Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-215-RC1, 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

O. A. Cirpka (Referee)

olaf.cirpka@uni-tuebingen.de

Received and published: 26 May 2016

In this opinion contribution, the authors take a recent scandal in cancer research, where the processing of molecular-biological data turned out to be so intransparent that the analysis could not be repeated, as an opportunity to call for equally strict rules of conduct in hydrological data processing as has now been called upon in the wake of the mentioned scandal in cancer research. Being an opinion paper, the paper need not be balanced, however, I believe that the authors should be adviced to (a) discuss counter-arguments and obstacles, (b) rethink whether repeatability is really more important than reproducibility, (c) broaden their perspective to reach out to hydrological researchers who are not modelers, and (d) separate issues of transparency from the

call of open-source programing.

On page 2, lines 9-11, the authors state: "Omics is a powerful field enabling cancer researchers to use large datasets to explore the efficacy of cancer treatments based on patient data and statistical modeling prior to conducting trials in humans." While this is not wrong, it gives the erroneous impression that omics (summarizing (meta)-genomics, measuring DNA, (meta)-transcriptomics, measuring messenger RNA, proteomics, measuring proteins, and metabolomics, measuring metabolites) was specific to cancer research. This is not the case. The mentioned molecular-biological techniques dominate all modern life sciences, including environmental microbiology, where it actually may come into contact with hydrological research. Modern omics-technologies, such as quantitative polymerase chain reaction (qPCR) as example for genomics, make it possible to test for 1000 genes with a single test. In chromatography coupled to high-resolution mass-spectroscopy, as used for metabolomics studies, thousands of molecular features are measured in a single sample. Analysis of some omics data is not possible without bioinformatics tools, operated by computer scientists rather than biologists and medical researchers.

Just to put this into its own context: Modern life sciences would be impossible without these techniques, and they are accompanied by other data-intensive technologies, such as practically all imaging techniques. Medical imaging and geophysical surveying are based on essentially the same measurement principles, some of which producing gigabytes of raw data. Life scientists are typically not considering the raw data at all and work with the images instead, which are inversion results prone to inversion artifacts, and which are often "enhanced" to highlight the features of interest (such as a brain tumor in an MRI scan).

The authors claim that hydrologists are more or less in the same situation. I would like to question that. Most hydrologists have a strong quantitative physical background. I would not take a hydrologist seriously, who takes a satellite-based soil-moisture map for granted. Everybody in the community knows that this is an interpreted satellite product

rather than a direct measurement of volumetric water content. In general, most hydrologists are closer to the physics of the measurements and more aware of the pitfalls of their quantitative interpretation than most life-scientists. In physical sciences, a measurement without an error analysis is considered useless, which is an attitude often lacking in life sciences and biogeochemistry. As such, there may be better chances to call for repeatable research, in the sense defined by the authors, in hydrology than in life sciences because the raw data are closer to our own way of thinking. This does not mean that we are better people or better researchers; it may actually mean that most our measurements are still so primitive that we ourselves can directly understand them.

A second very important set of differences between research on human health and hydrology is the question what is at stake and what are the commercial interests. Wrong interpretations by human-health researchers may eventually cost human lifes in a fairly direct sense. Despite all the yada yada regarding the importance of hydrology for human well-being, erroneous smoothing of a hydrograph most likely will not have the same direct effect. Conversely, there is also much less pressure on hydrologists to give the interpretation of their data a certain twist, because there is no equivalent to the pharmaceutical industry investing hundreds of million dollars to develop a drug. There are good reasons why the US Food and Drug Administration (FDA) has defined extremely strict codes on planning and documenting every decision made in developing, manufacturing, and distributing drugs. But it comes at the price of endless red tape, which honestly I don't want to see in hydrological research. For that I am freely willing to confess that my research is not directly saving lifes.

There are of course hydrology-related studies supporting high-risk decisions of high societal relevance where there is lots at stake. Analyzing potential sites for a nuclear-waste repository would fall into that category. Following an FDA-like protocol in such applications is highly recommendable, and most certainly done, simply because any decision taken will sooner or later end up in court anyway. Most hydrological studies,

however, don't fall into that category. This allows us to be somewhat more relaxed.

The authors try to work out differences between reproducibility and repeatability and make the bold statement that repeatability is more important. I like to doubt that. While the difference appears quite semantic to begin with, the authors highlight what they want: Given the same data, everybody should be able to follow the same steps of data processing to come to the same conclusions. But is lacking repeatability in this sense really the biggest problem?

I would claim that hydrologists can learn a lot from life scientists regarding the importance of reproducibility rather than repeatability. Andrew Binley tried to explain to hydrogeophysicists at a recent AGU fall meeting what he has learned in a collaboration with crop scientists where they used geophysical methods to test whether certain crop genotypes are more efficient in their water use than others, which involved geophysical imaging of soil moisture (Shanahan et al., 2015). While geophysicists (like hydrologists) like to get the last piece of information out of a single measurement campaign, life scientists (including cancer researchers, crop scientists, and ecologists) don't trust single experiments at all. You need several plots of the same experiment, you need several sites, and in a comparison study you have to make sure that either all other factors influencing the outcome of your experiment are identical or that the sample size is large enough for randomization, before you draw any conclusions. All of this has to do with reproducibility of experiments rather than data processing. If you cannot confirm your findings in a repeated experiment, you may have interpreted a freak event. Biologists take extra classes on experimental design, which discuss the statistical basics of reproducible research of complex systems. By contrast, there is a tremendous tendency in physical environmental sciences (including hydrology) to over-interpret single experiments. I don't think that coming to the same conclusions if you re-analyze the same data set is sufficient at all. Frankly, I also don't think that the latter has not happened in hydrology: There are dozens of papers on analyzing the same tracer tests in the aquifers of Borden, Cape Cod, and at the MADE site in Columbus, Mississippi. A

single experiment on mixing-controlled reactions on the bench-scale (Gramling et al., 2002) has been re-interpreted over an over again, whereas hardly any attempts were made to repeat the actual experiment.

So quite in contrast to the authors, I believe that reproducibility is a major concern in physical hydrology, which actually has consequences on the way how we do research: Should we intensify and refine measurements (and their associated interpretations) at single sites to gain the most mechanistic system understanding, or should we spread out cheaper measurements over many sites to gain reproducibility and confirmation of findings, possibly at the cost of refraining from a fully mechanistic process description? In inter-disciplinary research collaborations, these questions are heatedly debated, and putting down the necessity of reproducibility does not truly help.

I agree with the authors that transparency in data processing is mandatory. But this may be achieved by other measures than enforcing everybody to use open-source codes written in a free-of-charge scripting language. I want to highlight that transparency should start with the selection of the data. As a frequent reviewer, I am often bored by reading long explanations that the time series 3, 7, and 11 out of 15 time series measured were excluded for lengthy explained reasons, which is followed by an extended disucssion of outliers. However, much worse would be dropping the three time series and all outliers altogether and pretending that only 12 surprisingly consistent time series were taken. The same holds for measuring multiple chemical species but discussing only selected ones without mentioning the selection at all. We honestly don't know how much data has been droppped in hydrological studies, and how often this was justified (noisy probe, no flow through the piezometer, no gain of information at all, etc.) versus how often the data did simply not confirm the hypothesis put forward in the paper.

The authors are advocates of free software. But this is not the only way of guaranteeing transparency in data processing. I don't mind excel spreadsheets, if they are well documented. Often, it is absolutely sufficient to put the equations into an appendix

or the supporting informations, so that everybody can re-program the steps if wanted. Maybe more important, as a model-affine researcher I personally love matlab scripts to process data, but practically none of my colleagues in environmental chemistry, biogeochemistry, or geomicrobiology would ever do that, simply because programing is not part of their research. The approach suggested by the authors would be doomed to fail if you tried to impose it on the water-quality part of the hydrological community. For these colleagues, the quest of transparency requires different tools such as easy-to-use, flexible and yet standardized electronic lab journals and templates for excel spreadsheets linking raw data to data needed for calibration and meta information. In practice, the true intransparency lies not in handling large data streams, because people having to handle them are fully aware of data-handling and -documentation issues. The true intransparency lies in the type of research where all data still can be stored on the PC of an individual PhD student. Some funding organizations and publishers have formulated policies that data need to be stored in a repository and made be available to the public. However, for many people it's still an enigma how exactly this should be done. If there was an industry standard sold by microsoft for the management of lab data, I would not be happy that a private company makes all the money, but I would have hope that data won't simply vanish once the responsible doctoral student has left.

In contrast to the authors, I can understand strong resistance against making all source codes freely available. A specific argument against open-sorce codes relates to intellectual property rights. The development of the MODFLOW family of codes has been funded by USGS with the obligation to provide the source code. FEFLOW, by contrast has been developed by a company that has invested thousands of work hours into the development, and the return of investment relies on licencing the code. If you don't get a public-budget salary (like professors and federal administrators do), you have to sell the product of your work. If codes have been scrutinized enough by benchamrk tests etc. (which costs human resources, too), the users can rely on them without having access to the source code. So the transparency would lie in providing input files to

certified codes rather than providing the source codes.

A second reason for hiding source codes is quality assurance. Everybody can modify an open-source code, which can lead to erroneous behavior of the code. Who is ensuring that the wrongly-modified versions of the code are not disseminated? So-called community codes (like the community land-surface model - CLM, used in weather and climate modeling) require an institution that is willing to manage the official releases of new versions and perform the benchmarks. Somebody has to pay for that. The authors may believe to have reached paradise once every code is free. While I personally doubt that, it particularly is not a prerequisite for transparency in handling scientific data. So please, don't mix that up.

Let me come to my last ugly remark. If all steps of data processing was made publicly available, would the information actually be read? And - since the extra information would now be part of the publication - who would be willing to review it? As an associate editor of Water Resources Research, I once checked a matlab code provided by authors as supporting information, and I actually found an error. While on a psychologic level this gave me the illusionary ego-boost of feeling superior over the authors of that particular paper, I doubt that I would be doing that for all papers that I review or handle. Hence, the transparency by open codes requested by Mike Fienen and Marc Bakker would not guarantee that publication of erroneous analyses was prevented.

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