State of R in Hydrological Modelling

J. A. Torres, E. Pebesma

Institute for Geoinformatics, University of Münster, Weseler Str. 253, 48151 Münster, Germany

arturo.torres@uni-muenster.de, edzer.pebesma@uni-muenster.de

May 30, 2013

Keywords: *Hydrological modelling, modelling frameworks,* R

1 Introduction

Access to clean and safe water is a human right and to grant it under a sustainable development point of view is paramount for well being of the future generations. Therefore, it is important to understand the interaction between the material flows to coastal waters that are constrained by catchment boundaries and the human activities therein, and those materials that are tied to trade and other trans-boundary processes (e.g. residence time, transport and fate of physical, chemical and microbiological water-quality determinants) and their global implications on preserving the quality of the natural environment.

In consequence, in order to address an objective hydrological modelling, specifically looking forward on the validation of the hydrological cycle, and the transport and fate of sediments and solutes in surface water resources, it is paramount to recognize that in environmental modelling all model structures, regardless of their complexity, are to some extent in error (K. J. Beven, 1989; Grayson, Moore, & McMahon, 1992; Freer, McMillan, McDonnell, & Beven, 2004).

In this sense, model comparison in structure, calibration methods and simulation events is essential for choosing objectively the suitable configuration of the model for addressing a specific task related with hydrological modelling. To accomplish this, a novel, versatile, and open source programming language is provided by the R Project for Statistical Computing (Ihaka & Gentleman, 1996; R Development Core Team, 2013). R was influenced by two existing languages: S (Becker, Chambers, & Wilks, 1988) and Scheme (Steel & Sussman, 1975). R is very similar in appearance to S, but the underlying implementation and semantics are derived from Scheme (Ihaka & Gentleman, 1996). Being as open source software, several authors argue (Andrews, Croke, & Jakeman, 2011) that R provides for standardised tests, comparisons of models, reproducibility of methods and results, as is often required by science and research.

Several packages have been developed in R code for developing, implementing and evaluating different hydrological modelling tools and frameworks. Under this focus the authors believe that the programming platform that is growing based on the R language is a promissory field where research and development are the primary purpose. The present paper has the aim to introduce a state-of-art of hydrological modelling in R.

2 Hydrological modelling

Several applications in river and environmental engineering and science, related with hydrological modelling are related to the analysis of rainfall and hydrometric time series in order to implement rainfall-runoff and water-quality models as a conceptual mathematical basis for solute transport and fate assessment. Similarly, such applications and models are common in basin water resources and flood risk management e.g. in the study of the probable maximum precipitation and probable maximum flood. In this case, the application of hydrological models is done as the conceptual basis to simulate flood events in probable scenarios of storm events.

Regarding to the challenges for modern hydrological and environmental research as it is depicted in current researches –e.g. McDonnell et al. (2010); Swaney et al. (2011)– is essential to understand and develop a comprehensive modelling framework that includes as an important step the uncertainty analysis in order to identify primary physical controls, and henceforth for coupling, in a most suitable way, inland hydrological models with the coastal system at regional, transboundary and global scales. Henceforth, in hydrological applications a main aim is to consider a suitable and reproducible modelling framework that takes into account data input, spatial interpolation, calibration and simulation, and includes geospatial capabilities for querying, updating, sharing and visualization of data, methods and results. This focus is addressed in the following subsections where is presented in a succinct manner the existing open source data and software in R,

considering the input data, hydrological models and uncertainty analysis, and general tools as support in hydrological analysis.

3 Useful functionality in R

As R is an environment for statistical computing and graphics, it provides a large amount of generic functionality that can be useful from the perspective of hydrological modelling. This includes basic data manipulation (import, export, selection, joining of tables), handling and analyzing MCMC outputs, visualizing with high control and print quality, handling and analyzing time series, handling and analyzing spatial data and spatio-temporal data, geostatistics, downscaling, upscaling, aggregation and disaggregation, and so on. Higher level functionality comes from the fact that R is cross platform, provides easy ways to disseminate new extension packages, enables reproducible research, and provides support through open mailing lists forums, commercial companies, and books. The highest level functionality is that of its community: a large, open and active community is growing because many students now obtain experience with using and/or programming R during their studies.

4 Data management and R

Input datasets are required in hydrological modelling. The reproducibility issue in research implies to share a common data model as input independent of the modelling framework used. Some efforts in this way has been done to supply standardized open data for research purpose, e.g. ECRINS and the EOBS datasets, and the web services that are discussed below.

4.1 The ECRINS dataset

The European Environment Agency (EEA) has been developed the Catchments and Rivers Network System (ECRINS) version 1.1. The ECRINS is the hydrographical system currently in use at the European level as well as widely serving as the reference system for the Water Information System (WISE) (European Environment Agency - EAA, 2012).

In order to attend the several sectors and potential users of the database, ECRINS has been delivery in different layers and ancillary tables in MS Access® Personal GeoDatabases (PGDBs) format. The PGDBs, in Microsoft® proprietary format, can be handed with both MS Access® and ArcGIS ®, which is the software used to build ECRINS v1.x. However, the PGDBs can also be used with most open source GIS and database managers (European Environment Agency - EAA, 2012). In this sense packages as `foreign` (R Core Team et al., 1999-2013) includes the method `read.dbf` for importing a .dbf file into a R dataframe; and the package `sp` (E. Pebesma & Bivand, 2005-2012), has i.a. the class `SpatialPolygons` which is a data object equivalent to an ESRI polygon shapefile containing information for polygons, and additional similar definitions for spatial points and lines are defined through the objects `SpatialPoints` and `SpatialLines`, respectively.

Equally, a binding package for the Frank Warmerdam's Geospatial Data Abstraction Library (GDAL, http://www.gdal.org) is available in R through the package `rgdal` (Bivand et al., 2003-2013). It allows to deploy multiple classes defined in the `sp` package and access to the projection/transformation operations from the PROJ.4 library (https://trac.osgeo.org/proj/) and to the OGR library. The OGR Simple Features Library is a C++ open source library for reading, and in some cases writing, a variety of vector file formats including ESRI Shapefiles and PGDBs .mdb files via ODBC (Warmerdam, 2013). Therefore, using `rgdal` both GDAL raster and OGR vector map data can be imported into R and exported, and the ECRINS database could be handled properly.

4.2 The EOBS dataset

The E-OBS (Haylock et al., 2008) is a daily gridded observational dataset for precipitation, temperature and sea level pressure in Europe, prepared and developed by the European Climate Assessment & Dataset project, ECA&D (http://eca.knmi.nl). The datasets contain information on changes in weather and climate extremes, as well as the daily data needed to monitor and analyse these extremes at a resolution of 0.25°. These datasets are available for downloading from the web page of the project ENSEMBLES (ECA&D, 2012).

The format of the gridded data is NetCDF (network Common Data Form) which is a set of interfaces for array-oriented data access and a freely-distributed collection of data access libraries for C, Fortran, C++, Java, and other languages. The netCDF libraries support a machine-independent format for representing scientific data (Unidata, 2012). The Open Geospatial Consortium membership has approved the Enhanced Data Model Extension to the OGC Network Common Data Form (netCDF) Core Encoding Standard (http://www.unidata.ucar.edu/blogs/news/entry/ogc_adopts_netcdf_enhanced_data).

The access to NetCDF files for reading data into R and for creating new netCDF dimensions, variables, and files, or manipulating existing netCDF files from R, is possible through the `ncdf4` package (Pierce, 2010-2012) and the `RNetCDF` package (Michna, 2004-2012).

4.3 Web services and interoperability

R provides generic interfaces to `curl` and `libxml`, through packages `RCurl` (Lang, 2004-2013) and `XML`

(Lang, 2000-2013) respectively. These can be used to access web services, and building on these packages the package `sos4R` (Nüst, Stasch, & Pebesma, 2011) can be used to retrieve sensor data from an OGC sensor observation service, e.g. to retrieve data in `waterML`.

In the opposite direction, any R functionality can be accessed through a web service interface, for instance using the `Rserve` package (Urbanek, 2006-2013). A higher-level, generic OGC WPS interface was developed in the context of the 52°North WPS, called WPS4R (Hinz, Nüst, Proß, & Pebesma, 2013).

5 Hydrological analysis in R

5.1 Spatial prediction of rainfall in R

According to Lanza, Ramírez, and Todini (2001) traditionally areal rainfall has been estimated applying some kind of interpolation method and aggregation technique assuming that the "real punctual precipitation" corresponds to the in-situ measured data. Such approaches can be traced from the subjectively method of the isohyets (Linsley et al., 1949) to the more objectively geometrical interpolators based on Thiessen polygons (Linsley et al., 1949), (WMO, World Meteorological Organization, 1986), to the mathematical surface interpolators based on splines (Matheron, 1981), to the new branch of the statistical theory: the geostatistics introduced by Matheron (1971).

Currently in R there are several methods for spatial prediction which are recognized as part of the state-of-art in rainfall prediction (Hengl, AghaKouchak, & Tadic, 2010). Package `gstat` (E. J. Pebesma, 2004) provides methods for spatial and spatio-temporal geostatistical modelling, prediction and simulation. It contains methods for variogram modelling and description; simple, ordinary and universal point or block (co)kriging, sequential Gaussian or indicator (co)simulation; and additional utility methods for variogram map plotting.

The package `geoR` (Ribeiro Jr & Diggle, 2001-2013) contains methods for Geostatistical analysis including traditional, likelihood-based and Bayesian methods. Additional packages available are `intamap` and `psgp`.

5.2 Hydrological models in R

An important contribution in hydrological modelling is done by Andrews et al. (2011) with the R package `hydromad` (`http://hydromad.catchment.org`). It is based loosely on the unit hydrograph theory of rainfall-runoff modelling. More than a single hydrological model `hydromad` is a framework with several options of configurations that includes different Soil Moisture Accounting (SMA) models and objective calibration methodologies. In consequence, it can be used cohesively with workflows based on R. Two areas of focus for the package are discrete event separation and the design of fit statistics, and how event-based data analysis can be useful in a modelling context (Andrews et al., 2011).

Topmodel (K. Beven, Lamb, Quinn, Romanowicz, & Freer, 1995; K. J. Beven, 1997) more than a semi-distributed hydrological model is a compendium of hydrological concepts and a simple approach to predict spatial pattern of responses in a catchment. It may be seen as a product of two objectives (K. J. Beven, 2012, p. 190): (1) to develop a pragmatig and practical forecasting and continuous simulation model; (2) to develop a theoretical framework within which is perceived hydrological processes, issues of scale and realism, and model procedures may be researched. *Topmodel* is a reference against which other modelling concepts could be compared (see Buytaert et al., 2008). The implementation of the 1995 Fortran version of *Topmodel* has been done recently by Buytaert (2012) in R language within the `topmodel` package. The new functionality is being developed as part of the `RHydro` package on R-Forge.

The grid-based Water Flow and Balance Simulation Model WaSiM-ETH (Schulla (2012), `http://www.wasim.ch/en/`) is a distributed, deterministic, mainly physically based hydrologic model. It is a well-established tool for investigating the spatial and temporal variability of hydrological processes in complex river basins. An interface for the model is done through the homonym package `wasim` (Reusser & Francke, 2008-2011), which provides methods for visualising and analysing output files of the model.

There are several additional packages published at CRAN in the area of hydrology. The package `HydroMe` (Omuto, 2012) is suitable for estimating the parameters in infiltration and water retention models, commonly applied in soil science, by curve-fitting method. The package `hydroTSM` (Zambrano-Bigiarini, 2010-2013b) allows S3 functions for management, analysis, interpolation and plotting of time series used in hydrology and related environmental sciences; particularly oriented to hydrological modelling tasks.

5.3 Uncertainty analysis

According to Freer et al. (2004), the extent in error in environmental modelling is an important issue. Principally, this error can be attributed to two main factors (K. Beven, Leedal, & Alcock, 2010): firstly, an imperfect perceptual model due to an imperfect knowledge, i.e. epistemic uncertainty. Secondly, the highly simplified definitions in mathematical models required which commonly cannot represent the complexity of the natural interacting processes under modelling. As a consequence, tools and objective methodologies are required in hydro-

logical modelling.

In this sense, it is recommended that the model predictions and particularly predictions of the impacts change should be associated with an uncertainty analysis so that the significance of the difference between simulations can be assessed including the concept of equifinality (K. J. Beven, 2006, 2012) in parameter identification, which is related with the identification of the adequate and well identified parameters avoiding over-parametrisation or redundant parameter definition that fuzzes the real physical sense of them.

Therefore, a versatile open source, multiple platform programming language as R is an appropriated ground to address it. A useful tool for model calibration and sensitivity analysis processes in R is the `hydroPSO` package (Zambrano-Bigiarini & Rojas, 2013). This package implements several state-of-the-art enhancements and fine-tuning options to the Particle Swarm Optimisation (PSO) algorithm. `hydroPSO` interfaces the calibration engine to different model codes through ASCII files and/or R wrapper functions for exchanging information on the calibration parameters. The optimisation is based on evaluating the goodness-of-fit functions until a maximum number of iterations or a convergence criterion are met. The evaluation of the calibration process is supported by plotting functions that facilitate the interpretation of results (Zambrano-Bigiarini & Rojas, 2013).

5.4 Further hydrological functionality

The package `hydroGOF` (Zambrano-Bigiarini, 2010-2013a) has S3 functions implementing both statistical and graphical goodness-of-fit measures between observed and simulated time series, it is functional during the tasks of calibration, validation, and simulation of hydrological models. Similar packages are `tiger` (Reusser, 2009-2013) and `qualV` (Jachner, van den Boogaart, & Petzoldt, 2007), the former allows to determine and visualise temporally resolved groups of typical differences (errors) between two time series, the later contains functions for qualitative model comparison.

The `EcoHydRology` package (Fuka, Walter, Archibald, Steenhuis, & Easton, 2011-2013) presents a "flexible foundation for scientists, engineers, and policy makers to base teaching exercises as well as for more applied use to model complex eco-hydrological interactions". It includes methods as `BaseflowSeparation` for reading a streamflow dataset and producing a baseflow dataset. Also the package has the method `calib_swat_ex` for calibration of the SWAT model (Neitsch, Arnold, Kiniry, & J.R., 2011).

6 Conclusions

We present an introduction to the state-of-art of hydrological modelling as is developed in the open source software R. This constitutes a starting point for research because it grants a suitable programming platform where standardised tests and comparisons of models is possible, searching for reproducibility of methods and results as is often required in the context of science and research. The discussed packages let a common framework in R for developing, implementing and evaluating different hydrological modelling tools. Therefore, this is a promissory field where research and development in hydrological modelling are possible and constitutes a primary purpose.

References

Andrews, F. T., Croke, B. F. W., & Jakeman, A. J. (2011). An open software environment for hydrological model assessment and development. *Environmental Modelling and Software*, *26*. Retrieved from http://hydromad.catchment.org/ doi: doi:10.1016/j.envsoft.2011.04.006

Becker, R., Chambers, J., & Wilks, A. (1988). *The New S Language*. Pacific Grove, CA Wadsworth.

Beven, K., Lamb, R., Quinn, P., Romanowicz, R., & Freer, J. (1995). TOPMODEL. In V. P. Sing (Ed.), *Computer models of watershed hydrology. Water Resources Publications, Colorado. pp. 627-668.*

Beven, K., Leedal, D., & Alcock, R. (2010). Uncertainty and good practice in hydrological prediction. *VATTEN*, *66*, 159-163.

Beven, K. J. (1989). Changing ideas in hydrology. the case of physically-based models. *Journal of Hydrology*, *105 (1-2)*, 157-172.

Beven, K. J. (1997). *Distributed hydrological modelling: Applications of the TOPMODEL Concept*. Wiley.

Beven, K. J. (2006). A manifesto for the equifinality thesis,. *J. hydrology,*, *320*, 18-36.

Beven, K. J. (2012). *Rainfall-runoff modelling: The Primer* (Second, Ed.). Wiley-Blackwell. (Lancaster University, UK)

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., & Hijmans, R. (2003-2013, May). Package "rgdal": Bindings for the Geospatial Data Abstraction Library (.8-9 ed.) [Computer software manual].

Buytaert, W. (2012, February). Package 'topmodel' [Computer software manual].

Buytaert, W., Reusser, D., Krause, S., & Renaud, J.-P. (2008). Why can't we do better than TOPMODEL? *Hydrological Processes*, *22*, 4175-4179.

ECA&D. (2012). *E-obs datafiles 1950-01-01 until 2012-06-30*. Retrieved from http://eca.knmi.nl/download/ensembles/download.php

European Environment Agency - EAA. (2012). *EEA catchments and rivers network system, ECRINS v1.1. rationales, building and improving for widening uses to Water Accounts and WISE applications* (EEA Technical report No. 7/2012). (Luxembourg: Publications Office of the European Union)

Freer, J. E., McMillan, H., McDonnell, J. J., & Beven, K. J. (2004). Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology*, *291*, 254-277.

Fuka, D., Walter, M., Archibald, J., Steenhuis, T., & Easton, Z. (2011-2013). EcoHydRology: A community modeling foundation for Eco-Hydrology (.4.7 ed.) [Computer software manual].

Grayson, R. B., Moore, I. D., & McMahon, T. A. (1992). Physically based hydrologic modelling 2. is the concept realistic. *Water Resources Research*, *28 (10)*, 2659-2666.

Haylock, M. R., Hofstra, N., Tank, A. M. G. K., Klok, E. J., Jones, P. D., & New, M. (2008). A European daily

high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *Journal of Geophysical Research*, *113*. doi: 10.1029/2008JD010201

Hengl, T., AghaKouchak, A., & Tadic, M. P. (2010). Methods and data sources for spatial prediction of rainfall. In F. Y. Testik & M. Gebremichael (Eds.), *Rainfall: State of the science* (p. 287). American Geophysical Union (AGU).

Hinz, M., Nüst, D., Proß, B., & Pebesma, E. (2013). *Spatial Statistics on the Geospatial Web*. short paper, Agile 2013 proceedings.

Ihaka, R., & Gentleman, R. (1996). R: a languague for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*, 299-314.

Jachner, S., van den Boogaart, K. G., & Petzoldt, T. (2007). Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay (R Package qualV). *Journal of Statistical Software*, *22(8)*, 1-30. Retrieved from http://www.jstatsoft.org/v22/i08/

Lang, D. T. (2000-2013, March). Package "XML": Tools for parsing and generating XML within R and S-Plus (3.96-1.1 ed.) [Computer software manual].

Lang, D. T. (2004-2013, March). Package "RCurl": General network (HTTP/FTP/...) client interface for R (1.95-4.1 ed.) [Computer software manual].

Lanza, L. G., Ramírez, J. A., & Todini, E. (2001). Stochastic rainfall interpolation and downscaling. *Hydrology and Earth System Sciences*, *5(2)*, 139-143.

McDonnell, J. J., McGuire, K., Aggarwal, P., Beven, K. J., Biondi, D., Destouni, G., ... Wrede, S. (2010). How old is streamwater? Open questions in catchment transit time conceptualization, modelling and analysis. *Hydrological Processes*, *24*, 1745-1754.

Michna, P. (2004-2012, July). Package "RNetCDF": R Interface to NetCDF Datasets (1.6.1-2 ed.) [Computer software manual].

Neitsch, S., Arnold, J., Kiniry, J., & J.R., W. (2011). Soil and Water Assessment Tool, theoretical documentation, version 2009 [Computer software manual].

Nüst, D., Stasch, C., & Pebesma, E. J. (2011). Connecting R to the Sensor Web. In S. Geertman, W. Reinhardt, & F. Toppen (Eds.), *Advancing Geoinformation Science for a Changing World* (p. 227 - 246). Springer Lecture Notes in Geoinformation and Cartography.

Omuto, C. T. (2012, July). Package "HydroMe": Estimation of Soil Hydraulic Parameters from Experimental Data (1.0 ed.) [Computer software manual].

Pebesma, E., & Bivand, R. (2005-2012, May). Package "sp": classes and methods for spatial data (.9-99 ed.) [Computer software manual].

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, *30*, 683-691.

Pierce, D. (2010-2012, April). Package "ncdf4": Interface to Unidata netCDF (version 4 or earlier) format data files (1.6.1 ed.) [Computer software manual].

R Core Team, Bivand, R., Carey, V. J., DebRoy, S., Eglen, S., Guha, R., ... Free Software Foundation, Inc. (1999-2013, May). Package "foreign": Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase,... (.8-54 ed.) [Computer software manual].

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Reusser, D. (2009-2013, February). Package "tiger": TIme series of Grouped ERrors (.2.2 ed.) [Computer software manual].

Reusser, D., & Francke, T. (2008-2011). Package wasim: Visualisation and analysis of output files of the hydrological model WASIM (1.1.2 ed.) [Computer software manual].

Ribeiro Jr, P. J., & Diggle, P. J. (2001-2013). Package "geoR": Analysis of geostatistical data (1.7-4 ed.) [Computer software manual]. Retrieved from geoR

Schulla, J. (2012). Model description wasim (water balance simulation model) [Computer software manual]. Retrieved from http://www.wasim.ch/the_model.html

Steel, G., & Sussman, G. (1975). *Scheme: An Interpreter for the Extended Lambda Calculus*. MIT Artificial Intelligence Laboratory.

Swaney, D., Humborg, C., Emeis, K., Kannen, A., Silvert, W., Tett, P., ... Nicholls, R. (2011). Five critical questions of scale for the coastal zone. *Estuarine, Coastal and Shelf Science*, *xxx(2011)*, 1-13.

Unidata. (2012). *NetCDF FAQ*. Retrieved from http://www.unidata.ucar.edu/software/netcdf/docs/faq.html#whatisit

Urbanek, S. (2006-2013, March). Package "Rserve": Binary R server (.6-8.1 ed.) [Computer software manual].

Warmerdam, F. (2013). *GDAL - Geospatial Data Abstraction Library*. Internet. Retrieved from http://www.gdal.org/index.html (Consulted in 24.05.2013)

WMO, World Meteorological Organization. (1986). *Manual for estimation of probable maximum precipitation* (Operational Hydrology Report Nos. 1, WMO N.332). Geneva.

Zambrano-Bigiarini, M. (2010-2013a). Package "hydroGOF": Goodness-of-fit functions for comparison of simulated and observed hydrological time series (.3-5 ed.) [Computer software manual].

Zambrano-Bigiarini, M. (2010-2013b, February). Package "hydroTSM": Time series management, analysis and interpolation for hydrological modelling (.3-6 ed.) [Computer software manual].

Zambrano-Bigiarini, M., & Rojas, R. (2013). hydroPSO: A Model-independent Particle Swarm Optimization Software for Model Calibration. *Environmental Modelling & Software*, *43*, 5-25.