Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-215-RC5, 2016

Hydrology and Earth System Sciences
Discussions

Interactive comment on "HESS Opinions: Repeatable research: what hydrologists can learn from the Duke cancer research scandal" *by* Michael N. Fienen and Mark Bakker

S. Geiger (Referee)

s.geiger@hw.ac.uk

Received and published: 13 June 2016

The paper by Fienen and Bakker is an important opinion piece that argues that data analysis, and more generally scientific findings, in hydrology need to be both, repeatable and reproducible (providing a careful distinction between the two terms), to avoid future scientific scandals such as the infamous Duke cancer study. The authors view scripting languages, published communally through an Open Source approach, as a key enabling technology that can help to support repeatable and reproducible science. They further argue that academia needs to develop a proper reward structure where the publications of well-tested and carefully documented code is equally recognised as the publication of a peer-reviewed paper.

I think that this opinion piece is a timely contribution to HESS, and part of a larger groundswell that calls for repeatable and reproducible science across the entire research field of porous media flow. For example, a forthcoming editorial statement in the SPE Journal will make similar arguments for repeatable and reproducible science in the field of petroleum engineering.

I think this opinion piece should definitely be published but would welcome if the authors take a more differentiated view that also addresses some of the wider challenges related to generating repeatable and reproducible science by considering (some of) the following points:

1. While the Duke Cancer study is one "good" example of a high-profile scientific scandal, there are unfortunately other high-profile examples where experiments, data, and analysis were not properly documented and publications selectively used the results to satisfy foregone conclusions, and in some cases even fabricated the results. The largest scientific fraud in the field of physics involving the German wunderkind Jan Hendrik Schoen at Bell Laboratories at the turn of the millennium springs to mind. Improper use of data is a much wider issue that goes beyond the field of "omics", although the consequences may be very different: Cancer research directly impacts human life and inappropriate use of data could be a life-or-death decision; the Schoen scandal involved a research field were scientific breakthroughs could have been rewarded with a Nobel Prize. Hydrology, or more generally porous media research, normally does not have this kind of impact, for better or worse. Where the stakes are high (e.g. when simulating the performance of hydrocarbon reservoirs to help supporting drilling multi-million dollar wells), stakeholders normally have strict protocols in place that aim to assess the quality of the work and mitigate risks. However, what all scientific scandals have in common is that they adversely impact the scientific careers of innocent bystanders, especially the careers of PhD students and postdocs who repeatedly and unsuccessfully try to reproduce the fraudulent results as a basis for starting their own research.

2. Scripting and Open Source code development is just one means to reach the goal of reproducible and repeatable science. Equally important is the stewardship of the underlying data, which may have been collected as part of a larger interdisciplinary study and is published along with the analysis. The Research Councils UK (RCUK) are now enforcing that UK universities develop policies that guarantee that all data from all publicly funded research appropriately managed and archived at the university that originated it (for example, see the guidelines for the field of engineering and physical sciences at https://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatama
Only by making the underlying (experimental and/or field) data available as well as the analytical tools (scripts and/or code) available, science will become truly reproducible and repeatable. An important question, however, then becomes who will manage the data; as we are currently finding in the UK, this is often not an easy and cheap task to accomplish. A repository like GIT may provide a convenient solution, but who will manage the repository once the PhD student or postdoc who wrote the scripts has moved on or the funding has ceased? And how should we store gigabytes or even terrabytes of raw data?

3. I do not believe that the use of GUIs or, indeed, the use of commercial software is the key problem when it comes to reproducible and repeatable science. As it should be common with laboratory experiments, simulation experiments should also maintain a "lab book" that documents how certain simulation experiments were run (e.g. which input parameters were used in a simulation as well as the decision making process that has led to the particular choice of parameters, which may well include a large number of "failed" simulation attempts – but see the above comments on how to manage potentially vast quantities of data) . Such metadata may be documented directly in the script itself or electronically (including the version of the software that was used). I would further add that the generation of data, be it in the lab or field, can be the main focus of a study and the analysis is only a minor part of the research that relies on existing software packages with an easy-to-use GUI; not every student will become an expert

in experimental research AND scripting languages.

4. Although I am a big advocate of Open Source code, practice often shows that Open Source code is not as well documented as it could, and that a simulation with certain parameters that was running in, say, version 1.1 no longer runs in version 1.2. The nature of this is perfectly understandable: The code is often developed by students and postdocs who have other targets than documenting and testing the code to the level where it can be readily run by a large number of scientists. The authors rightly state that academia needs to change its reward system to recognise significant code developments, but in the current highly competitive "publish or perish" climate I am not overly optimistic that this change will happen anytime soon. To this end, I would hence welcome if the authors actually discuss some examples of best practice in Open Source code development such as the Open Porous Media initiative, and herein in particular the Matlab Reservoir Simulation Toolbox, where all input scripts and input datasets are provided along with the scientific paper (see http://opm-project.org/).

5. A possibly contentious issue surrounding the provision of Open Source code is plagiarism: Where code is made available at the time of publication, it becomes much easier to repeat an analysis – perhaps with a negligibly small change – and pass the "research" off in another (perhaps less well-read) journal as your own work. As stated in the editorial of Water Resources Research a few years ago, and as I have witnessed myself as the associate editor for two journals, plagiarism is on rise. In theory, publishing code and data along with the paper provides could offer evidence as to where the original research was conducted, but still plagiarism it is probably the most frequent counter argument that I have heard when it comes to making code available via an open source route.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-215, 2016.