Theory and Practice of Reproducible Research

Riccardo Rigon, Francesco Serafin, Marialaura Bancheri

OGRS, Perugia, October 12, 2016
Antonio Canova gypsum statues bring a series of little signs. They served the stonemasons to reproduce “industrially” the opera. Art became “reproducible” for the first time.
I have been frustrated often with statisticians and computer scientists who write papers where they develop new methods and seem to demonstrate that those methods *blow away* all *their competitors*. But then no software is available to actually test and see if that is true. ... In my mind, new methods/analyses without software are just *vaporware* ... *If there is no code, there is no paper.*

By Jeff Leek*
Science must be reproducible (i.e. repeatable)

It is the fundamental. It means that everyone (in principle) should be able to take what you write, the experiment you did, the mathematics you drew, and doing it again with his own resources.

“In principle means” that science is often not is not shared … for various reasons …
“In principle” means that

Not anyone can reproduce scientific achievements

S/he must be trained to do it (there are problems of transmission of information here). And, in fact, more advanced results, can be difficult to grab, even for the very same authors.
Reproducibility vs Replicability

Getting more: Replicability

Replicability: Reproduction of the original results using the same tools:
- by the original author on the same machine
- by someone in the same lab/using a different machine
- by someone in a different lab

Reproducibility: Reproduction using different software, but with access to the original code:

Completely independent reproduction based only on text description, without access to the original code.
Analysing a paper for reproducibility

(the case of Formetta et al., 2011)

The JGrass-NewAge system for forecasting and managing the hydrological budgets at the basin scale: models of flow generation and propagation/routing

G. Formetta1, R. Mantilla2, S. Franceschi3, A. Antonello1, and R. Rigon1

1University of Trento, 77 Mesiano St., Trento, 38123, Italy
2The University of Iowa, C. Maxwell Stanley Hydraulics Laboratory, Iowa 52242-1585, USA
3Hydrolis S.r.l., Bolzano, BZ, Italy

Abstract. This paper presents a discussion of the predictive capacity of the implementation of the semi-distributed hydrological modeling system JGrass-NewAge. This model focuses on the hydrological budgets of medium scale to large scale basins as the product of the processes at the hillslope scale with the interplay of the river network. The part of the modeling system presented here deals with the: (i) estimation of the space-time structure of precipitation, (ii) estimation of runoff production; (iii) aggregation and propagation of flows in channel; (vi) estimation of evapotranspiration; (vi) automatic calibration of the discharge with the method of particle swarming.

The system is based on a hillslope-link geometrical partition of the landscape, combining raster and vectorial treatment of hillslope data with vector based tracking of flow in channels. Measured precipitation are spatially interpolated with the use of kriging. Runoff production at each channel link is estimated through a peculiar application of the Hymod model. Routing in channels uses an integrated flow equation and produces discharges at any link end, for any link in the river network. Evapotranspiration is estimated with an implementation of the Priestley-Taylor equation. The model system assembly is calibrated using the particle swarming algorithm. A two year simulation of hourly discharge of the Little Washita (OK, USA) basin is presented and discussed with the support of some classical indices of goodness of fit, and analysis of the residuals. A novelty with respect to traditional hydrological modeling is that each of the elements above, including the preprocessing and the analysis tools, is implemented as a software component, built upon Object Modelling System v3 and gisrastools prescriptions, that can be cleanly switched in and out at run-time, rather than at compiling time. The possibility of creating different modeling products by the connection of modules with or without the calibration tool, as for instance the case of the present modeling chain, reduces redundancy in programming, promotes collaborative work, enhances the productivity of researchers, and facilitates the search for the optimal modeling solution.

1 Introduction

Hydrological forecasting over time has focused on different issues. Determining the discharge of rivers during flood events has been a central topic for more than a century; firstly through the rational model of Malvane (1851), later through the use of instantaneous unit hydrograph models (Sherman, 1932; Dooge, 1959), and more recently including the geomorphological approach (i.e. GIUH; Rodriguez-Iturbe and Valdés, 1979; Gupta and Waymire, 1980; Rosso, 1984; D’Odorico and Rigon, 2003). Even models of runoff generation such as Topmodel (Beven and Kirkby, 1979; Beven, 2001; Franchini et al., 1996) have been used mainly for this purpose.

Subsequently, however, the water resource and river management required the need to estimate a whole set of hydrological quantities such as discharge, evapotranspiration, and soil moisture, bringing very soon to the implementation of more comprehensive modeling systems, like the pioneering Stanford watershed model (Crawford and Linsley, 1966), the Sacramento model (e.g. Burnash et al., 1973), and the PRMS model (Leavesley et al., 1983). They were usually based on the idea of intercommunicating compartments (reservoirs), each representing a process domain, each one with its residence time.
This is a paper, which I co-authored, dealing with a model for rainfall runoff, it is mostly which presents a hydrological model, with an application to a case study.
Reproducible, in this case requires first

Consistency of notation

For what regards to this, the paper is certainly consistent (it is part of the peer-review process to guarantee it).

A more strong statement would require consistency of notation through series of companion papers.

But this paper, in particular, is not a heavy theoretical treatment of some topic, and notation is not really crucial here.
Different story for this paper
(the case of Botter et al., 2010)

1. Introduction

[1] The age of water (or residence time) represents the time spent by water molecules ideally sampled from a given hydrologic system within the reference control volume (measured since the entry through rainfall). Thus, the age of water blends in a single, quantitative attribute information about hydrological and chemical storages, flow pathways, and water sources (e.g., McGuire and McDonnell, 2006). Several field observations (especially built through extensive rainfall/runoff dating by isotope hydrology) and a few theoretical results have established the so-called "old water paradox," according to which a sizable part of the runoff within the hydrologic response of catchment transport volumes is constituted by aged water particles (i.e., by water particles injected at times preceding the event causally related to the observed runoff) (e.g., Maloszewski and Zuber, 1982; McDonnell, 1990; McDonnell et al., 1991; Stewart and McDonnell, 1991; Wilson et al., 1991a, 1991b; Leeney et al., 1993; Reddy et al., 1994; Cermos and McDonnell, 1998; Nyberg et al., 1999; Peters and Ratcliffe, 1998; Burns et al., 1998; Weiler et al., 2003; McGuire et al., 2007; Botter et al., 2007, 2008a, 2009). The release of old water has been explained by the propagation of pressure waves induced by precipitation inputs with a celerity exceeding the pore water velocity (e.g., Beven, 1981, 1989b), including displacement of water previously immobilized within the soil matrix into preferential flow pathways (e.g., Beven and Germann, 1982). However, some of the physical processes controlling the release of pre-event water from catchments are still poorly understood or roughly modeled, and the observational data do not suggest either universal behaviors, nor do they support linear and time-invariant behaviors as assumed by unit hydrograph schemes (e.g., Weiler and McDonnell, 2006). The complexity of the mixing patterns involving event- and pre-event waters in hillslopes is partly a byproduct of the structural complexity of subsurface environments, which are typically characterized by pronounced heterogeneity and time- and space-variable connectivity of flow pathways. For this reason, it is inappropriate to use the point-scale physical laws determining the movement of water and solutes within hillslopes to make predictions at larger scales because of the nonlinearity of flow processes and the uncertain distribution of hydrologic, geological and morphological properties of control volumes (e.g., Beven, 1989a, 2006; Kirchner, 2009). Hence, lumped approaches are frequently employed to describe in an effective manner the overall behavior of hillslopes/catchments. In particular, the water travel time...
This is an outstanding paper dealing with transport for residence time, which I read several times during the last months, in order to reproduce their research (with my own tools)
The Hymod component is applied for each HRU and the runoff production is then propagated in the channel network. A new runoff propagation component is implemented and presented in the next subsection. To study the role and the importance of the channel routing component, a test is performed. Two river basins are used for the test and modeled in three different delineations by using one (DL1), three (DL3), and twenty (DL20) HRUs. Two modeling solutions were set up: Hymod and RHymod in Fig. (7.9).

The modeling solution RHymod includes: the Pristley-Taylor component for the evapotranspiration estimate, the ordinary kriging algorithm for the rainfall spatialization, the Hymod model for the runoff production of the hillslope, and finally the new channel routing component presented in the next section. The modeling solution Hymod differs from the model solution RHymod by only turning off the channel routing component and the discharge for each HRU are just added downstream. LUCA (66) was selected as calibration component for both the modeling solutions. The objective function is the Kling-Gupta efficiency (KGE) function as presented in (63).

The test is performed on two different river basins: Fort Cobb and Little Washita. The simulation period covered 2006-2007 in the case Fort Cobb and 2002-2003 in the case of Little Washita river basin; one year was used for calibration and one year for verification. The simulations time step was hourly.
Therefore, to reproduce JGrass-NewAGE 1.0 results, one has to know the theory of any of the above components. Unfortunately, this is only the first impression. You have to know actually more.
5.3 Motivation for Semivariogram modelling and providing krigings tools in NewAge-JGrass.

5.3.3 The kriging tools in the NewAge-JGrass system

After the variogram assessment, we are able to apply it for kriging interpolation of a dataset. The flow chart of the kriging algorithm is presented in fig.(5.3). The input data are: i) the shape file of the measurement stations, ii) the .csv file of the measured data, iii) the shape file or the raster map of the interpolations points, iv) the semivariogram model to use for the interpolation. The model parameters are: a flag to specify the working mode (raster or vector), the semivariogram model parameter, a flag to specify the kriging type (ordinary, local, or detrended) and some control parameters related to the selected kriging algorithm (maximum distance for local kriging, threshold of the correlation between elevation and measurements for detrended kriging). Within kriging model configuration, different variogram models can be used for different time steps. The outputs could be or a .csv file or a raster map with the interpolated values.

Comparisons with the R-package Gstat (115) are presented in Appendix 1 in order to test the implemented algorithms (ordinary and local kriging).
JGrass-NewAGE 1.0: even more than more

4.2 Catchment analysis

Whatever the conceptualisation, the challenge is to deploy the ideas in robust and correct code. This is accomplished in NewAge-JGrass by using the GEOtools libraries and their implementation of the geographic features which seamlessly integrate with OMS3 programming and uDig. The Horton Machine (127) and (128) is built on top of these libraries which are the modelling components that are actually being used.

To obtain this hierarchical structure it is necessary to first process the raster data from a digital elevation model which is summarised below.

4.2.1 Geomorphological analysis

Starting from the digital terrain model (DTM), the "Horton Machines" (128) components as provided by the GIS uDig-JGrass are used. In sequence, those are:

- **RasterReader**
  - @In: path to the dtm
  - @Out: GridCoverage2D

- **VectorReader**
  - @In: path to the .shp
  - @Out: Feature Collection

- **Pitfiller**
  - @In: dem
  - @Out: depitted dem

- **FlowDir**
  - @In: depitted dem
  - @Out: flow dir. D8

- **DrainDir**
  - @In: depitted dem
  - @Out: drain dir. LAD Tca

- **HackLength**
  - @In: drain dir. LAD Tca
  - @Out: Hack length

- **HackStream**
  - @In: Network
  - @Out: Hackstream

- **ExtractNetwork**
  - @In: drain dir. LAD Tca
  - @Out: Network

- **Mode, Threshold**

- **BasinShape**
  - @In: depitted dem
  - @Out: Subbasins

- **NetNumbering**
  - @In: drain dir. LAD Network
  - @Out: NetNumbered Subbasin

- **Pfafstetter**
  - @In: drain dir. LAD depitted dem
  - @Out: Pfaf net F.C.
Help me!

Scared Enough?
JGrass-NewAGE 1.0: Sorry, I forgot a piece

You need the same data!

In this case, you are lucky. We used open data ... but this is not always the case.
Assuming you are bold and smart

This will take for you at least a **couple of years** for putting all the parts together for your own and just following verbatim the indication you can get from the paper. (We think we put all of the information in the paper necessary: but, you know, this is practically unverifiable)
Our paper is theoretically reproducible … but practically not: it requires a trained person to do it, having all the right tools in her hands (including programming skills)…

If you are a Ph.D. student that starts from the scratch you cannot afford it! Almost nobody goes back and repeats something that's already been published, though.*

So are we doing science or just cheating of doing science?

Theoretically reproducible ... but practically not: means that theoretically we are doing science, but practically not?
This is even worse than believed in today sciences

Because of the massive use of computation.

Computation is now central to the scientific enterprise and it adds a further layer of complexity to the science visible in papers.

Some paper that comes out from computation are out of any control
“Computation is now central to the scientific enterprise, and the emergence of powerful computational hardware, combined with a vast array of computational software, presents novel opportunities for researchers. Unfortunately, the scientific culture surrounding computational work has evolved in ways that make it difficult to verify findings, efficiently build on past research, or even apply the basic tenets of the scientific method to computational procedures.”

By Victoria Stodden, Jonathan M. Borwein, David H. Bailey, SIAM news

To keep out any doubt

I decided to make public any code (any source code, actually) under a copyleft license (GPL v 3.0). See at:

http://abouthydrology.blogspot.it/2015/03/jgrass-newage-essentials.html

So we reduced a couple of years of work to three months (with instructions)
An we plan to make our work Replicable in any paper not only Reproducible but we are not alone.
Editorial: The publication of geoscientific model developments v1.0

one of the EGU’s Open Access journals, i.f. 3.6
Reproducible Research in Vadose Zone Sciences

T.H. Skaggs,* M.H. Young, and J.A. Vrugt

A significant portion of present-day soil and Earth science research is computational, involving complex data analysis pipelines, advanced mathematical and statistical models, and sophisticated computer codes. Opportunities for scientific progress are greatly diminished if reproducing and building on published research is difficult or impossible due to the complexity of these computational systems. Vadose Zone Journal (VZJ) is launching a Reproducible Research (RR) program in which code and data underlying a research article will be published alongside the article, thereby enabling readers to analyze data in a manner similar to that presented in the article and build on results in future research and applications. In this article, we discuss reproducible research, its background and use across other disciplines, its value to the scientific community, and its implementation in VZJ.

Abbreviations: NIH, National Institutes of Health; RR, Reproducible Research; VZJ, Vadose Zone Journal.

A hallmark of the scientific method is that research results must be reproducible. Although the reproducibility requirement has always existed, technological advances over the last few decades have changed the way science is practiced and communicated, creating for researchers and publishers new opportunities and challenges with respect to openness and reproducibility.
Make our source code open source (actually not necessary just the executable could serve the scope) and available through GitHub

https://github.com/
2

a) **Documenting our code** as best as possible, according to a standard format (still to define … but we are working on it).

b) Documenting **our algorithms**.

c) Using the **Object modelling System v3** (David et al., 2013, Formetta et al, 2014)
Using the appropriate building tools

```gradle
dependencies {
  compile 'org.slf4j:slf4j-api:1.7.21'
  compile name: 'fizzbuzzer'

  // Using the appropriate building tools
  compile group: 'javax.media', name: 'jai_core', version: '1.1.3'
  compile group: 'javax.media', name: 'jai_cyst', version: '1.1.3'
  compile group: 'javax.media', name: 'jai_imageio', version: '1.1.3'

  // https://mvnrepository.com/artifact/org.jgrasstools/jgrass
  compile group: 'org.jgrasstools', name: 'jgrasstools', version: '3.7.8'
  compile group: 'org.jgrasstools', name: 'jgrasstools', version: '3.7.8'
  compile group: 'org.jgrasstools', name: 'jgrasstools', version: '3.7.8'
  compile group: 'org.jgrasstools', name: 'jgrasstools', version: '3.7.8'

  textCompile group: 'junit', name: 'junit', version: '4.8'
}
```

https://gradle.org/
Use standard names for hydrological variable. For instance use the Basic Model Interface standards BMI

http://csdms.colorado.edu/wiki/BMI_Description
Using Authorea for uploading complementary material and documentation.

https://www.authorea.com/

You can use also Jupyter or Beaker
A strategy to make our paper replicable

Using as much as possible **Open Data** in our research and making available openly our data*.  


* [http://www.nature.com/sdata/](http://www.nature.com/sdata/) is a Nature Journal
Other (more valuable experiences)

The R community

(https://cran.r-project.org/web/views/ReproducibleResearch.html)
Request for Review: ESIP's Software Guidelines

ESIP (the Federation of Earth Science Information Partners) has been developing research code/software guidelines for the earth observation and geosciences communities, and would appreciate feedback on the current draft before the end of October. If you have suggestions or feedback for:

- Interoperability
- Jupyter notebooks as code or documentation
- The proposed progression model
- Sustainability or adoption/reuse

please chime in. (Look in the right margin of the browser page for the hypothes.is controls that will let you add and view comments.) As a taste of what they’re doing, here’s a table of their stakeholders’ user cases and desired outcomes:

http://software-carpentry.org/
R based reproducible research on Coursera

Reproducible Research

Part of the Data Science Specialization »

Learn the concepts and tools behind reporting modern data analyses in a reproducible manner. This is the fifth course in the Johns Hopkins Data Science Specialization.

https://www.coursera.org/course/repdata?from_restricted_preview=1&course_id=973513&r=https%3A%2F%2Fclass.coursera.org%2Frepdata-012%2Fclass#
Bancheri M. et al., Research reproducibility and replicability: the case of JGrass-NewAge

**Source code**

**Project examples**

**Community blog**

**Documentation**


R.Rigon, M.Bancheri, F. Serafin, W.Abera, G.Formetta
Become a Reproducible Research Warrior!

The \( \mathbf{R}^2 \) stairs

Do not wait! Make your stuff available on the Web (whatever format) under an open license*.

*Same as Tim Berners-Lee - Waiting to have it in better shape will delays the publication forever, and your contribution will be lost (like tears in rain): [http://5stardata.info/](http://5stardata.info/)
Make it available with documentation (e.g. a README file for any data set and for any model)
For yourself

The $R^2$ stairs

Provide examples of runs, and give some reference. Structure your documentation. Include figures and their making.
Use URLs and providers like Github to store code and data, so people can point at your stuff, and browse it freely.*
The R² stairs

Maintain a user group (and answer to questions when asked). Provide any run you do on the web with the appropriate metadata.

**: https://earthsystemcog.org/projects/es-doc-models/**

***http://abouthydrology.blogspot.it/2014/10/naming-things-in-hydrological-models.html***
Maintain a user group (and answer to questions when asked). Provide any run you do on the web with the appropriate metadata.

Use URLs and providers like Github to store code and data, so people can point at your stuff, and browse it freely.

Provide examples of runs, and give some reference. Structure your documentation. Include figures and their making.

Make it available with documentation (e.g. a README file for any data set and for any model).

Do not wait! Make your stuff available on the Web (whatever format) under an open license.

* http://5stardata.info/

** https://earthsystemcog.org/projects/es-doc-models/

*** http://abouthydrology.blogspot.it/2014/10/naming-things-in-hydrological-models.html
See Also

Science Code Manifesto

| Manifesto | Discussion | Endorse | Resources | About |

Software is a cornerstone of science. Without software, twenty-first century science would be impossible. Without better software, science cannot progress.

But the culture and institutions of science have not yet adjusted to this reality. We need to reform them to address this challenge, by adopting these five principles:

**Code**
All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.

**Copyright**
The copyright ownership and license of any released source code must be clearly stated.

**Citation**
Researchers who use or adapt science source code in their research must credit the code's creators in resulting publications.

**Credit**
Software contributions must be included in systems of scientific assessment, credit, and recognition.

**Curation**
Source code must remain available, linked to related materials, for the useful lifetime of the publication.

http://sciencecodemanifesto.org/
In conclusion

- Research must be reproducible
- In many case it would be better it is replicable
- Making our research replicable can be an advantage
- It can favour the progress of science
- **Do not be shy: share your research**
- Nobody is going to hurt you

Find your own way to Reproducible Research
Find this presentation at


Other material at

http://abouthydrology.blogspot.com
For the web references, see the slides.


References right to the point

