
METROPOLITAN WATER RECLAMATION DISTRICT OF GREATER CHICAGO

Argonne Microbiome Study
Biosolids
Chicago Area Waterways User Attainability Analysis (UAA)
Effluent Disinfection Task Force Results
Endocrine Disrupters & Other
M&R Numbered Reports
M&R Seminar Series
Research Reports
TARP Groundwater Monitoring Reports
Water Quality Monitoring
Water Reclamation Plant Data
Other Reports
2010 Medication Disposal Survey Report

Reports >> M&R Data and Reports >> Water Reclamation Plant Data

Water Reclamation Plant Data

» Final Effluents (list by WRP and year)
» Biomonitoring Reports
» Raw Influents (list by WRP and year)
» TARP Return Flow (list by plant and year)
» Organic Compounds



ILLINOIS STATE WATER SURVEY
PRAIRIE RESEARCH INSTITUTE

Water and Atmospheric Resources Monitoring Program (WARM)

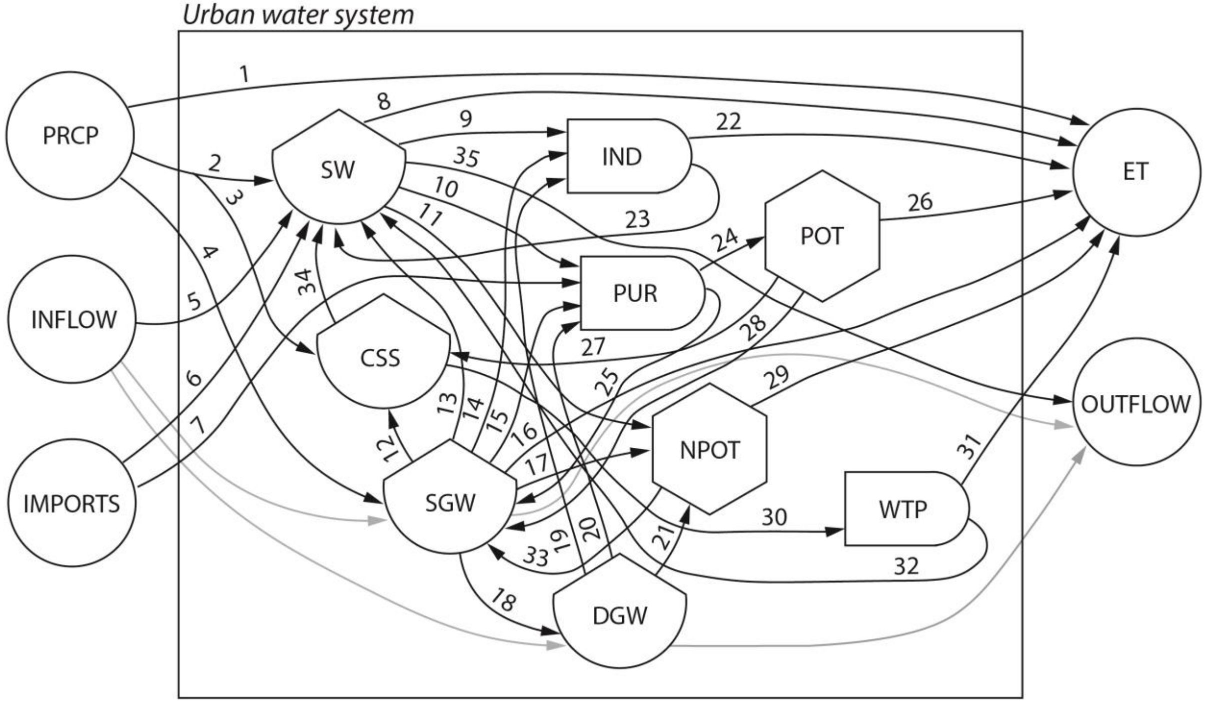
Groundwater Data
Background
Stations Map
Publications

Shallow Groundwater Wells Network

Inside *CityWaterBalance*

Usage:

- estimate unmeasured flows by mass balance



Key

○ Global flows: precipitation (PRCP), streamflow in (INFLOW), imports (IMPORTS), evapotranspiration (ET), streamflow out (OUTFLOW)

⬠ Storages: surface water (SW), combined sewer system (CSS), shallow groundwater (SGW), deep groundwater (DGW)

▭ Producers: purification plants (PUR), industrial facilities (IND), wastewater treatment plants (WTP)

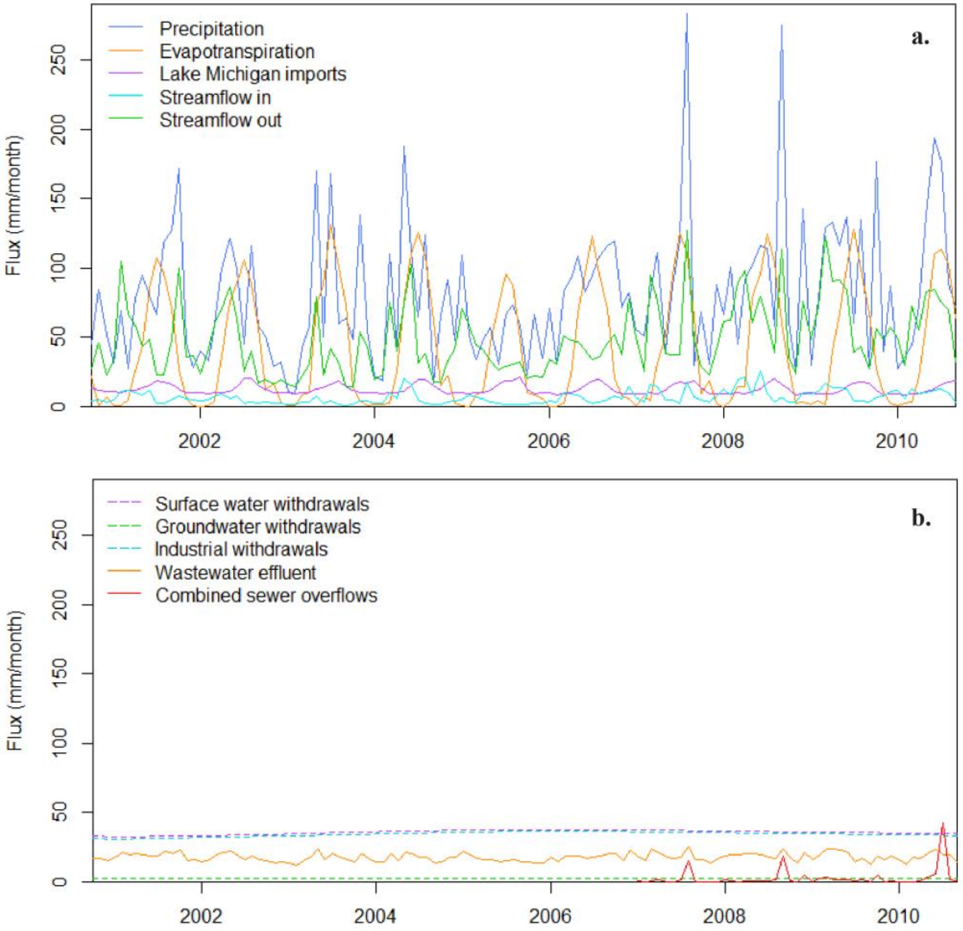
⬡ Consumers: potable use (POT), non-potable use (NPOT)

dS/dt?

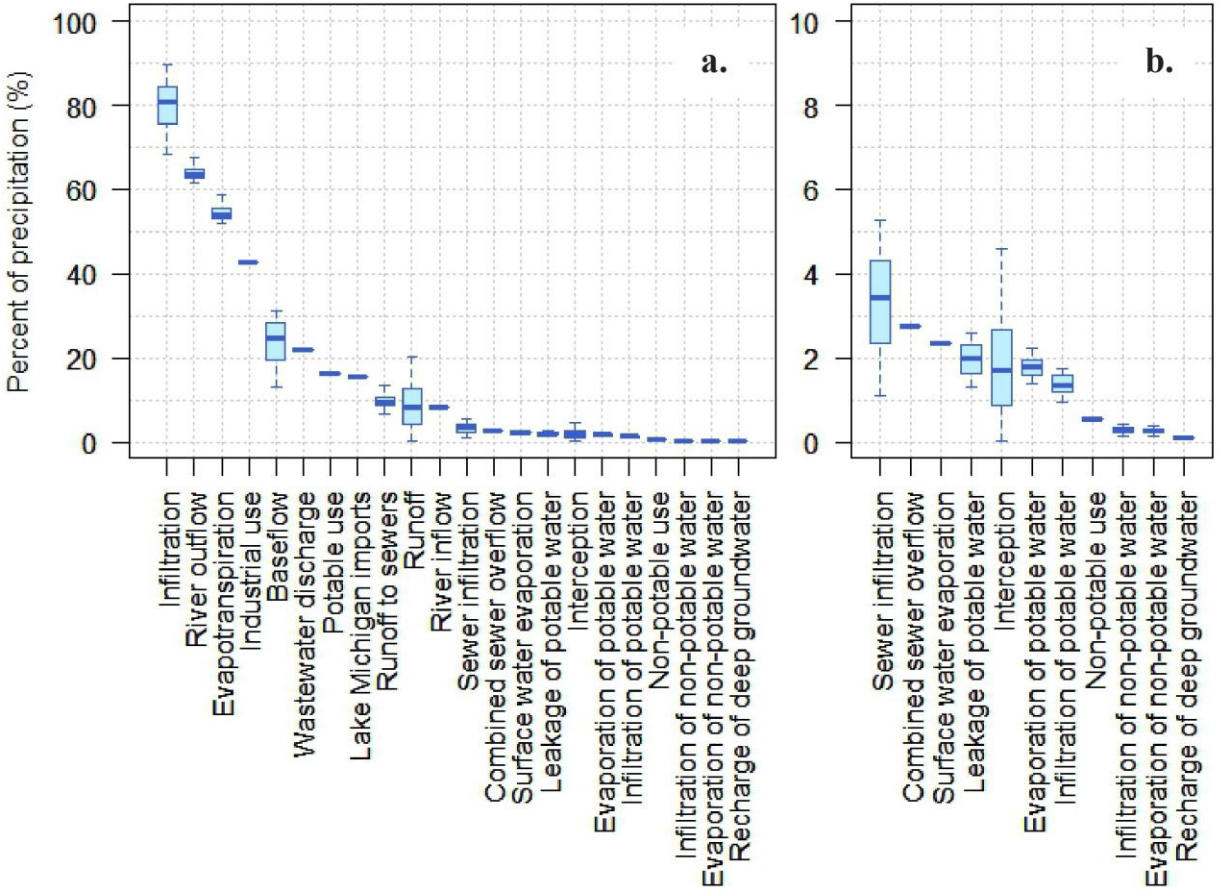
Erban et al., 2018 *Open Water Journal*

CityWaterBalance output, from two perspectives

1) time

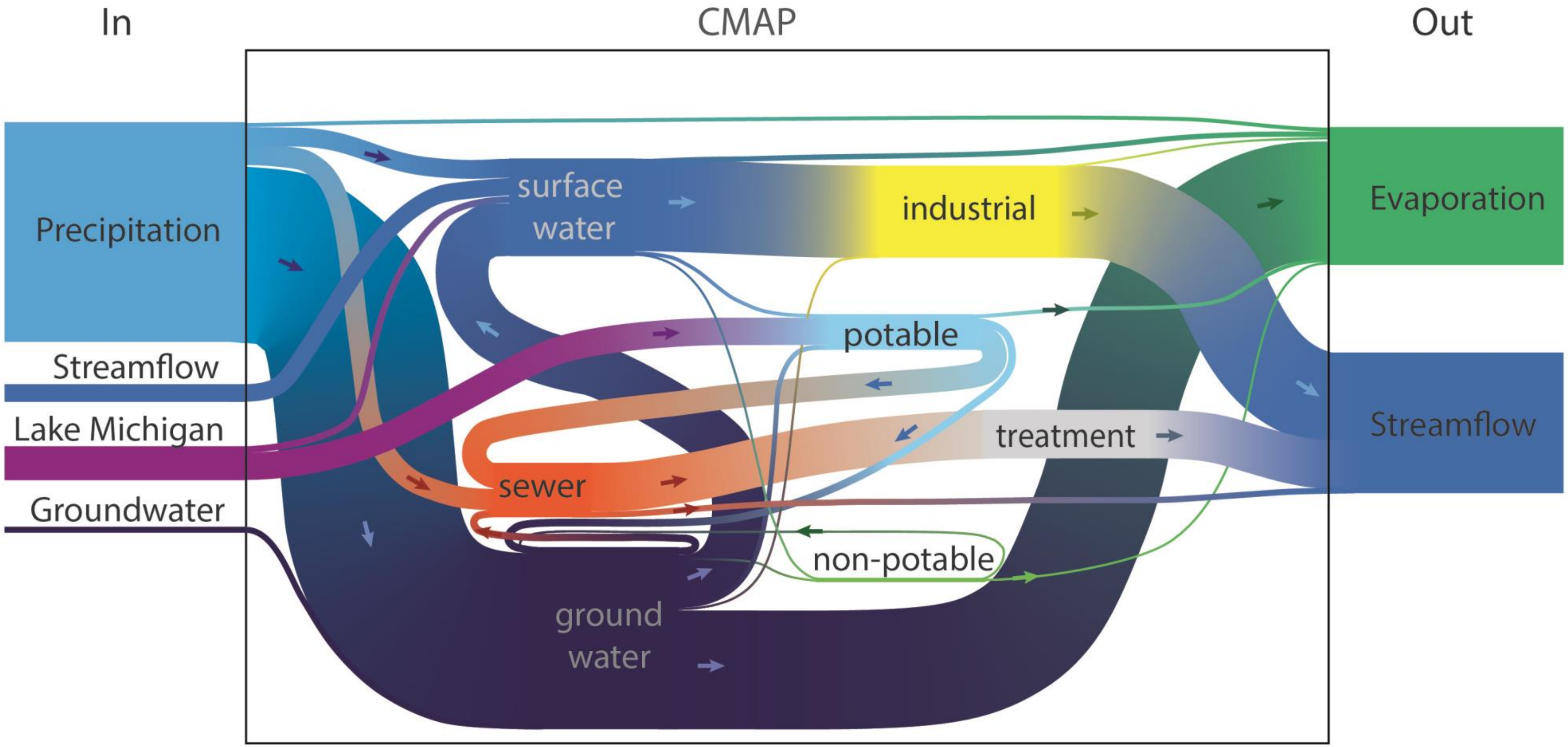


2) uncertainty



Erban et al., 2018 *Open Water Journal*

A quantitative portrait



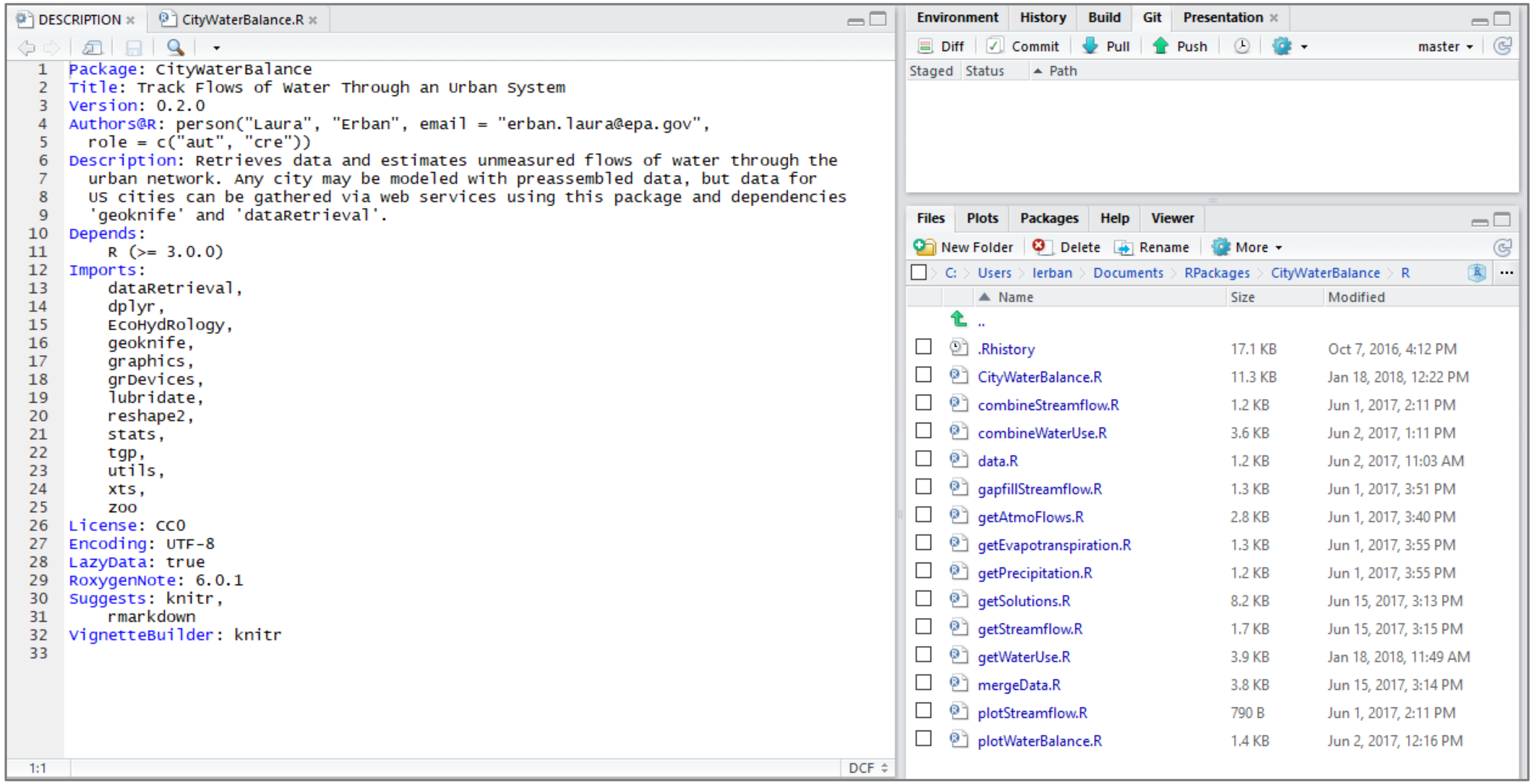
Average annual flows: water years 2001-2010

Erban et al., 2018 *Open Water Journal*

Packaging your research problem: generalizable lessons

Writing a package can help you..

1. organize your functions



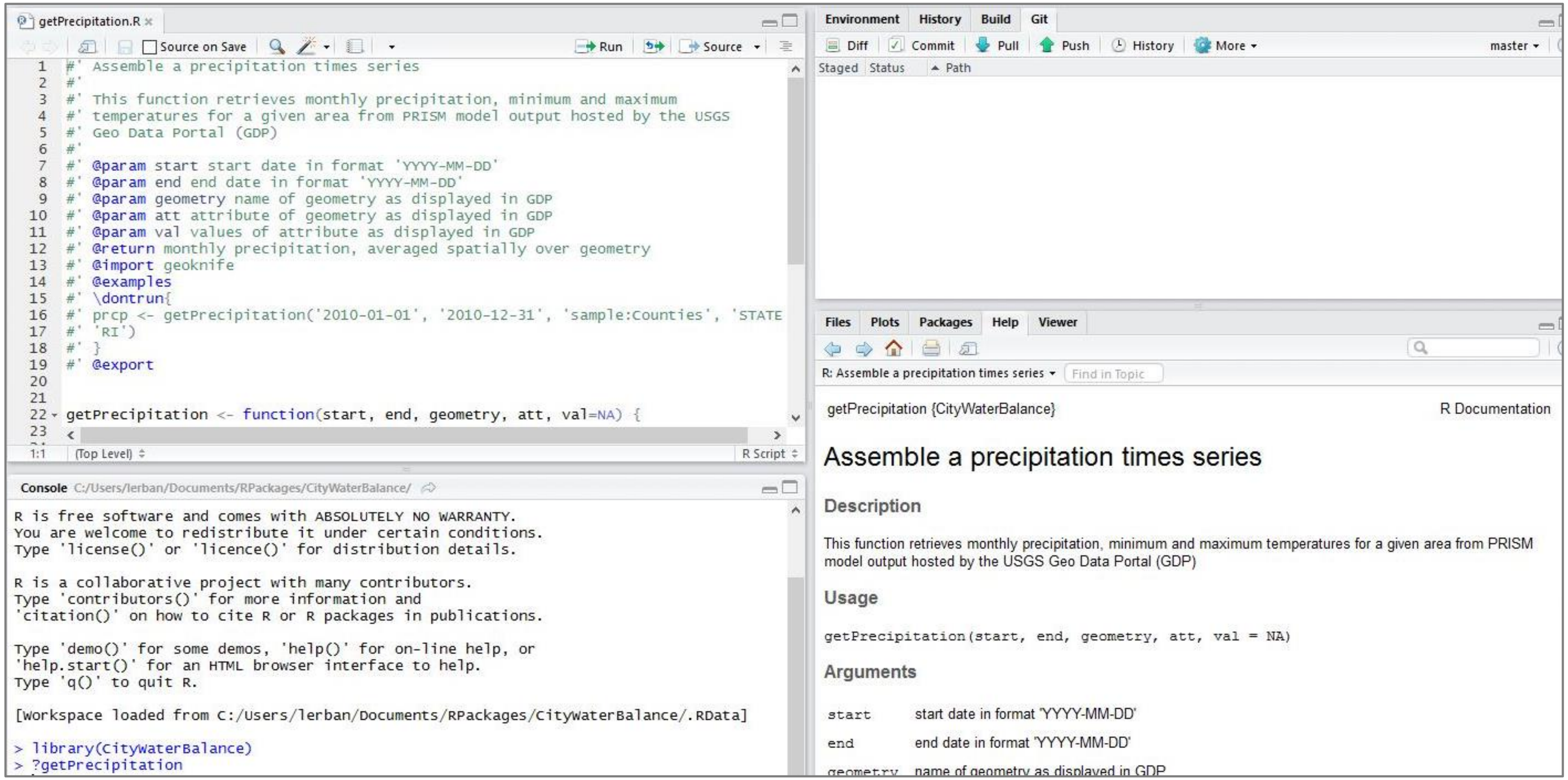
```
Package: CityWaterBalance
Title: Track Flows of Water Through an Urban System
Version: 0.2.0
Authors@R: person("Laura", "Erban", email = "erban.laura@epa.gov",
  role = c("aut", "cre"))
Description: Retrieves data and estimates unmeasured flows of water through the
  urban network. Any city may be modeled with preassembled data, but data for
  US cities can be gathered via web services using this package and dependencies
  'geoknife' and 'dataRetrieval'.
Depends:
    R (>= 3.0.0)
Imports:
    dataRetrieval,
    dplyr,
    EcoHydRology,
    geoknife,
    graphics,
    grDevices,
    lubridate,
    reshape2,
    stats,
    tgp,
    utils,
    xts,
    zoo
License: CC0
Encoding: UTF-8
LazyData: true
RoxygenNote: 6.0.1
Suggests: knitr,
    rmarkdown
VignetteBuilder: knitr

```

Name	Size	Modified
..		
.Rhistory	17.1 KB	Oct 7, 2016, 4:12 PM
CityWaterBalance.R	11.3 KB	Jan 18, 2018, 12:22 PM
combineStreamflow.R	1.2 KB	Jun 1, 2017, 2:11 PM
combineWaterUse.R	3.6 KB	Jun 2, 2017, 1:11 PM
data.R	1.2 KB	Jun 2, 2017, 11:03 AM
gapfillStreamflow.R	1.3 KB	Jun 1, 2017, 3:51 PM
getAtmoFlows.R	2.8 KB	Jun 1, 2017, 3:40 PM
getEvapotranspiration.R	1.3 KB	Jun 1, 2017, 3:55 PM
getPrecipitation.R	1.2 KB	Jun 1, 2017, 3:55 PM
getSolutions.R	8.2 KB	Jun 15, 2017, 3:13 PM
getStreamflow.R	1.7 KB	Jun 15, 2017, 3:15 PM
getWaterUse.R	3.9 KB	Jan 18, 2018, 11:49 AM
mergeData.R	3.8 KB	Jun 15, 2017, 3:14 PM
plotStreamflow.R	790 B	Jun 1, 2017, 2:11 PM
plotWaterBalance.R	1.4 KB	Jun 2, 2017, 12:16 PM

Writing a package can help you..

2. document your functions in a common format



```
#' Assemble a precipitation times series
#'
#' This function retrieves monthly precipitation, minimum and maximum
#' temperatures for a given area from PRISM model output hosted by the USGS
#' Geo Data Portal (GDP)
#'
#' @param start start date in format 'YYYY-MM-DD'
#' @param end end date in format 'YYYY-MM-DD'
#' @param geometry name of geometry as displayed in GDP
#' @param att attribute of geometry as displayed in GDP
#' @param val values of attribute as displayed in GDP
#' @return monthly precipitation, averaged spatially over geometry
#' @import geoknife
#' @examples
#' \dontrun{
#' prcp <- getPrecipitation('2010-01-01', '2010-12-31', 'sample:Counties', 'STATE
#' 'RI')
#' }
#' @export


getPrecipitation <- function(start, end, geometry, att, val=NA) {
```

```
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.

[workspace loaded from C:/Users/lerban/Documents/RPackages/CityWaterBalance/.RData]

> library(CityWaterBalance)
> ?getPrecipitation
```

getPrecipitation {CityWaterBalance} R Documentation

Assemble a precipitation times series

Description

This function retrieves monthly precipitation, minimum and maximum temperatures for a given area from PRISM model output hosted by the USGS Geo Data Portal (GDP)

Usage

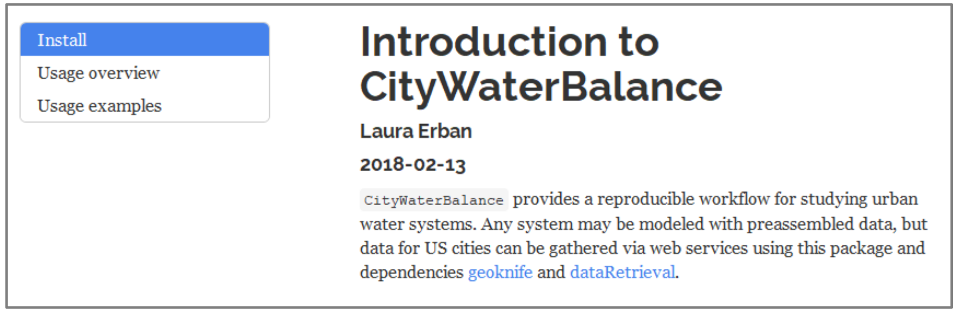
```
getPrecipitation(start, end, geometry, att, val = NA)
```

Arguments

start	start date in format 'YYYY-MM-DD'
end	end date in format 'YYYY-MM-DD'
geometry	name of geometry as displayed in GDP

Writing a package can help you..

3. share your work

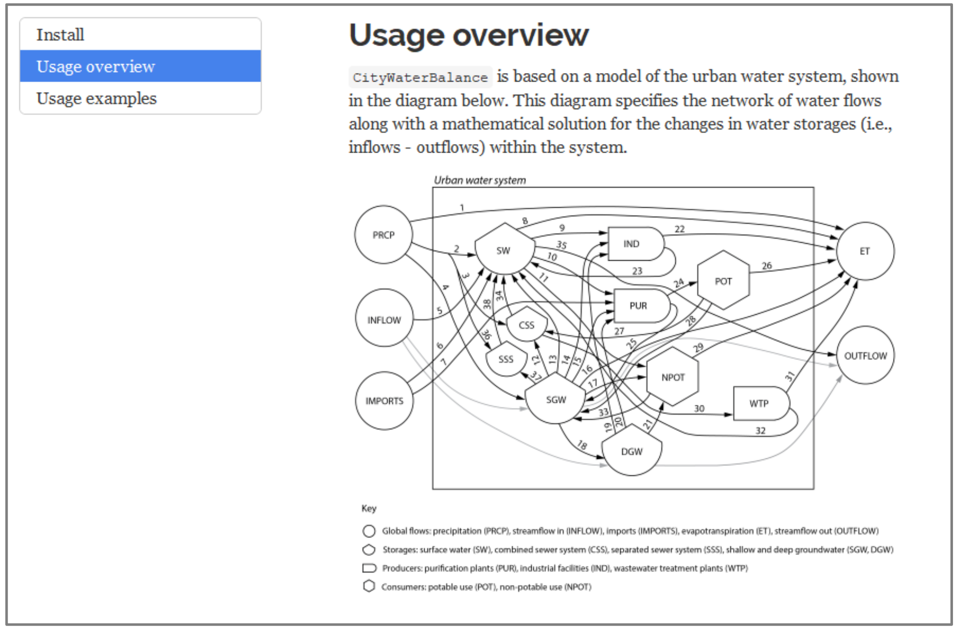


Introduction to CityWaterBalance

Laura Erban

2018-02-13

`CityWaterBalance` provides a reproducible workflow for studying urban water systems. Any system may be modeled with preassembled data, but data for US cities can be gathered via web services using this package and dependencies geoknife and dataRetrieval.



Usage overview

`CityWaterBalance` is based on a model of the urban water system, shown in the diagram below. This diagram specifies the network of water flows along with a mathematical solution for the changes in water storages (i.e., inflows - outflows) within the system.

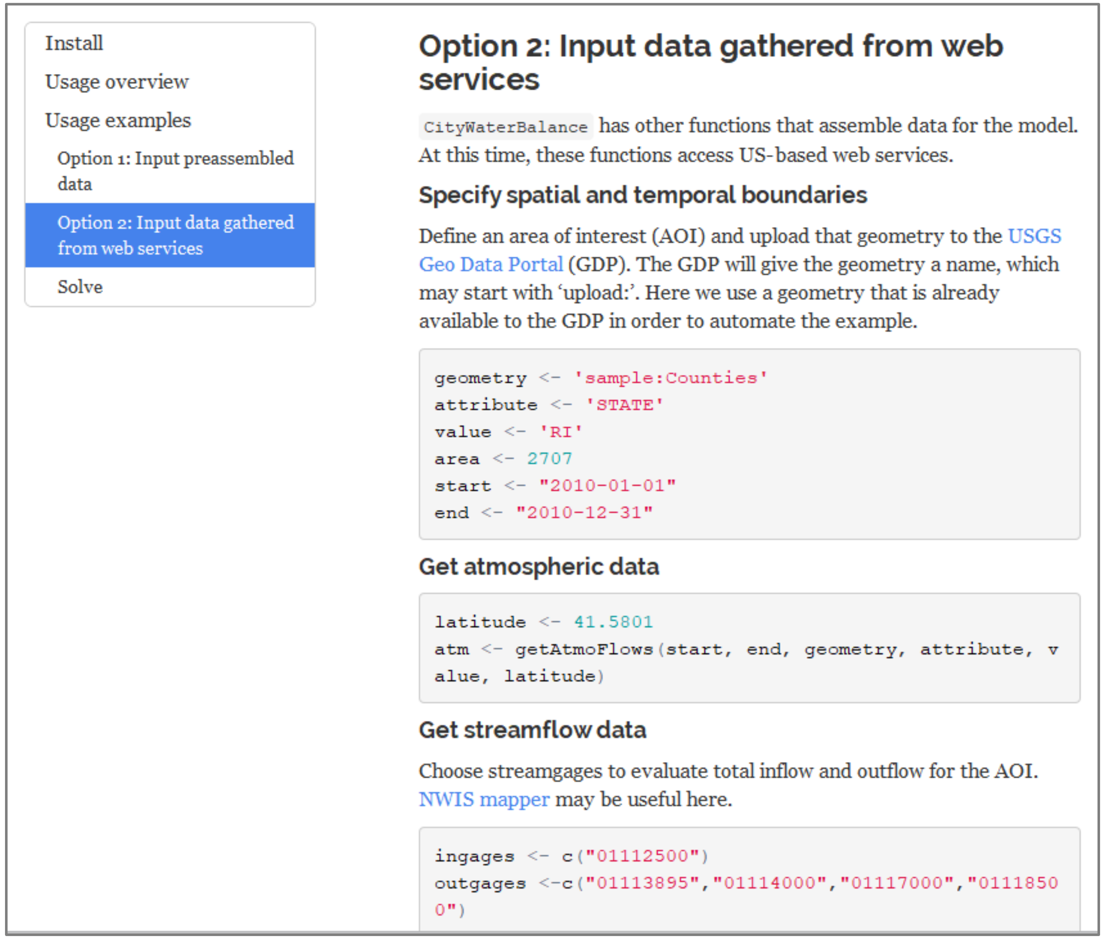
Key

○ Global flows: precipitation (PRCP), streamflow in (INFLOW), imports (IMPORTS), evapotranspiration (ET), streamflow out (OUTFLOW)

⬠ Storages: surface water (SW), combined sewer system (CSS), separated sewer system (SSS), shallow and deep groundwater (SGW, DGW)

▭ Producers: purification plants (PUR), industrial facilities (IND), wastewater treatment plants (WTP)

⬡ Consumers: potable use (POT), non-potable use (NPOT)



Install
Usage overview
Usage examples
Option 1: Input preassembled data
Option 2: Input data gathered from web services
Solve

Option 2: Input data gathered from web services

`CityWaterBalance` has other functions that assemble data for the model. At this time, these functions access US-based web services.

Specify spatial and temporal boundaries

Define an area of interest (AOI) and upload that geometry to the USGS Geo Data Portal (GDP). The GDP will give the geometry a name, which may start with 'upload:'. Here we use a geometry that is already available to the GDP in order to automate the example.

```
geometry <- 'sample:Counties'
attribute <- 'STATE'
value <- 'RI'
area <- 2707
start <- "2010-01-01"
end <- "2010-12-31"
```

Get atmospheric data

```
latitude <- 41.5801
atm <- getAtmoFlows(start, end, geometry, attribute, v
alue, latitude)
```

Get streamflow data

Choose streamgages to evaluate total inflow and outflow for the AOI. NWIS mapper may be useful here.

```
ingages <- c("01112500")
outgages <-c("01113895","01114000","01117000","0111850
0")
```

Considerations

- Lots of help available for writing a package for R
- Not every workflow needs to be a package
- R notebooks as one possible alternative
- Do make your steps traceable, in whatever format you choose

Thank you!

erban.laura@epa.gov

@leerban

lerban