Break Free from Your Inertia

Victor Muñoz

SRK Consulting (Canada) Inc., 1066 W Hastings Street, Vancouver, BC, Canada V6E 3X2

Abstract Innovation is the coveted fruit of all consultancies. Unfortunately, innovation most often exists within the world of academia, where theoretical concepts appear to supply simple solutions to complex problems. Meteorological data have increasingly complemented data from ground-based sources to the extent of real-time information. R statistical computing language, provide powerful data access, compilation, analysis, and graphing tools. In addition, the time required to perform typical hydrological analyses is significantly reduced by several orders of magnitude when compared to traditional methods. R can bridge the gap between academia and consultancies—marrying innovation with simplicity, affordability, and practical application—and help improve our understanding of the global environment.

Key words Hydrology, R, tools, spreadsheets, innovation

Introduction

Innovation is the coveted fruit of all consultancies. Yet, it appears restricted to academia, where theoretical concepts are applied to complex problems and appear to provide simple solutions. In reality, these solutions are rarely simple, practical, or affordable. Herein lies the great paradox of innovation.

Information Paradigm – from Spreadsheets to Satellite Information

Over the past 20 years, the complexity of data collection and analysis tools has evolved with the exponential increase of available data.

Most experienced mining consultants graduated when consultancy work was just beginning to become technique-driven and discipline-specific. The fields of water management and hydrology have since seen a significant shift in methodology. For example, hydrology in the 1980s was based mostly on monthly or daily data collected from local meteorological stations; most of the information was compiled as isohyets maps of simplified yearly or monthly tables.

In the 1980's, applying satellite technology in a meteorological context was introduced, but information was scarce and not yet public. During this time, the bulk of hydrologic analyses relied on simple spreadsheet software such as Lotus 1-2-3 and early versions of Microsoft Excel. The growth of the internet in the 1990s, carried the benefits of access to increasing amounts of information. Today, meteorological information from satellite databases can be used to complement data from ground-based sources such that real-time hydrological information is accessible.

The publication of the First Assessment Report by the Intergovernmental Panel on Climate Change (IPCC) in 1990 spurred increased modelling of climate data (IPCC 1990). The amount of information has since increased exponentially. Available data sources include satellite, ground stations, and projections resulting from climate modelling. Examples of some of these worldwide data sources are listed in Appendix A.

The Problem

Sources such as those listed in Appendix A produce large volumes of data. For these data to be effective in understanding regional and local hydrology, they must be processed efficiently. However, the consultancy environment presents two main challenges to efficient data capture and analysis:

1) **Drain on time:** Too much time is spent in data capture and repetitive data processing, instead of on data analysis and problem-solving. The budget is consumed by simple repetitive tasks instead of scientific analysis and potential innovation.
2) **Data processing limitations:** The sheer size of spreadsheets become dif ficult to process, creating computer processing delays and software crash es. This is also an unnecessary drain on budget and resources.

About R—a Possible Solution for Hydrology

Most end users of climate data are not engineers nor work in a mining-related field, but rather work in academia, where time pressures differ from those in the consultancy environment. Script-based programming tools allow consultants to reduce data processing time to a matter of seconds, thus enable hydrological analyses to extend beyond the scope of conventional consultancy work.

The R statistical language is an example of script-based programming. *R* is a free, open-source software environment that was released in 1993. It has since become the standard problem-solving tool for a growing body of researchers in industry, government, and academia who work with large quantities of data. Statistical methods and other functionalities can be modified and combined by the final users, therefore *R* is always pertinent and up-to-date. *R* has built-in access to public information from government institutions, including the United States Geological Survey (USGS 2016), NOAA (NOAA 2016), and Google tools (Gesmann, Castillo and Cheng 2017) and APIs (Google 2017).

R is based on a simple lean environment, and its capacities can be extended through thousands of libraries obtained from CRAN (Comprehensive R Archive Network, one of the most valuable package repositories) (CRAN 2017), as well as from other sites. CRAN's libraries are public and free, and most of them directly reference public papers and specific library manuals. More than 10,300 libraries are available as of April 2017 (CRAN 2017).

Some particular libraries relevant to hydrological analyses are:

- **HydroTSM**: Provides compilation of tools for hydrological data and time series analyses (Zambrano-Bigiarini 2016).
- **EcoHydRology**: Provides sets of hydrology-related tools for engineers and scientists, from snowmelt models to baseflow analyses (Fuka et al. 2014).

- **Evapotranspiration:** Provides methodology to calculate evapotranspiration from Penman-Monteith, Penman, FAO, Morton (Guo and Westra 2016).
- **ncdf4:** Enables reading of netCDF files, the typical format used for climate change or world climate data; this tool can read files containing several gigabytes of information in a few seconds (Pierce 2015).
- **nsRFA**: Provides a collection of statistical tools for predictive (non-supervised) applications of regional analysis methods in hydrology, including L-moments applications (Viglione et al. 2016).
- **caret**: Provides the compilation of several functions (more than 40 packages) to create predictive models. These tools include: artificial neural networking, machine learning, parallel processing, bootstrapping, and resampling (Kuhn 2016).
- **Parallel:** Allows the division of tasks along all the CPU cores available. Simply applying three lines of code can divide a task between multicores (i.e. divide the workload into 8 or 20 cores simultaneously) (R Core Team 2017).

Personal Contribution – Hydro Library

Using this rich programming environment, I created several short scripts that grew in complexity over time to now form a local library. The compilation of these scripts and these saving time increased my productivity and allows me to invest in more complex hydrological problems and obtain more challenging sources. This positive feedback loop started with the curiosity to use new sources and methodologies beyond spreadsheets to improve quality, productivity, and consistency.

Currently, I am already using more than 40 different other public libraries that are the building blocks upon which I base my work necessities. These facilitate data access, compilation, and analysis. These custom-made scripts can be reused and therefore can be applied to different projects, improving hydrologic studies in any region of the world.

Most of the actual hydrological sources presented in Appendix A are impossible to manage and to use with typical spreadsheet tools, because most of these sources work with climate files such as netCDF files (Unidata 2017). netCDF files are the standard for climate information, and are typically used as an output from satellite information, reanalysis or climate change outputs. One netCDF file can contain hourly precipitation information for the entire planet from 1979 to now, and can be as big as 15GB, which is beyond what any spreadsheet can process or even read.

In addition, the time required to perform typical hydrological analyses is greatly reduced. Figure 1 below compares the time frame of data processing for standard spreadsheets and *R,* at both linear and logarithmic scale. In all cases, the time reduction is higher than 95%, which demonstrates that this is a standard time reduction for R-scripts.

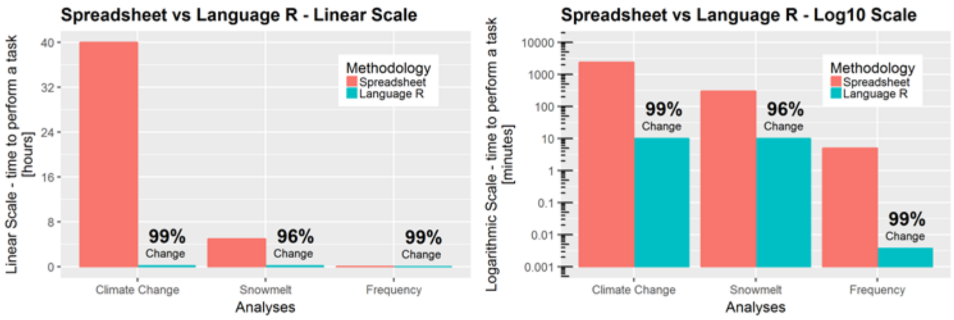

***Figure 1**: Comparison in processing time between Spreadsheet and R.*

The time required for three types of analyses using spreadsheets is compared with time required using R language:

- **Climate change analysis time frame**: The evaluation of climate change data, anywhere on the planet, used to take one week. It now takes only two minutes of processing time with the new methodology. The analysis captures and processes current information (35 GB of information for IPCC Assessment Report AR5), reads historical information from reanalyses (10 GB) and then resolves in one engineer design parameter (Muñoz, Shapka-Felps, and Rykaart 2017).
- **Snowmelt analysis time frame**: Traditional snowmelt models, based on a simplified temperature index, take around of five hours to generate. The EcoHydRology Library (Fuka et al. 2016) presents an energy snowmelt model (Walter et al. 2005) that is typically more representative than the temperature index. It used worldwide meteorological information obtained directly from re-analysis ERA-Interim (ECMWF 2017), and accessing worldwide topographical information from the USGS (USGS 1996), such as slope and topographical aspect. A simple snowmelt model for any place on the planet can be presented in less than 10 minutes.
- **Frequency analysis time frame**: Based on the public nsRFA library (Viglione et al. 2016), the approach reduces the frequency analysis time from five minutes to around one second or less. This methodology includes an L-moment result, which is less sensitive to outliers (Hosking and Wallis 2005).

These few examples present a reduction in processing time that can be invested in broader and deeper analyses (if required) of meteorological parameters, such as correlation matrixes, cluster analyzes, artificial neural network, and Bayesian statistics. Additionally, these parameters are already implemented in dozens of libraries in *R* (CRAN 2017).

These types of analyses are outside of the scope of any spreadsheet-based analysis and potentially gives the consultancy sector an improved capacity for worldwide meteorological estimations. Overall, this change from spreadsheets to *R* which is completely free for everybody, brings the consultancy environment more definitive results that up-to-date with worldwide meteorological estimations.

Academia has a proactive advantage, where effective research and study are encouraged and can create innovation. In contrast, the consultancy environment supports the development of practical and familiar tools that lead to simple results. We can learn from academia and integrate new tools, information, and technology into our consultancy work. Breaking off from the inertia of familiar protocols offers new opportunities that can give rise to a more efficient use of information in a fraction of the time needed for traditional methods.

Conclusion

The amount of data available for hydrological analyses has increased dramatically, thanks to the availability of vast public databases and powerful data management tools. It is time to progress beyond limited traditional tools such as spreadsheets that have kept us in a technological comfort zone. Adopting current databases and tools can heal the gap between consultancy and academia—marrying simplicity and innovation—thereby helping to improve our understanding of the global environment.

References

[CRAN] Comprehensive R Archive Network (2017) Download and install R. The Comprehensive R Archive Network. https://cran.r-project.org

[ECCC] Environment and Climate Change Canada (2017) The Canadian Regional Climate Model. Environment and Climate Change Canada. http://www.ec.gc.ca/ccmac-cccma/default.asp?lang=en&n=82DD0FCC-1

[ECMWF] European Centre for Medium-Range Weather Forecast (2017) Reanalysis ERA-Interim. http://www.ecmwf.int/en/research/climate-reanalysis/era-interim ERA-Interim

Fuka, D.R., Walter, M.T., Archibald, J.A., Steenhuis, T.S. and Easton, Z.M. (2016) R library EcohydRology: A community modelling foundation for Eco-Hydrology. https://cran.r-project.org/web/packages/EcoHydRology/EcoHydRology.pdf

Gesmann, M., Castillo, D. and Cheng, J. (2017) R library googleVis: Interface to Google Charts. https://cran.r-project.org/web/packages/googleVis/googleVis.pdf

Google. Accessed (2017) Google Maps APIs. https://developers.google.com/maps/

Guo, D. and Westra, S. (2016) R library Evapotranspiration: Modelling Actual, Potential and Reference Crop Evapotranspiration. https://cran.r-project.org/web/packages/Evapotranspiration/Evapotranspiration.pdf

[IPCC] Intergovernmental Panel on Climate Change (1990) IPCC first assessment climate change 1990. IPPC.

Hosking, J.R.M. and Wallis J.R. (2005) Regional Frequency Analysis: An Approach Based on L-moments. Cambridge University Press.

Kuhn, M. (2016) R library Classification and Regression Training. https://cran.r-project.org/web/packages/caret/caret.pdf

[NASA] National Aeronautics and Space Administration (2017) Prediction of Worldwide Energy Resource. https://power.larc.nasa.gov/

[NOAA] National Oceanic and Atmospheric Administration (2016) R library rnoaa: 'NOAA' Weather Data from R. https://cran.r-project.org/web/packages/rnoaa/rnoaa.pdf

[NOAA] National Oceanic and Atmospheric Administration (2017) National Centers for Environmental Information. https://gis.ncdc.noaa.gov/maps/ncei

Muñoz, V., Shapka-Felps, T. and Rykaart, M. (2017) Integrating Climate Change into Engineering Design. IMWA 2017 – Mine Water & Circular Economy Conference

Pierce, D. (2015) R library ncdf4 Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files. https://cran.r-project.org/web/packages/ncdf4/ncdf4.pdf

R Core Team. (2017) R library parallel Support for Parallel computation in R. https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf

Reanalyses.org. (2017) Atmospheric Reanalyses Comparison Table. Advancing Reanalysis. http://reanalyses.org

Unidata (2017) Network Common Data Form (NetCDF). Unidata. http://www.unidata.ucar.edu/software/netcdf

[USGS] U.S. Geological Survey (1996) Global 30 Arc-Second Elevation (GTOPO30). https://lta.cr.usgs.gov/GTOPO30

[USGS] U.S. Geological Survey (2016) R library dataRetrieval: Retrieval Functions for USGS and EPA Hydrologic and Water Quality Data.SGS r info libraries. https://cran.r-project.org/web/packages/dataRetrieval/dataRetrieval.pdf

Viglione, A., Hosking, J.R.M., Laio, F., Miller, A., Gaume, E., Payrastre, O., Salinas, J.L., N'Guyen, C.C. and Halbert, K. (2016) R library nsRFA Non-supervised Regional Frequency Analysis. https://cran.r-project.org/web/packages/nsRFA/nsRFA.pdf

Walter, M.T., Brooks, E.S,, McCool, D.K., King, L.G., Molnau, M., and Boll, J. (2005) Process-based snowmelt modeling: does it require more input data than temperature-index modeling? Journal of Hydrology 300(1): 65–75, doi:10.1016/j.jhydrol.2004.05.002

Zambrano-Bigiarini, M. (2016) R library hydroTSM Time series management, analysis and interpolation for hydrological modelling. https://cran.r-project.org/web/packages/hydroTSM/hydroTSM.pdf

Zambrano-Bigiarini, M., Nauditt,A., Birkel, C., Verbist, K. and Ribbe, L. (2017) Temporal and spatial evaluation of satellite-based rainfall estimates across the complex topographical and climatic gradients of Chile. Hydrol. Earth Syst. Sci., 21, 1295–1320, 2017; doi:10.5194/hess-21-1295-2017

Appendix A – Climate Data Sources

Each of the following worldwide meteorological sources provide an understanding of climatic parameters in a given area on the planet such as precipitation, temperature, wind speed, solar radiation, relative humidity, dew point temperature, and evaporation. All of these sources are freely available for everybody:

- **National Climatic Data Center from NOAA (Source: land surface stations):** NOAA databases provide access to worldwide, ground-based meteorological information. Two important databases are the Global Surface Summary of the Day (GSOD), with more than 9,000 stations, and the Global Historical Climatology Network (GHCN), with in excess of 75,000 stations located in more than 180 countries (NOAA 2017). Records span up to data 150 years. The NOAA databases are excellent sources for historical and current meteorological data.
- **Satellite-based rainfall estimates (Source: satellite data):** There are several satellite sources of precipitation data which cover a variety of assumptions, spatial resolutions, and coverage. Most of then cover latitudes between 60° to 50° degrees North and 60° to 50° degrees South, with a variable temporal resolution from 0.04° to 0.25° (~4 to 28 km). As an example of these sources or dataset are: CMORPH, PERSIANN, PERSIANN-CDR, PERSIANN-CCS, TRMM, CHIRPSv2, MSWEPv1.1 and PGFv3. (Zambrano- Bigiarini et al 2017).
- **Reanalysis tools (Source: processed ground and satellite data):** Reanalysis uses a process called data assimilation to combine data from satellites, land surface stations, and numerical models to simulate the earth's climate. The time step for reanalysis can be as short as eight hours and can be applied to data collected since 1979. Two good examples of reanalysis tools are Modern-Era Retrospective Analysis for Research and Applications (MERRA) (NASA 2017) and ERA-Interim from the European Centre for Medium-Range Weather Forecasts (ECMWF 2017). MERRA and ERA-Interim encompass more than 100

meteorological parameters, including total precipitation, rainfall, snowfall, wind speed, air temperature, and relative humidity.

- **Global climate change models (Source: climatic models):** Global climate change models and scenarios are presented as assessment reports by member meteorological institutes of the Intergovernmental Panel on Climate Change (ECCC 2017). Each model and scenario presents projections of meteorological parameters for the entire planet, typically up to the year 2100. The combination of climate models and scenarios can produce hundreds of projections for any given location on earth. This variation among projections can be incorporated into long-term climate predictions for a given area.