AGU PUBLICATIONS

Water Resources Research

COMMENTARY

10.1002/2016WR020116

Key Points:

- The basic structure of the scientific method is widely championed as a recipe for scientific progress
- Hypothesis formulation and testing rarely correspond to the idealized model of the scientific method
- Hypothesis tests have played an essential role in spurring major advances in hydrological theory

Correspondence to:
L. Pfister,
laurent.pfister@list.lu

Citation:
Pfister, L., and J. W. Kirchner (2017), Debates—Hypothesis testing in hydrology: Theory and practice, *Water Resour. Res.*, *53*, 1792–1798, doi:10.1002/2016WR020116.

Received 11 NOV 2016
Accepted 9 FEB 2017
Accepted article online 27 FEB 2017
Published online 29 MAR 2017

Debates—Hypothesis testing in hydrology: Theory and practice

Laurent Pfister[1] **and James W. Kirchner**[2,3,4]

[1]Catchment and eco-hydrology research group, Environmental Research and Innovation Department, Luxembourg Institute of Science and Technology, Luxembourg, [2]Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland, [3]Mountain Hydrology Research Unit, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland, [4]Department of Earth and Planetary Science, University of California, Berkeley, California, USA

Abstract The basic structure of the scientific method—at least in its idealized form—is widely championed as a recipe for scientific progress, but the day-to-day practice may be different. Here, we explore the spectrum of current practice in hypothesis formulation and testing in hydrology, based on a random sample of recent research papers. This analysis suggests that in hydrology, as in other fields, hypothesis formulation and testing rarely correspond to the idealized model of the scientific method. Practices such as "p-hacking" or "HARKing" (Hypothesizing After the Results are Known) are major obstacles to more rigorous hypothesis testing in hydrology, along with the well-known problem of confirmation bias—the tendency to value and trust confirmations more than refutations—among both researchers and reviewers. Nonetheless, as several examples illustrate, hypothesis tests have played an essential role in spurring major advances in hydrological theory. Hypothesis testing is not the only recipe for scientific progress, however. Exploratory research, driven by innovations in measurement and observation, has also underlain many key advances. Further improvements in observation and measurement will be vital to both exploratory research and hypothesis testing, and thus to advancing the science of hydrology.

> There's two possible outcomes. If the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery.
>
> Enrico Fermi

> The great tragedy of science is the slaying of a beautiful hypothesis by an ugly fact.
>
> Thomas H. Huxley

> Scientists have odious manners, except when you prop up their theory; then you can borrow money off them.
>
> Mark Twain

1. Introduction

It is almost axiomatic that hypotheses and hypothesis testing are the central pillar of science. In recent years, hydrologists have made repeated calls for more rigorous hypothesis testing in their field [e.g., *Beven*, 2001, 2010, 2012; *Kirchner*, 2006; *Clark et al*., 2011; *Kavetski and Clark*, 2011; *Burt and McDonnell*, 2015; *Westerberg and Birkel*, 2015], echoing earlier calls for strengthening hydrology's scientific basis [*Dooge*, 1986; *Klemes*, 1986, 1988; *Bras and Eagleson*, 1987; *Lee*, 1992]. But what counts as a useful hypothesis, and a useful hypothesis test, in hydrology? Here we survey the broad spectrum of hypothesis formulation and testing in hydrology using a small random sample of research papers, give concrete examples of remarkable progress made in our field through hypothesis testing, and provide arguments for more widespread inclusion of hypothesis testing in hydrologic research.

The process of framing and testing hypotheses is so fundamental to the scientific enterprise that it is known simply as "the scientific method," a single recipe that is thought to be universal across scientific fields. Many practicing scientists are familiar with the broad outlines of this recipe, as described below:

First, frame your hypotheses, as tentative conjectures based on prior observations or prevailing theory. The more strongly a hypothesis links observable phenomena to underlying theory, the more scientifically

consequential it is: if the hypothesis stands or falls, the underlying theory stands or falls with it. Ideally, hypotheses are framed as pairs or multiples [*Chamberlin*, 1890] that are mutually exclusive (only one can be true) and jointly exhaustive (one must be true, because taken together, they span the universe of all possibilities). Such pairs or sets of hypotheses are particularly useful because the rejection of one of them necessarily implies support for the other(s).

Second, deduce the consequences of the hypotheses for things that you can observe or measure. If a particular hypothesis is true, what should we observe? If it is false, what should we observe? Are these two sets of predictions observably or measurably different from one another? If not, then stop: if hypotheses predict the same measurements, no data can decide among them, and empirical tests are useless.

The first two steps of the conventional "recipe" are the crucial ones, because they determine the logical connections between hypotheses, the underlying theories, and the universe of possible observations. From this point on, the recipe is relatively straightforward: Third, specify the decision rule that you will use to determine whether the observations support or refute the hypothesis. Fourth, now (and only now) collect or analyze the data, and apply the decision rule. Then modify the theory or hypothesis accordingly, and go back to step 1. Rinse and repeat.

2. Common Deviations From the Standard Model of Hypothesis Formulation and Testing

The method outlined above is simple in principle, but subtle in practical application. For one thing, the textbook scientific method assumes that theories are always tentative whereas observations are never in doubt, but the practical reality is not so simple. Observations are often ambiguous, and measurements always have errors; worse still, these errors are often incompatible with conventional statistical tools [*Beven and Westerberg*, 2011]. Thus, when a well-founded theory is challenged by poorly supported data, it can sometimes be reasonable to reject the data and accept the theory, and Bayesian inference can provide a framework for deciding when it is rational to do so [*Kavetski et al*., 2006]. But even in such difficult cases, the process of explicitly framing and testing hypotheses provides a useful strategy for scientific inquiry. . . as long as it is actually followed.

Several common forms of post hoc "cherry-picking" undermine this strategy, however. For example, many researchers try multiple statistical tests, and even multiple variables and data sets, until they find a combination of data and statistical tests that meets the statistical significance threshold (typically $p < 0.05$) required for a publishable result. This practice, known as "p-hacking," "a statistical fishing expedition," or "researcher degrees of freedom" [*Simmons et al*., 2011; *Head et al*., 2015] makes it laughably easy to manufacture a statistically "significant" result; among any 10 totally random data series, there is a 90% chance that at least one pair exhibits a spurious correlation that is "significant" at $p < 0.05$. One can make the risk of spuriously "significant" results even worse by using standard statistical methods where they do not apply (by using Pearson correlations on badly skewed data, for example, or by failing to account for serial correlation, which is nearly ubiquitous in environmental measurements).

A somewhat less obvious (and thus perhaps more serious) problem is "HARKing," or "Hypothesizing After the Results are Known" [*Kerr*, 1998]. HARKing arises when researchers formulate hypotheses based on examining a set of observations, then test the same hypotheses using the same observations, and then interpret the result as if it came from an a priori hypothesis test instead. A related problem arises with grand, overarching hypotheses that must be made precise for each individual case, and thus can be redefined to match whatever behavior needs to be explained [e.g., see *Kirchner*, 1990, 2003a]. But if any conceivable phenomena can be explained by an overarching "theory of everything," then that theory is logically empty, because it says only that "anything is possible." More generally, when one engages in cherry-picking among possible theories, one risks manufacturing a spuriously "significant" result, because almost any data set will appear to confirm some theory, as long as one has complete freedom to reinterpret that theory to fit the data. This is why one is not supposed to design one's hypothesis based on the same data that the hypothesis will be tested against. Of course, in exploratory research one will inevitably frame one's hypothesis based on the available observations; indeed, in exploratory research, that is the whole point. Thus the central problem in HARKing is not post hoc hypothesis formulation, but rather misrepresenting a post hoc hypothesis as an a priori one (or, equivalently, misrepresenting exploratory research as confirmatory hypothesis testing).

The third, and perhaps most blatant, form of cherry-picking arises when researchers adjust the scope of their analysis to exclude data that would otherwise be problematic for their favored hypothesis. As obviously illegitimate as this practice is, it does occur (we have seen it in papers we have reviewed), and is sometimes explained away as simply removing unreliable data (we have seen that too).

One might assume that researchers engaged in any of these forms of cherry-picking would be aware that they are doing it, particularly in extreme cases that amount to "torturing the data until it gives up." Unfortunately, at least from our anecdotal experience as reviewers, this does not seem to be the case. One possible explanation is confirmation bias, the pervasive psychological tendency to favor evidence that supports one's prior beliefs over evidence that refutes them [*Nickerson*, 1998]. Scientists are not immune to confirmation bias, and we are not aware of any evidence that the review process can reliably detect and counteract it. Indeed, studies have shown that reviewers have their own confirmation biases, rating papers more favorably when they confirm previously held beliefs [e.g., *Mahoney*, 1977; *Koehler*, 1993].

The published record may also be distorted by publication bias (also known as "the file-drawer problem"), in which spectacular results are readily published, and unspectacular results are instead filed away. Sometimes this publication bias arises from the peer review process itself. For example, a recent study showed that peer reviewers were more likely to recommend publication (and less likely to detect errors) when reviewing a fabricated paper that reported a statistically significant result, than when reviewing an otherwise identical paper reporting a nonsignificant result [*Emerson et al*., 2010]. The interaction between publication bias and confirmation bias may be particularly pernicious; confirmation bias may lead authors to cherry-pick in ways that inflate the apparent significance of their findings, and publication bias may make it more likely that these same exaggerated findings will be published.

3. Hypothesis Formulation and Testing in Hydrology

The foregoing discussion has addressed hypothesis formulation and testing as practiced in any branch of science. What, then, about hypothesis formulation and testing in hydrology? We searched the Web of Science database for papers published between 1991 and 2015 that referred to "runoff" or "streamflow" in their titles, abstracts, or metadata, and then tallied the fraction of these that also included words beginning with "hypoth," as a catch-all for terms like "hypothesis," "hypothesize," "hypothetical," and so forth. For comparison, we also tallied the fraction that included words beginning with "model."

The results of this analysis (Figure 1) show that: (a) the total number of papers mentioning runoff or streamflow grew roughly exponentially, (b) papers mentioning models grew as a fraction of the total, from roughly 37% 25 years ago to roughly 55% today, and (c) very few papers mentioned hypotheses, making up only 4% of the total, a percentage that remained nearly constant over this 25 year period. Of course, there are

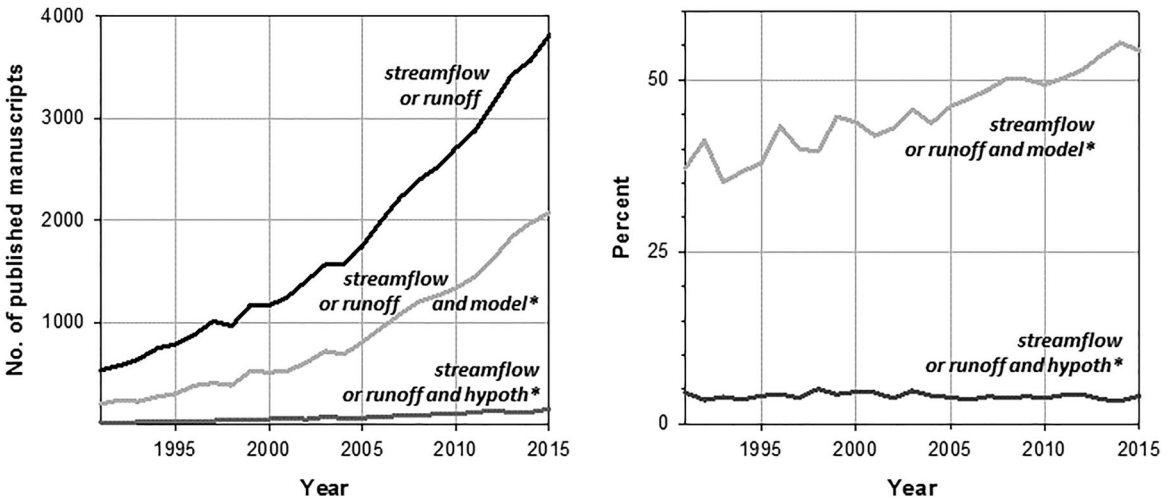

Figure 1. Hydrology papers published from 1991 to 2015 matching Web of Science queries for specific words in titles, abstracts, or metadata (left: number of published manuscripts; right: proportion of papers mentioning models or hypotheses, as a percentage of total number of papers mentioning streamflow or runoff; * indicates a wildcard).

likely to be many papers that test hypotheses but do not mention the word in their titles, abstracts, or metadata, and these would be missed in our tally. But it is also likely that many papers that mention hypotheses (and thus would be counted) do not actually test them.

We therefore randomly sampled 69 papers from the 646 published between 2011 and 2015 that mentioned hypotheses, and read them. Nearly half (34) of those 69 papers did not state a hypothesis in the sense of the "scientific method," but rather used the term to indicate a hunch or an ad hoc estimate (e.g., hypothetical climate scenarios, hypothetical reservoirs). Of the remaining 35 papers, 9 presented purely statistical tests (e.g., for trends) without linkage to mechanistic hypotheses. Only 15 of the 69 papers clearly deduced the consequences of hypotheses for things that can be observed or measured, but usually only for the authors' preferred hypotheses, and almost never for any alternative hypotheses. Thus, the hypothesis "tests" usually looked only for confirmation, leaving open the possibility that the observations were consistent with both the preferred and alternative hypotheses. Such tests are inconclusive.

4. Models as Hypothesis-Testing Platforms

A further 11 of the 69 papers defined their working hypothesis by a simulation model, and evaluated that hypothesis by the model performance against observed time series. Models generate specific predictions from mechanistic assumptions, so they would seem to be ideal hypothesis-testing platforms, and have been proposed as such [e.g., *Clark et al*., 2011; *Beven*, 2012]. The limitations of this approach need to be recognized, however.

Model predictions do not depend on a single hypothesis, but on multiple interdependent hypotheses (for multiple hydrological processes), along with multiple assumptions about algorithms, parameters (and their estimation procedures), and input data (which are subject to error). Thus if the modeled time series fails to match the data, which of these many hypotheses is falsified? Or if the model matches the data, how many of its underlying hypotheses could still be wrong, perhaps in offsetting ways? And if adding one particular hypothesis improves model performance, does that mean the hypothesis is correct, or that it is simply needed to counteract other problems elsewhere among the model's many assumptions? Fully exploring the "hypothesis space" of any realistically complex model would seem to be an infeasible n-factorial problem.

Uncertainties inherent to both observations and process representations may lead to equivalent predictions by different models (the "equifinality problem") [see, e.g., *Hornberger and Spear*, 1981; *Beven and Freer*, 2001]. The fact that different models often generate nearly equivalent predictions, particularly after parameter calibration, poses a fundamental challenge to hypothesis testing because, as pointed out in section 1 above, hypotheses that predict the same observations cannot be tested against one another.

Furthermore, models are often calibrated to fit the same data that they are tested against, or functionally equivalent data, such as different years from the same sites. This is equivalent to HARKing (because the parameters determine the relative influence of the different hypotheses in the model), with an equivalent risk of generating spuriously "good" results.

5. Hypothesis Tests and Exploratory Research Drive Hydrological Advances

One might infer that we are skeptical, or even cynical, about the role of hypothesis testing in hydrology. Far from it. Although hypothesis testing is rarely as systematic as the procedure outlined in section 1 above, hypothesis tests have nonetheless been pivotal in moving the field forward by revealing the limits of established theories, and in some cases overthrowing them outright. Here are just four examples:

1. The strongly damped behavior of water isotopes in stream water compared to rainfall [*Dansgaard*, 1964; *Sklash et al*., 1976] clearly refuted the hypothesis that runoff is composed of recent rainfall that travels rapidly to the stream via overland flow [*Horton*, 1933; *Betson*, 1964], subsurface storm flow [*Hewlett and Hibbert*, 1963], and return flow [*Dunne*, 1978]. The implications of this discovery are still being felt, as hydrologists work to understand how catchments store water for weeks, months, or even years, but release that water in minutes or hours after the onset of rainfall [*Kirchner*, 2003b].

2. More recently, the spectral scaling of tracer time series has been shown to have a different slope than the one predicted by well-mixed box models [e.g., *Kirchner et al.*, 2000; *Godsey et al.*, 2010; *Kirchner and Neal*, 2013], overthrowing a fundamental assumption underlying many conceptual catchment models.
3. A further example comes from multiple lines of experimental evidence showing strongly preferential flow in both saturated and unsaturated media [*Beven and Germann*, 1982, 2013; *Uhlenbrook*, 2006], thus challenging the Darcy-Richards paradigm (at least at the scales that it has been typically applied), and calling for a new theory of flow through structured soils [*Beven and Germann*, 2013].
4. The discovery 25 years ago that streams and their riparian trees draw on isotopically distinct pools of water [*Dawson and Ehleringer*, 1991] launched what has more recently become known as the "two water worlds" hypothesis [*Brooks et al.*, 2010; *McDonnell*, 2014], which directly challenges the implicit homogeneity assumptions underlying almost every model of catchment-scale flow and transport.

These are just four examples, and readers can probably think of many more. The common feature of field-altering hypothesis tests is that clear, general predictions are derived from the reigning theory, and then shown to be clearly inconsistent with empirical evidence. We suspect that there are many such hypothesis-testing opportunities waiting to be discovered. Making the most of these opportunities will require more attention to best practices in formulating and testing hypotheses [*Head et al.*, 2015], and a clearer separation between exploratory work and hypothesis-testing research. Recent proposals for improving hypothesis testing, such as the limits-of-acceptability approach [*Beven*, 2012], the multiple hypothesis testing framework [*Clark et al.*, 2011], intensively monitored observatories [*Zehe et al.*, 2014; *Blöschl et al.*, 2016; *Bogena*, 2016], and community hydrological models [*Weiler and Beven*, 2015], deserve careful consideration.

In closing, we want to emphasize that hypothesis testing is not the only valuable form of scientific activity. For example, hydrologists often confront practical problems that demand practical solutions. In many cases, these solutions do not require new discoveries (and thus hypothesis tests), but instead the clever application of existing hydrological knowledge.

Conversely, a lot of hydrological research is essentially exploratory in character, and is aimed primarily at generating hypotheses rather than rigorously testing them. It is essential for future progress in hydrology that this exploratory research continues to be supported, even if the immediate scientific payoff is sometimes unclear. Most of the discoveries that we have outlined above, for example, had their origins in exploratory research rather than the targeted pursuit of the theory that was eventually overthrown.

There is a recent trend among some funding agencies to require strictly formulated hypothesis tests in every grant application, which may pose an insurmountable hurdle for work that is truly exploratory; just imagine if Galileo had been required to know in advance that he would see the moons of Jupiter, in order to get funding for his telescope! Many success stories in hydrologic research began, not as a priori hypothesis tests, but as exploratory research driven by innovations in measurement and observation. Further advances in measurement and observation will be vital in framing new hypotheses and testing them rigorously, and thus in advancing the science of hydrology.

6. Postscript

Although hypothesis testing is widely recognized as a cornerstone of good scientific practice, the contributions to this debate illustrate diverse views of whether, how, and why hypotheses should be tested in hydrology. D. M. McKnight suggests that one reason to frame hypotheses is that they can help to attract funding [*McKnight*, 2017]. Framing hypotheses, and figuring out how to test them, can certainly help in clarifying what we are trying to investigate, and why. However, in our view, funding agencies should also be willing to support exploratory research in cases where productive hypotheses are likely to emerge, but not enough is known to frame rigorous hypotheses yet.

V. R. Baker points out that computer modeling "has become the principal operating paradigm for modern hydrology," and I. Neuweiler and R. Helmig give examples of how model-based hypothesis tests have been important in moving that paradigm forward [*Baker*, 2017; *Neuweiler and Helmig*, 2017]. However, we share Neuweiler and Helmig's concern about the challenge of testing multiple overlapping hypotheses that are encoded into complex models with multiple interdependent assumptions.

Neuweiler and Helmig point out a number of cases where observations are unable to distinguish between different competing hypotheses. That naturally leads to the question of what observations would be needed to accomplish this goal. If these observations are impractical, then the hypotheses will remain unresolved. But in some cases the key observations can be made, and can be decisive, like McKnight's speciation of ferrous versus ferric iron in her mine drainage study.

Currently in hydrology, theorizing and modeling often proceed with little recognition of what data are actually available (or reasonably obtainable). Conversely, data collection efforts are often designed without explicit recognition of alternative hypotheses, and what data would be particularly powerful in testing them. The science of hydrology will be advanced by designing instruments and observations specifically to test theories and models, and conversely, by generating theories and models with an eye to the data that will actually be available to test them.

Acknowledgment
We thank the anonymous reviewer for the very helpful comments on an earlier version of this manuscript.

References

Baker, V. R. (2017), Debates—Hypothesis testing in hydrology: Pursuing certainty versus pursuing uberty, *Water Resour. Res.*, *53*, doi: 10.1002/2016WR020078.
Betson, R. P. (1964), What is watershed runoff?, *J. Geophys. Res.*, *69*, 1541–1552, doi:10.1029/JZ069i008p01541.
Beven, K., and P. Germann (2013), Macropores and water flow in soils revisited, *Water Resour. Res.*, *49*, 3071–3092, doi:10.1002/wrcr.20156.
Beven, K. J. (2001), On hypothesis testing in hydrology, *Hydrol. Processes*, *15*, 1655–1657, doi:10.1002/hyp.436.
Beven, K. J. (2010), Preferential flows and travel time distributions: Defining adequate hypothesis tests for hydrological process models, *Hydrol. Processes*, *24*(12), 1537–1547.
Beven, K. J. (2012), Causal models as multiple working hypotheses about environmental processes, *C. R. Geosci.*, *344*(2), 77–88.
Beven, K. J., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, *249*, 11–29, doi:10.1016/S0022-1694(01)00421-8.
Beven, K. J., and P. F. Germann (1982), Macropores and water flow in soils, *Water Resour. Res.*, *18*, 1311–1325.
Beven, K. J., and I. Westerberg (2011), On red herrings and real herrings: Disinformation and information in hydrological inference, *Hydrol. Processes*, *25*, 1676–1680, doi:10.1002/hyp.7963.
Blöschl, G., et al. (2016), The Hydrological Open Air Laboratory (HOAL) in Petzenkirchen: A hypothesis-driven observatory, *Hydrol. Earth Syst. Sci.*, *20*, 227–255.
Bogena, H. R. (2016), TERENO: German network of terrestrial environmental observatories, *JLSRF*, *2*, A52, 1–8, doi:10.17815/jlsrf2-98.
Bras, R., and P. S. Eagleson (1987), Hydrology, the forgotten Earth science, *Eos Trans. AGU*, *68*, 227.
Brooks, R., R. Barnard, R. Coulombe, and J. J. McDonnell (2010), Ecohydrologic separation of water between trees and streams in a Mediterranean climate, *Nat. Geosci.*, *3*, 100–104, doi:10.1038/NGEO722.
Burt, T. P., and J. J. McDonnell (2015), Whither field hydrology?, The need for discovery science and outrageous hydrological hypotheses, *Water Resour. Res.*, *51*, 5919–5928, doi:10.1002/2014WR016839.
Chamberlin, T. C. (1890), The method of multiple working hypotheses, *Science*, *15*, 92–96 (reprinted 1965, *Science*, 148, 754–759).
Clark, M., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modelling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.
Dansgaard, W. (1964), Stable isotopes in precipitation, *Tellus XVI*(4), 436–468.
Dawson, T. E., and J. R. Ehleringer (1991), Streamside trees that do not use stream water, *Nature*, *350*, 335–337.
Dooge, J. C. I. (1986), Looking for hydrologic laws, *Water Resour. Res.*, *22*, 46S–58S, doi:10.1029/WR022i09Sp0046S.
Dunne, T. (1978), Field studies of hillslope flow processes, in *Hillslope Hydrology*, edited by M. Kirkby, John Wiley, Chichester, U. K.
Emerson, G. B., W. J. Warme, F. M. Wolf, J. D. Heckman, R. A. Brand, and S. S. Leopold (2010), Testing for the presence of positive-outcome bias in peer review, *Arch. Int. Med.*, *170*, 1934–1939.
Godsey, S. E., et al. (2010), Generality of fractal 1/f scaling in catchment tracer time series, and its implications for catchment travel time distributions, *Hydrol. Processes*, *24*, 1660–1671, doi:10.1002/hyp.7677.
Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions (2015), The extent and consequences of P-hacking in science, *PLoS Biol.*, *13*(3), e1002106, doi:10.1371/journal.pbio.1002106.
Hewlett, J. D., and A. R. Hibbert (1963), Moisture and energy conditions within a sloping soil mass during drainage, *J. Geophys. Res.*, *68*, 1081–1087, doi:10.1029/JZ068i004p01081.
Hornberger, G. M., and R. C. Spear (1981), An approach to the preliminary analysis of environmental systems, *J. Environ. Manage*, *12*, 7–18.
Horton, R. E. (1933), The role of infiltration in the hydrologic cycle, *Eos Trans. AGU*, *14*(1), 446–460, doi:10.1029/TR014i001p00446.
Kavetski, D., and M. Clark (2011), Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing, *Hydrol. Processes*, *25*, 661–670, doi:10.1002/hyp.7899.
Kavetski, D., G. Kuczera, and S. W. Franks (2006), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.
Kerr, N. L. (1998), HARKing: Hypothesizing after the results are known, *Pers. Soc. Psychol. Rev.*, *2*, 196–217.
Kirchner, J. W. (1990), Gaia metaphor unfalsifiable, *Nature*, *345*, 470.
Kirchner, J. W., X. Feng, and C. Neal (2000), Fractal stream chemistry and its implications for contaminant transport in catchments, *Nature*, *403*, 524–527, doi:10.1038/35000537.
Kirchner, J. W. (2003a), The Gaia hypothesis: Conjectures and refutations, *Clim. Change*, *58*, 21–45.
Kirchner, J. W. (2003b), A double paradox in catchment hydrology and geochemistry, *Hydrol. Processes*, *17*, 871–874, doi:10.1002/hyp.5108.
Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, *42*, W03S04, doi:10.1029/2005WR004362.
Kirchner, J. W., and C. Neal (2013), Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection, *Proc. Natl. Acad. Sci. U. S. A.*, *110*(30), 12,213–12,218, doi:10.1073/pnas.1304328110.
Klemes, V. (1986), Delettantism in hydrology: Transition or destiny?, *Water Resour. Res.*, *22*, S177–S188.

Klemes, V. (1988), A hydrological perspective, *J. Hydrol.*, *100*, 3–28.
Koehler, J. J. (1993), The influence of prior beliefs on scientific judgments of evidence quality, *Organ. Beh. Hum. Decis. Proces.*, *56*, 28–55.
Lee, J., (1992), The education of hydrologists, *Hydrol. Sci. J.*, *37*, 285–289, doi:10.1080/02626669209492588.
Mahoney, M. J. (1977), Publication prejudices: An experimental study of confirmatory bias in the peer review system, *Cognit. Ther. Res.*, *1*, 161–175.
McDonnell, J. J. (2014), The two water worlds hypothesis: Ecohydrological separation of water between streams and trees?, *WIREs Water*, *1*, 323–329, doi:10.1002/wat2.1027.
McKnight, D. M. (2017), Debates—Hypothesis testing in hydrology: A view from the field: The value of hydrologic hypotheses in designing field studies and interpreting the results to advance hydrology, *Water Resour. Res.*, *53*, doi:10.1002/2016WR020050.
Neuweiler, I. and R. Helmig (2017), Debates—Hypothesis testing in hydrology: A subsurface perspective, *Water Resour. Res.*, *53*, doi: 10.1002/2016WR020047.
Nickerson, R. S. (1998), Confirmation bias: A ubiquitous phenomenon in many guises, *Rev. Gen. Psychol.*, *2*, 175–220.
Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011), False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant, *Psychol. Sci.*, *22*, 1359–1366, doi:10.1177/0956797611417632.
Sklash, M. G., R. N. Farvolden, and P. Fritz (1976), A conceptual model of watershed response to rainfall, developed through the use of oxygen-18 as a natural tracer, *Can. J. Earth Sci.*, *13*(2), 271–283, doi:10.1139/e76-029.
Uhlenbrook, S. (2006), Catchment hydrology: A science in which all processes are preferential, *Hydrol. Processes*, *20*, 3581–3585, doi:10.1002/hyp.6564.
Weiler, M., and K. Beven (2015), Do we need a community hydrological model?, *Water Resour. Res.*, *51*, 7777–7784, doi:10.1002/2014WR016731.
Westerberg, I. K., and C. Birkel (2015), Observational uncertainties in hypothesis testing: Investigating the hydrological functioning of a tropical catchment, *Hydrol. Processes*, *29*, 4863–4879, doi:10.1002/hyp.10533.
Zehe, E., et al. (2014), HESS Opinions: From response units to functional units: A thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments, *Hydrol. Earth Syst. Sci.*, *18*, 4635–4655, doi:10.5194/hess-18-4635-2014.