Consolidated Replies to Online Comments

Mike Fienen and Mark Bakker

August 25, 2016

1 Introduction

In this document, we repeat comments made by the five reviews and present our replies. In the replies we also indicate where changes are made in the manuscript. Attached following that discussion is a marked-up version of the manuscript highlighting changes that were made in the revision process.

2 Reply RC1

We are pleased to have the detailed review from Dr. Olaf Cirpka of our opinion piece. Dr. Cirpka raises some important issues. There are a couple issues that warrant revision of our text to clarify our intent and meaning, while there are a few others we do not agree with. We have distilled Dr. Cirpka's extensive comments to a few salient issues which we address in turn.

1. *"The authors take a recent scandal ... as an opportunity to call for equally strict rules" requiring scripting of data analysis and modeling in Hydrology as was done by the Institute of Medicine:* Indeed, we suggest that repeatability is as much an issue in hydrology, but we do not suggest to define a set of rules that need to be strictly enforced. Rather, we suggest that the techniques mandated by IOM have relevance to hydrological science (and many other fields).

2. *"discuss counter-arguments and obstacles":* As an Opinion Piece, we prefer to rely on the online Interactive Discussion (such as this one) to form the other side of the discussion.

3. *"rethink whether repeatability is really more important than reproducibility":* This is an important topic and we regret that Dr. Cirpka interpreted our brevity on reproducibility as dismissal of its importance. We will expand the paragraph distinguishing repeatability from reproducibility to better explain why our focus is soundly on repeatability in this piece. Most hydrological studies are not fully controlled experiments, as in many other fields, but take place in the field, which means that results strongly

depend on the specific field site and the specific circumstances (temperature, rainfall, river discharge, etc.), which cannot be fully reproduced. Only a few hydrological experiments are truly reproducible in the classical sense, for example at the Borden and MADE sites mentioned by the reviewer. In contrast, many hydrologists are trying to understand and predict natural systems through measurement and modeling rather than performing controlled experiments that can be reproduced by colleagues.

4. *"broaden their perspective to reach out to hydrological researchers who are not modelers":* The paper extensively discusses both data analysis and modeling, and our opinion equally holds for projects that do data analysis without modeling. But, in a sense, we are all modelers. Experimental campaigns are commonly followed by data analysis, which includes some kind of modeling, even if it is "just" trying to determine a trend.

5. *"separate issues of transparency from the call of open-source programing":* We are (as our opinion) advocates of open-source programming. However, our focus is meant to be on the techniques of recording steps taken in analysis than on open-source software particularly. But, as we state in the paper, the two are related: "Without the availability of an executable code, the simulations can still not be repeated and without the availability of the code itself, the computational steps in the code cannot be understood and scrutinized." Nonetheless, we have tried to make it clearer that open-source programming is a factor but not the entire answer.

6. *"it gives the erroneous impression that omics was specific to cancer research":* We regret that Dr. Cirpka got this impression, but we feel the context is clear and indeed it was omics that was at issue in the cancer scandal. We propose to highlight that we are speaking of omics in the context of cancer research by adding words in bold type in the sentence "The fields of omics **as used in cancer research** and hydrology may seem as completely unrelated..." which we assume was the source of Dr. Cirpka's objection.

7. *Objection to the analogy between hydrology and life science on two accords. a) physical scientists have more connection to their raw data so the repeatability steps may be applied in hydrology, and b) there is less pressure to twist results in hydrology as lives are not at stake in the same way they are in pharmaceutical research.* On the first accord, whether hydrologists have more connection to their raw data is a judgement call that we are not able to make, but we agree that repeatability steps are equally valid for hydrology (and other sciences). On the second accord, we often hear similar arguments that hydrologic findings have less at stake (at least in the short term) than national security, health, etc. While this is true, we don't agree that such a case serves as an excuse for us (as hydrologists) to be less robust in conducting our science. If we wish to take ourselves seriously (and be taken seriously by others) we need to hold ourselves to

a high standard. However, and harkening back to point #1, we are not advocating for strictly enforced rules, but rather suggesting best practices that, ones adopted by the community, will form a standard that others want to comply with as best practices.

8. *"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.":* We will revise our language to make it clear that free-of-charge is not the principal criterion we are advocating, although it makes it obviously a lot easier to repeat a modeling effort when the code is free. However, Dr. Cirpka also discusses that he is bored reading about why certain data are selected and others rejected in papers, but he adds: "...much worse would be dropping the three time series and all outliers altogether and pretending that only 12 surprisingly consistent time series were taken." We agree completely! That's why scripts are much better than spreadsheets. In a spreadsheet, often one simply deletes data that are not carried forward in analysis, but in a script that operates on data with auditable provenance, every such decision to drop a member of a dataset (or perhaps make a judgement call about data quality, such as a unit or datum conversion) can be documented.

9. *"The authors are advocates of free software. But this is not the only way of guaranteeing transparency in data processing. I dont mind excel spreadsheets, if they are well documented.":* The paragraph following this comment has a fair bit to unpack. First of all, Excel spreadsheets certainly play a role, but with complicated data analysis, often the order of operations and details about the calculations is difficult to fit properly in the confines of a spreadsheet. We advocate scripting languages in this piece, but even Fortran or C source code can contain detailed comments that might make it easier to follow the progression of calculations made to process data. RScript, MATLAB, and Python notebooks are much better than that. We also strongly disagree that placing the equations in an appendix is "absolutely sufficient." It is one thing to work out equations properly and entirely another to actually implement them correctly. Many errors are not as egregious as using the wrong equation, but are things like a unit conversion not made (that's how to crash a spaceship into Mars!) or even a wrong sign. The scripts contain not only notes but the actual calculations and since the scripts have been written anyway, why not make them available? We totally agree that archiving data in a meaningful way is a challenge, but that's a weak excuse not to try. Forcing PhD students to be more transparent and create repeatable work would be a service to the community going forward. Finally, it's disappointing to hear Dr. Cirpka thinks that enforcing such requirements on the water-quality community is doomed to fail. Indeed, the omics research related to cancer research that motivated this piece has more in common with water-quality than

with quantitative hydrology. It seems like an excuse not to try rather than a solution.

10. *"On free software":* Just three additional points on free software to address here. First, to imply that because a code has been paid for it is bug-free is naive - patches are constantly made to commercial software partly because bugs are found by users. The inability for a user to inspect the code is an impediment to quality assurance. Sure, not many users will want to look at the code, but enough do so that it can enhance quality. That said, we do not mean to imply that all software must be free. FEFLOW is a solid code, as are many in the petroleum industry which are expensive and commercial. But, it's simply not true that "If codes have been scrutinized enough by benchamrk*[sic]* tests etc. (which costs human resources, too), the users can rely on them without having access to the source code." Scrutiny and analysis by users is key to any software development and hidden bugs in proprietary code can go unnoticed much longer than when external scrutiny is ongoing. Second, open-source and free software is not the same as unmoderated community software. All three exist in various combinations, but many open-source projects are maintained by a team (who often sell training and consulting services rather than shrinkwrapped boxes of software to pay the bills). In fact, many open-source codes can be purchased so are not technically free. Many tools have been developed for open software development to enforce rigorous testing and quality checking by a limited team. Git makes this available through putting the code online so anyone can modify their own copy, but the lead developers decide which proposed changes get accepted.

11. *Finally what the reviewer called an "ugly" comment on whether anyone will read the data analysis scripts:* It is not necessary that all the scripts be reviewed by journal editors and peer reviewers. It is fine in our view to assume that the calculations are correct while reviewing a paper. But...having such scripts and notes available to the reviewer can be valuable when results don't make sense or an error is suspected. They need not be read and scrutinized in every case to be valuable. Certainly many cases have little at stake and will not be reviewed, but in some cases the audit may be crucial.

3 Reply RC2

We appreciate the response from Dr. Wolfgang Nowak on our Opinion Paper. We are glad that he interpreted our intent to highlight the value of a documented path from original data, through analysis and modeling, to forecasts or model results. We appreciate that Dr. Nowak recognized we were not simply implying open-source software was the answer. Nonetheless, in response to the other review, and at the suggestion of Dr. Nowak, we will revise the paper to make that clearer and hopefully avoid the misunderstanding of our conclusions as "use open source and all is fine."

The second recommendation from Dr. Nowak was to disclaim the fact that we are addressing only one issue (data provenance and auditable pathways through data and analysis) but there are others that can enhance transparency. This is a good point, and we will revise the paper to incorporate a bit more context in that way.

4 Reply RC3

It is good to see a robust discussion developing around this opinion piece. Dr. Nowak agreed with our point that archiving data and processing through scripting is worthwhile even if not everyone will examine every step. We are glad he sees this point.

Dr. Nowak also raises another point that repeatability transcends a need for transparency and is useful for future researchers to revisit work done in the past. This is an excellent point! We respond here with a bit of discussion but will also briefly highlight that issue in the revised manuscript. In this context, however, we would like to reply more substantially.

When the public pays for research At the US Geological Survey, where the first author is employed, there are policies in place in groundwater hydrology (and rapidly expanding to other research) to maintain rigorous archives of models and data analysis. These archives are designed to make it possible for the taxpayers to obtain copies of models that are operational and match the results published in papers and USGS technical reports. These archives require a fair bit of work to assemble, require a peer review which takes time and energy, and many languish for years without any interest—until there *is* interest. Recent projects have focused on automatically scraping the archives to assemble, for example, all published USGS models in a large region of the United States. Without maintaining the archive, such new meta-projects would be impossible. Many consultants and researchers also, upon learning that USGS has created a model of an area they are interested in, request the archive and thus have a working model and (as we stated in the manuscript, possibly of more interest) the supporting data to launch their new project from. Using a scripting approach to also make the steps of data processing and analysis available serves two purposes: it makes the process of archiving easier and more transparent, and it provides the context of interpretations made by the original researchers as they evaluated the data and made the model.

Marketing We also agree that trying to enforce more rules and extra work is likely to be met with skepticism and disdain (indeed, the authors have experienced that both in conversations with colleagues and in discussion in this forum). However, Dr. Nowak makes a good point that even voluntary standards and protocols are of value if researchers can market their work as following them. Consumers of the results–be they other researchers, consultants, or the public–can demand higher standards even as they remain voluntary.

It is true that we are wandering into a different topic than the original intent of the piece, but these issues are useful and important so we will make mention of them briefly in the revisions to the manuscript.

5 Reply RC4

We thank Dr. Bellin for contributing to the discussion about our opinion piece. We are pleased that Dr. Bellin generally found our comments of interest and we welcome the suggestions for potential improvement of our presentation.

Considering uncertainty We are definitely advocates of considering uncertainty in all scientific endeavors. However, we have the opinion that discussion of uncertainty is really a topic in itself and not closely enough related to our main topic to add extensive discussion of it in the opinion piece. However, we will make strides to clarify (also in response to other comments) that repeatability and reproducibility are different. What Dr. Bellin identifies as a shortcoming by not accounting for the role of parameter and epistemic uncertainty really points to the difference between the two. Being able to repeat analysis and modeling with a specific parameter set may seem trivial since it doesn't consider the uncertainty of the process, but it is a necessary and often difficult to accomplish step! Even stochastic approaches should be repeatable and ensembles of parameter sets and resulting forecasts can be carried forward using scripts. This does not guarantee reproducibility in the case of epistemic uncertainty as it influences subjective decisions about the data/models, as another group with another model may come to different conclusions. If this is the case, it is crucial that the published results can be repeated, as it can at least be concluded that the difference is not due to errors in the published results but (likely) due to epistemic uncertainty. This highlights again that reproducibility is still an important issue that is not given enough attention in the hydrological sciences. We will clarify this in the revised version of our paper and highlight the importance of repeatability in the case of epistemic uncertainty.

The more general issue is to include more details than "letters" style articles Indeed, there are multiple aspects to how more detailed information leading to greater repeatability can be incorporated into scientific discourse. This was also an issue raised by Dr. Cirpka. However, we chose to highlight the response of the medical research community to the Duke Cancer Research scandal, being to require not just detail in writing, but an executable path through analysis via scripting. We use this as an example rather than insisting on this as the only solution. We are glad that the result has been a vigorous discussion so far and we will incorporate more about the general issue of repeatability in the revised manuscript.

Scarcity of hydrologic data It is true that the example from omics often are cases with large datasets that must be trimmed while in hydrology data are often scarce so trimming is less an issue. We can clarify this in our paper. However, the analogy between the fields is more basic in our view. Whether the issue is trimming a large omics dataset or interpreting noisy and sparse hydrologic data, in both cases subjective decisions must be made about suitability of data. Since they are subjective, other researchers must be able to understand, assess, and, possibly, overrule such interpretations. By clearly documenting them in a scripted path through the analysis, other researchers can change, add, or subtract their interpretations of the data and rerun the analysis. Such transparency can also enhance the level of collaboration Dr. Bellin hopes for. Using tools that are freely available further enhances that ability.

6 Reply RC5

We thank Dr. Geiger for his support of our Opinion Paper and for his suggestion on how to improve it. Dr. Geiger raises the following five points:

Other scandals There are unfortunately many other scandals on scientific studies where things went awry. The Netherlands (the home country of the second author), the scientific community was shocked in the past few years by a high-profile social scientist that had made up all kinds of data, which indeed had detrimental effects on his PhD students and the credibility of the scientific community. Every country seems to have its own high-profile cases, but many of these cases concern deliberate activities to falsify data. Such cases are very difficult to catch if done 'well' and are not the topic of our Opinion paper, and we will revise our paper to indicate that. We have built our Opinion Paper around the Duke Cancer Research scandal, as this is a prime example where the researchers did not deliberately falsify data, but published results that could not be repeated as they were based on a few questionable choices. The protocols developed by the Institute of Medicine are intended to make it possible to repeat published results, and we were inspired to address similar practices in the field of Hydrology.

Stewardship of underlying data The issue of how to store and make data and code available is an important one. In the United States, this has been recognized at the highest level (see White House memorandum referenced in the Opinion Piece). It can certainly be cumbersome to follow rigorous data-handling protocols with large datasets but the reward is large in the future and there is an obligation for the public to have access to data that society pays for. Our main point in this work, however, picks up where the data stewardship leaves off. Documenting the path from original data through analysis and potentially forecasts is the context in which we write.

GUIs and commercial codes vs. scripts One of the main points of our paper is that research is not repeatable when a GUI is used, unless every button-push (and the order) are recorded. Our suggestion was to record such button-pushes in a script. As mentioned in the paper, several GUIs already have this capability, which makes the analysis instantly repeatable. Such a 'spit out a script' option will make it possible for researchers to produce repeatable research without becoming scripting experts.

Open Source The documentation of Open Source codes is indeed an issue. Writing documentation is considered (at least somewhat) boring by many code developers (including the authors of this paper), but, obviously, crucial. In that respect we will emphasize this when discussing that the development of codes *and documentation* needs to be rewarded more appropriately by academia. We will think about if we can add a discussion of examples of best practices to make source codes available.

Plagiarism We are aware that some researchers don't want to make their code available, because they don't want others to change it a bit, then use it on their own problems, and then publish it. Luckily this can be regulated with the choice of an appropriate Open Source license, which gives authors the ability to specify what can and can not be done with their code. Further, the more detail of work is documented, the easier plagiarism can potentially be detected. It is indeed good to mention these issues in our paper.

HESS Opinions: Repeatable research: what hydrologists can learn from the Duke cancer research scandal

Michael N. Fienen[1] and Mark Bakker[2]

[1]U.S Geological Survey Wisconsin Water Science Center, Middleton, Wisconsin, USA
[2]Water Resources Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, Netherlands

Correspondence to: Michael N. Fienen (mnfienen@usgs.gov)

Abstract. In the past decade, difficulties encountered in reproducing the results of a cancer study at Duke University resulted in a scandal and an investigation which concluded that tools used for data management, analysis, and modeling were inappropriate for the documentation of the study, let alone the reproduction of the results. New protocols were developed which require that data analysis and modeling be carried out with scripts that can be used to reproduce the results and are a record of all decisions and interpretations made during an analysis or a modeling effort. In the hydrological sciences, we face similar challenges and need to develop similar standards for transparency and repeatability of results. A promising route is to start making use of open source languages (such as R and Python) to write scripts and to use collaborative coding environments (such as Git~~and github.com~~) to share our codes for inspection and use by the hydrological community. An important side-benefit to adopting such protocols is consistency and efficiency among collaborators.

1 Introduction

In hydrology, we face increasing amounts of data that we use to build and calibrate models, which are ultimately used for forecasts. Many subjective and interpretive steps go into the translation of data to models, sometimes referred to as the “art of hydrology” (Savenije, 2009). Hydrological science always involves judgements and interpretations so it is unrealistic to expect a single path from original data to models (Fienen, 2013). However, we can certainly do a better job of documenting our interpretations, and make it easier for others to repeat, if not reproduce, our results. The ~~seemingly unrelated fields of omics and~~ field of cancer research faced a scandal in the past decade, related to applications of omics, that offers lessons for hydrology both in the nature of the scandal and in the response by institutions involved in and overseeing cancer research.

In this Opinion Paper, we provide background about the Duke cancer scandal, highlight how repeatability and reproducibility were at the center of the solutions, and relate lessons from the scandal to the field of hydrology. Unfortunately, other high-profile scientific scandals have taken place—sometimes due to neglect, and sometimes due to intentional fraud—but we focus on the Duke cancer scandal to highlight requirements that came out of the scandal which have relevance to hydrology.

2 The Duke Cancer Scandal

In 2007, a comment on a paper in Nature Medicine pointed out difficulties in reproducing a cancer study at Duke University in the research group of Anil Potti (Coombes et al., 2007). This spiraled into "the Duke cancer scandal" that included allegations of improper methods and inflated credentials. The scandal led to an internal inquiry (Califf and Kornbluth, 2012) and later a set of guidelines by the Institute of Medicine (Institute of Medicine, 2012) highlighting the shortcomings of the studies and putting forth protocols to avoid such problems in the future. A key element of the guidelines was that an unreproducible path through data using graphical user interfaces, spreadsheets, and other such tools would no longer suffice to document the data management that necessarily precedes analysis and modeling. Computations should be "locked down" and repeatable using scripting languages so that, given an original set of data, all steps of analysis can be repeated and documented (Institute of Medicine, 2012).

The field of omics in which the Potti group performed research refers to fields in ~~biology~~ life sciences ending in "-omics", and is defined as "…the scientific disciplines comprising the study of global sets of biological molecules such as DNAs (genomics), RNAs (transcriptomics), proteins (proteomics), and metabolites (metabolomics)…" (Carlson, 2012). Omics is a powerful field with many applications in the life sciences including 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. The large datasets require processing to remove unsuitable data for a particular experiment. However, if too many data are removed in the process, overfitting can result "which unintentionally exploits characteristics of the data that are due to noise, experimental artifacts, or other chance effects not shared between data sets rather than to the underlying biology" (Carlson, 2012). As a result, the provenance of the data ultimately used for experiments is a critical element to the overall work, and the analysis path can be tedious and involve subjective judgement, especially with large, complicated datasets. Indeed, "guaranteeing robust data provenance and reproducible data management" (Califf and Kornbluth, 2012) was cited as a major recommendation by the Duke University internal inquiry. Key elements were to establish data provenance are the use of scripting languages and the sharing of code (Califf and Kornbluth, 2012).

3 Reproducible or Repeatable?

The National Institute of Standards and Technology in the USA defines "reproducible" as "closeness of the agreement between the results of measurements of the same measurand carried out under *changed* conditions of measurement" and repeatability as "closeness of the agreement between the results of successive measurements of the same measurand carried out under *the same* conditions of measurement" (Taylor and Kuyatt, 1994). These definitions are very similar, but the subtle distinction (highlighted in ~~italic~~italics) is important. For a process to be reproducible, it implies that a different group given the same data and following the same protocols will interpret and process them the same way, resulting in the same outcome as another group.

On the other hand, a repeatable process is one in which all steps are documented and the exact steps of data processing can be repeated. In fields such as omics and hydrology, where judgement and interpretation are part of the process, the goal is often more repeatability than reproducibility. For a repeatable path through the data, with judgements properly documented, another research group can evaluate each judgement and decide whether to agree with it or not.

The call for repeatable research has echoed through the computational sciences for several decades (Fomel and Claerbout, 2009), although the terms reproducible and repeatable are often used interchangeably. Peng (2011) presents a spectrum of reproducibility from solely publication of results (not reproducible) to inclusion of code, code plus data, or linked and executable code and data (full reproducibility, which should probably be called repeatability). Some journals have adopted policies to encourage repeatability of results, varying from a requirement to state where or how the data can be obtained to the submission of code that can be run to actually repeat the results, including "kite marks" that indicate which level of repeatability/reproducibility a paper achieves (Peng, 2011).

Reproducibility may be seen as a higher goal than repeatabilty. Unfortunately, hydrological field experiments are typically not made under controlled conditions such as bench experiments in chemistry or physics, but rather depend on natural variability in conditions like precipitation, river stage, and others, which may make reproducibility an elusive goal. Furthermore, many quantities are measured only indirectly and strongly depend on interpretation and inverse modeling, including remotely sensing and geophysical imaging. Other data sources are less quantitative but more descriptive, such as land use, boring logs, and outcrop analysis. Given the uncertain nature of all these data sources, it is understandable that conclusions drawn from hydrological models can be highly uncertain. Quantification of the uncertainty and problems of equifinality are very important and beyond the scope of this Opinion Paper, but they are certainly not an excuse to play down the importance of repeatability. On the contrary, repeatability seems to be the first step to tackle the problem of uncertainty and equifinality.

4 How does this relate to hydrology?

The fields of ~~omics and~~ omics—as used in cancer research—and hydrology may seem as completely unrelated, but the way data are handled and processed, and the ramifications of such data handling are actually quite similar. Hydrological and omics datasets can both be noisy and require trimming or even adjustment of some values based on quality control, interpretation, and appropriateness for the ~~modeling tasks~~ analysis at hand. Hydrological datasets come in an incredible variety of data types and formats, such as meteorological data, water levels, flow measurements, soil types, lithological logs, surface water diversions, ~~and~~ groundwater extractions, and remote sensing data. Much of this information is provided in spreadsheets, graphical documents, databases, and web-queries. At the raw data stage, the provenance is generally known but between data acquisition and creating model inputs and outputs, an unknown series of steps takes place that breaks the provenance and can hide the interpretations and judgements that took place.

Beyond interpreting the ~~samen~~ same spreadsheets and databases, many hydrologists use graphical user interfaces (GUIs) to organize and manipulate the information used in models. In a GUI, data are interpreted spatially and temporally, boundary conditions are specified, grids are generated, parameters are selected or specified, etc., while typically none of these steps can be repeated without going through the same sequence of mouse clicks, menu selections, and entries made in boxes. Repeating all these steps is tedious, prone to errors, and does not include documentation of interpretations made.

As time passes after the completion of a modeling or analysis project, the collection and interpretation of the original data is often of more lasting use than the actual model files. Modeling technology changes but the data are persistent. Access to the original data and a detailed documentation of the analysis path may be the most useful record of a project in the future (e.g. Anderson et al., 2015).

5 What can be done?

In the same spirit as the recommendations of the Institute of Medicine report above, scripting languages such as R and Python can replace much of the GUI and spreadsheet data manipulations in hydrology and hydrological modeling. Scripting languages have many features and access to specialized libraries. They also have facilities for making comments in which the subjective elements of data processing can be clearly stated. In this way, common tasks (e.g., unit conversions), specific decisions (e.g., identification of outliers), and algorithms (e.g., spatial interpolation or regularization of time intervals) can be reviewed and understood. ~~These~~ Scripting languages are interpreted so they do not need to be compiled, making them work on many different platforms easily. Tools like Jupyter Notebooks (formerly IPython Notebooks; Pérez and Granger, 2007) and RStudio (RStudio Team, 2015) provide seamless integration of written documentation and executable code. In addition to repeatability, an important benefit of these tools is increased efficiency. Note that several Python packages are specifically designed for hydrologists, for example for watershed modeling (Lampert and Wu, 2015) and groundwater modeling (Bakker, 2013; Bakker et al., 2016).

Of course, this implies that everything can be done without a GUI, but that is not necessarily true. GIS software and model GUIs provide a valuable set of tools to enable model creation and data analysis. We suggest, however, that an auditable scripting path through the GUI logic is a necessary feature of a GUI to record the many steps taken in the model-building process. For example, ArcGIS (ESRI, 2011) provides a Python application programming interface that can be used to perform any operation using a script. Furthermore, it is possible to record all the steps while clicking and selecting in the GUI as a Python script that serves as a record of the performed analysis that can be evaluated and run later, mitigating the hurdle of programming expertise for practitioners to improve repeatability in their work.

6 What else can be done?

In hydrological modeling, the documentation of a data analysis and modeling effort in a script is only one side of the coin. The other side of the coin is the model that is used to perform the computations. Without the availability of an executable code, the simulations can still not be repeated and without the availability of the code itself, the computational steps in the code cannot be understood and scrutinized. The code is also necessary to run the program on another platform than the authors used or a future version of the same platform. Harvey and Han (2002) already recognized the increasing value of open-source codes in hydroinformatics. Ince et al. (2012) make a strong case that “anything less than the release of source programs is intolerable for results that depend on computation”.

Over the past decade and a half, open-source codes have risen in prominence, as illustrated in an analysis of data analytics job postings in 2015, showing more requests for open-source coding experience than experience with proprietary analytics codes (http://r4stats.com/articles/popularity/). Unfortunately, many research groups don't make/have time to go through the extra effort to extensively test and document their code and make it available to the public. Merali (2010) suggests that more open-source software may be developed at universities when the value of such developments are rewarded more appropriately (e.g., similar to research papers in peer-reviewed journals). The road for sharing computational codes is paved by the emergence ~~of~~ collaborative coding environments such as Git (Chacon, 2009), an easy to use and free application for version control of (collaborative) coding efforts, the success of github.com, ~~a free hosting service~~ bitbucket.org, and other free hosting services for the dissemination of source code, and the availability of free and open-source compilers for many languages.

It is noted that open-source software is not always free and the open-source aspect of the code is not a panacea—indeed, proprietary software may also be used to improve repeatability. However, the more open all aspects of the analysis are, the more transparent are the findings. Both open-source and proprietary software used to enhance repeatability and transparency should be documented in enough detail to allow benchmarking and comparisons by the community to ensure consistency between documented processes and their outputs.

7 Conclusion

This paper began with a short review of a cancer scandal, which started when difficulties were encountered in the reproduction of a cancer study at Duke University. On the face of it, the fields of hydrology and omics may seem unrelated. However, both fields need to make important forecasts, whether it is the response of patients to cancer treatment, high water levels in rivers, droughts, or contaminant plume migration in groundwater systems. Both fields depend on drawing conclusions from models based on large datasets. In both fields, processing, trimming, and validating these datasets require judgement and a certain degree of art and interpretation. The specific interpretations and decisions can make the difference between high-quality forecasting and overfitting where the model chases noise in the dataset at the expense of generalization. Uncertainty in the entire data analysis process contributes to nonunique solutions in modeling and analysis. It is crucial to understand all decisions made in research that lead to a conditionally unique solution or an ensemble of solutions.

For decades, both ~~fields~~ omics and hydrology have seen a variety of techniques for data analysis and interpretation, including GUIs, custom programs, manipulation of spreadsheets, and hand calculations. GUIs and spreadsheets typically do not provide an auditable path through the process and some custom programs, once compiled, are opaque to review if source code is not provided. The result is a lack of transparency and repeatability that may ~~cover up~~ cover-up mistakes, judgements based on thinking that can change over time, and, at worst, manipulation or fraud.

The cancer research problems were encountered when one group tried to confirm the analysis and modeling of another ~~group – a~~ group—a scientific tradition that is not conducted frequently in the hydrological sciences. During the investigation of the Duke cancer scandal, it became apparent that mistakes of overfitting were made. The response of ~~the~~ academia and the Institute of Medicine was to require data provenance and documentation of data processing and modeling in ~~open-source codes such~~

~~that the~~ scripts such that all steps could be repeated independently and the analysis path through the data was well documented. These new requirements caused a major shift in approach for many researchers. The field of hydrology has not experienced such a high-profile scandal, but we must learn preemptively and adopt similar standards of transparency and repeatability for our work. ~~Open source tools~~ Scripting languages (such as R and Python) and collaborative coding environments (such as Git and online hosting such as github.com and bitbucket.org) make it practical to improve the repeatability and documentation of our research. Furthermore, transparency and reproducibility are enhanced by the application of open-source software.

Open data are also the subject of an initiative in the US at the direction of the White House (https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf). This initiative has created an environment in which researchers employed by the US government must now adhere to much higher standards of repeatability and data stewardship (similar open data initiatives are explored by the Horizon 2020 research program of the EU). Such requirements come at a cost of time and energy. To make it more realistic for such standards to be adopted, the academic systems of rewards must evolve to properly reward the extra effort required. It is up to individual scientists, journals, stakeholders, and funding agencies to demand it and create meaningful standards of repeatability. ~~We~~

It is not fully necessary to hold all research to exactly the same standard, but if we, as a community, assign value to repeatability and transparency, then even voluntary standards can gain currency. The entire community can benefit from the ability to build on each others' prior work when both data and auditable code are available. Important advances in science are made when results are confirmed or falsified in subsequent research. In any case, we must learn from the Duke cancer research scandal to prevent our field of hydrology from falling in the same trap.

Disclaimer

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Acknowledgements. The authors thank Chandra Miller Fienen (formerly BIOARRAY Therapeutics) for informing us about the Duke cancer research scandal, Yu-Feng Lin (Illinois State Geological Survey), and Randall Hunt (U.S. Geological Survey) for valuable conversations on the topic, and the online reviewers for participating in a robust discussion about the manuscript.

References

Anderson, M. P., Woessner, W. W., and Hunt, R. J.: Applied Groundwater Modeling, Academic Press, San Diego, second edn., 2015.

Bakker, M.: Semi-analytic modeling of transient multi-layer flow with TTim, Hydrogeology Journal, 21, 935–943, 2013.

Bakker, M., Post, V., Langevin, C., Hughes, J., White, J., Starn, J., and Fienen, M.: Scripting MODFLOW model development using Python and FloPy, Groundwater, doi:10.1111/gwat.12413, 2016.

Califf, R. M. and Kornbluth, S.: Establishing a Framework for Improving the Quality of Clinical and Translational Research, Journal of Clinical Oncology, 30, 1725–1726, doi:10.1200/JCO.2011.41.4458, http://jco.ascopubs.org/content/30/14/1725.short, 2012.

Carlson, B.: Putting oncology patients at risk, Biotechnology healthcare, 9, p. 17–21, 2012.

Chacon, S.: Pro Git, Apress, Berkeley, CA, USA, 1st edn., 2009.

Coombes, K., Wang, J., and Baggerly, K.: Microarrays: retracing steps, Nature Medicine, 13, p. 1276–1277, doi:doi:10.1038/nm1107-1276b, 2007.

ESRI: ArcGIS Desktop: Release 10, 2011.

Fienen, M. N.: We speak for the data, Groundwater, 51, 157, doi:10.1111/gwat.12018, 2013.

Fomel, S. and Claerbout, J. F.: Reproducible research, Computing in Science & Engineering, 11, 5–7, 2009.

Harvey, D. and Han, D.: The relevance of Open Source to Hydroinformatics, Journal of Hydroinformatics, 4, 219 – 234, publisher: IWA, 2002.

Ince, D. C., Hatton, L., and Graham-Cumming, J.: The case for open computer programs, Nature, 482, 485–488, 2012.

Institute of Medicine: Evolution of translational omics : lessons learned and the path forward, Tech. rep., Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Board on Health Sciences Policy, Institute of Medicine of the National Academies, National Academies Press, Washington, D.C., 2012.

Lampert, D. and Wu, M.: Development of an open-source software package for watershed modeling with the Hydrological Simulation Program in Fortran, Environmental Modelling & Software, 68, 166–174, 2015.

Merali, Z.: Computational science: Error, why scientific programming does not compute, Nature, 467, 775–777, 2010.

Peng, R. D.: Reproducible research in computational science, Science (New York, Ny), 334, 1226, 2011.

Pérez, F. and Granger, B. E.: IPython: a System for Interactive Scientific Computing, Computing in Science and Engineering, 9, 21–29, doi:10.1109/MCSE.2007.53, http://ipython.org, 2007.

RStudio Team: RStudio: Integrated Development Environment for R, RStudio, Inc., Boston, MA, http://www.rstudio.com/, 2015.

Savenije, H. H. G.: HESS Opinions "The art of hydrology"*, Hydrology and Earth System Sciences, 13, 157–161, doi:10.5194/hess-13-157-2009, http://www.hydrol-earth-syst-sci.net/13/157/2009/, 2009.

Taylor, B. N. and Kuyatt, C. E.: Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, Tech. rep., NIST Technical Note 1297, 1994.