Behind every robust result is a robust method: Perspectives from a case study and publication process in hydrological modelling

Fabrizio Fenicia1 | Dmitri Kavetski2

1Eawag, Dübendorf, Switzerland
2University of Adelaide, Adelaide, Australia

Correspondence
Dmitri Kavetski, University of Adelaide, Adelaide, Australia.
Email: dmitri.kavetski@adelaide.edu.au

1 MODELS AS EXPLORATORY TOOLS

Models are commonplace in hydrological and environmental sciences. In scientific contexts, models are used to encapsulate our current understanding of environmental processes and to test new scientific propositions. In applied contexts, models are used to predict future environmental conditions and hence are a cornerstone of the decision-making chain in civil protection (e.g., real-time flood forecasting) and environmental management and planning (e.g., seasonal streamflow forecasts to inform water allocation, ecological health assessment, and so forth). The importance of models—and public expectations of them—are only likely to increase, as science's attention turns to the formidable environmental challenges of the XXI century.

The increased reliance on models places a corresponding expectation on model reliability and robustness. Environmental model failures can be costly—from loss of time of a PhD researcher who adopts a published model that turns out to be non-robust, to potentially catastrophic consequences during natural disasters such as floods and droughts if models incorrectly anticipate such events.

These considerations might lead an impartial observer to expect to see an increasing scientific emphasis on model design and verification. Paradoxically, there is evidence of an opposite trend taking place. For example, modern publication practices in natural sciences seem to place an ever-increasing emphasis on the presentation of “results,” rather than on the “methods” used to achieve them. Some journals minimise their methods sections by presenting them in reduced font size, or even relegate them to supplementary materials that do not form part of their formal printed versions. The motivation for these changes is perhaps understandable: results and their interpretation are usually seen as the more exciting aspects of scientific endeavours.

But such degree of prioritisation of results and interpretations over their methodological foundation brings real risks to longer term scientific rigour and credibility. The risks are particularly high as environmental sciences and applications are increasingly dominated by heavily abstracted mathematical models, which are used to simulate variables far beyond those we can measure (Cunge, 2003). Seemingly routine methodological choices can significantly affect the conclusions of case studies and investigations—making it a problematic practice to skimp over methodological details in an enthusiastic rush towards the next great discovery that models are able to “demonstrate.”

Our commentary emphasises the value of three methodological instruments that arguably remain under-utilised in hydrological modelling practice and broader environmental applications: (1) benchmarking model results with a null-hypothesis model, (2) testing model predictions in space–time validation, and (3) carrying out controlled model comparisons. A schematic representation of these instruments is shown in Figure 1, in the representative context of developing a catchment scale rainfall-runoff model. These instruments play a key role in the classic scientific method, and are certainly not new in hydrological modelling. For example, the importance of establishing meaningful benchmark models has been emphasised by many hydrologists (Pappenberger et al., 2015; Seibert et al., 2018). The need of model validation tests was notably advocated by Klemeš (1986) followed by numerous discussions on the topic (e.g., Andréassian et al., 2009; Refsgaard, 1997); see also earlier
discussion by Burges (1984). The value of systematic model comparisons and hypothesis testing was reviewed and elaborated by Clark et al. (2011) (see also Baker, 2017; Pfister & Kirchner, 2017). Yet, these instruments are not necessarily included in what is generally perceived as “good modelling practice” (Cunge, 2003). For example, the common modelling protocol, adopted in many/most hydrological modelling studies, relies on calibration-validation (Blöschl & Sivapalan, 1995; Cunge, 2003; Refsgaard, 1997), but does not explicitly consider comparisons with alternative models. Validation in space, although considered important for spatially distributed models, is not implemented routinely (Refsgaard, 1997). Research in broader environmental sciences, including climate modelling, mirrors these experiences, with the need for carefully designed model comparisons being stressed in several recent publications (e.g., Eyring et al., 2019; Rood, 2019).

So, multiple opinion papers and discussions notwithstanding, why are these instruments not routinely adopted in environmental

**FIGURE 1** Key methodological instruments discussed in this commentary: (a) comparison with a null hypothesis to “weigh” the merits of the proposed model; (b) space–time validation of model simulations, and (c) controlled stepwise comparisons to appraise the effect of individual model choices. The schematic illustrates these instruments in the context of a catchment-scale rainfall-runoff model. The specific implementation of the instruments will depend on the application.
modelling? The reason might be twofold. First there is still insufficient appreciation and guidance on what should constitute “good modelling practice.” An appropriate modelling protocol is not necessarily unique, but may depend on the type of model and its application. For example, in some cases model calibration may not be advisable (Cunge, 2003). Second, there may be an insufficient appreciation of how much these instruments matter in practice. With some exceptions (e.g., Refsgaard & Hansen, 2010), previous discussions have revolved mainly around theoretical aspects. We therefore refrain from restating the “theoretical” importance of these instruments and instead offer a practical example.

Our story is inspired by a modelling study in a Luxembourgish catchment with a surprisingly wide range of streamflow dynamics (Fenicia et al., 2016). This commentary revisits that earlier study and presents it as a Hypothetical Paper Submission and Review Process. We start from a common choice of modelling methods and illustrate how progressively more stringent techniques change the modelling conclusions. Our characters are the humble Authors, Reviewers 1–3, and the Editor. Our disclaimer is that, in this story, the model results are real, but any semblance to actual persons or actual events is purely coincidental.

2 | A HYDROLOGICAL MODELLING STUDY AND ITS REVIEW PROCESS

2.1 | The story begins ... the original submission

The story begins with the Authors submitting a manuscript on the hydrology of the 300 km\(^2\) Attert catchment in Luxembourg. In that catchment, the diversity of hydrograph dynamics observed across multiple locations beseeched the question of which climatic or landscape properties acted as dominant controls on streamflow generation. This research question was approached from a modelling perspective, seeking to find a distributed model structure and parameter values that reproduced—or as commonly put, “explained”—the observed streamflow variability.

The reader will appreciate that developing and setting up a distributed hydrological model can be a challenging endeavour. At the very least, one has to develop a conceptual model, translate it into a computer code, gather and format the data, interface the model with a calibration algorithm, and spend a non-negligible amount of time to debug and run the entire machinery. The Authors did all this, and when methodological choices were necessary, they did their best to “do things the right way,” following the recommendations and techniques from the hydrological modelling literature. After a quick description of the methods, the paper focused its attention on the results.

As examples of methodological choices, the Authors implemented the model using a robust time stepping scheme to avoid numerical instabilities (Kavetski & Clark, 2011), applied the square-root streamflow transformation to account for error heteroscedasticity/skew (Mclnerney et al., 2017; Sorooshian & Dracup, 1980), and employed global optimisation strategies (Duan et al., 1992; Qin et al., 2018). The model was validated on a separate time period not used in calibration (Blöschl & Sivapalan, 1995; Klemes, 1986; Refsgaard, 1997).

To ascertain the model's ability to simulate streamflow, the Authors had followed common “best practices” and evaluated the model not only in the time domain, but also in the signature domain (Hrachowitz et al., 2014; Kavetski et al., 2018). In particular, in addition to the Nash-Sutcliffe (NS) efficiency of streamflow time series simulations at all subcatchments—an imperfect choice for a model evaluation metric (Schaefli & Gupta, 2007), but also a familiar and still widely used one—the Authors considered three streamflow signatures: the runoff coefficient, the baseflow index, and the flashiness index (e.g., Addor et al., 2018; Baker et al., 2004). This selection of data and performance metrics was argued to provide a suitably comprehensive model evaluation.

The proposed model will be referred to as M1 (and the numbering already foreshadows it will not be the final one...). Like many distributed models, M1 was based on the concept of Hydrological Response Units (HRUs). In this case, 3 HRUs were defined, based on catchment topography. Following standard practice, the same “generic” model structure was assigned to each HRU. The main selling point of this model was its ability to capture the spatial variability of streamflow in the study area.

Figure 2a presents the performance of model M1 in terms of streamflow time series predictions and their NS efficiency in each of the subcatchments. In discussing this figure, the Authors noted that model performance in calibration was “satisfactory” or “acceptable” in all the subcatchments but one, with NS efficiencies between 0.60 and 0.93. The only rogue subcatchment was the Huewelerbach headwater catchment, where the model had a NS efficiency of 0.34. The Authors contended this was a “minor” problem representative of “idiosyncrasies” of headwater catchments (McDonnell, 2003). In terms of streamflow signatures, the Authors noted that their absolute values were difficult to match, but the dots generally followed the same pattern, with linear correlation values (Pearson correlation, R) in the range of 0.77–0.94 (Figure 2b–d). Moreover, time validation resulted in only a moderate decrease in model performance.

Overall, the model was seen to provide a “credible” indication that streamflow diversity in the Attert catchment was controlled by topography. Satisfied with these findings, the Authors submitted their work for publication. They were certainly not anticipating the odyssey ahead.

2.2 | Review round 1: A benchmark “null-hypothesis” model

Then came the reviews. Reviewers 1 and 2 were generally positive. They only had some general remarks on improving referencing and presentation. Reviewer 3, instead, was surprisingly critical. In particular, Reviewer 3 did not appreciate statements such as “satisfactory” and “acceptable”—these were not only subjective, but lacked a frame of reference—“with respect to what is the model satisfactory?” The reviewer contended that a meaningful evaluation of model performance requires a credible benchmark, that is, a “null-hypothesis”
As Reviewer 3 argued, “Getting an NS efficiency above zero ain’t exactly newsworthy.” Setting up a meaningful benchmark model is not trivial, and the literature does not offer many examples. After some reflection, the Authors came up with a solution. The most relevant hypothesis in their model setup was the selection of HRUs, as this model decision controls how the model associates distinct dominant processes with distinct landscape elements. The null-hypothesis model requested by Reviewer 3 should serve to test this hypothesis, and therefore could be represented by a model with distributed rainfall inputs but uniformly distributed parameters (Figure 1a). In this setup, the only difference between the null-hypothesis model, here referred to as M0, and the proposed model M1 was the selection of HRUs: a single HRU for M0 versus 3 HRUs for M1. Models M0 and M1 made the same choices in all other modelling decisions.

The key results of the revised manuscript are shown in Figure 3. In the time domain, model M1 outperformed model M0 in 9 of 10 sub-catchments, which for the Authors was a reassuring result. Still, M0 failed similarly to M1 in the Huwelerbach, which was once again attributed to headwater catchment “idiosyncrasies.” In the signature domain, M1 clearly outperformed M0 for the baseflow and flashiness indices, for example, with R values of 0.94 versus 0.09 respectively for the flashiness index. Notably, the inclusion of the benchmark model also allowed to replace potentially subjective absolute statements (“good” or “bad”) with relative ones (“better than” or “worse than”).

That said, model M0 diminished some of the perceived merits of model M1. For example, M0 could match the runoff coefficient signature even better than M1 just by virtue of distributing the forcings, without also distributing the model parameters. Still, the Authors argued, this exception did not overturn the conclusion that the topography based model M1 offered a “realistic” representation of hydrological processes in the study catchment. The Authors therefore resubmitted their paper, convinced that their thorough revision would satisfy the most critical reviewer...

2.3 | Review round 2: Space–time validation

The Authors’ optimism lasted as long as the re-review process. Actually, the first two reviewers were now rather enthusiastic about the paper, in which they recognised many more improvements than they had themselves asked for. Reviewer 3, instead, came up with an additional set of criticisms. A particular concern was that “temporal split sample” validation was not sufficient for a spatially distributed model. Reviewer 3 asserted that “a distributed model should be able to make predictions not only in time, but also in space.”
The paper was now held back by a single tenacious reviewer. However the request for more stringent validation was eminently sensible and hard to dismiss. The Editor therefore asked the Authors to perform this last check, which "should not take much additional work."

The Authors, confident in their model, decided to meet rather than evade the challenge, and implemented the more stringent so-called "proxy-basins, split-sample" space–time validation test (Klemes, 1986). The model was calibrated on one group of subcatchments over a given time period, and validated on another group of subcatchments over another time period (Figure 1b).

Horror and damnation! Model M1 performed considerably worse in space–time validation, as seen in Figure 4. In the time domain (panel a), the model not only continued to fail in the now infamous Huewelerbach, but also had worse NS efficiency than the benchmark model M0 in another three subcatchments. More regrettable, the pride and joy of model M1—its high correlation between simulated and observed signatures—was also lost, with R values dropping from 0.77-0.94 in time validation (Figure 3), to 0.46-0.53 in space–time validation.

The Authors contemplated defending their model, but after this experiment themselves began to question it. Instead of arguing that their hitherto favoured model M1 still provided an improvement over the benchmark (e.g., at simulating streamflow in 6 out of 10 catchments), the Authors embarked on the arduous journey of revisiting the model assumptions. After several failed attempts, and as many requests to extend the re-submission deadline ... Eureka! A new model, M2, was found that reproduced the observed data much better than models M0 and M1!

The performance of model M2 is shown in Figure 4. In terms of NS efficiency of streamflow time series (panel a), model M2 outperformed the benchmark model M0 in all subcatchments. Notably, M2 did well in the Huewelerbach—no need to blame headwater catchment idiosyncrasies no more! In terms of signatures (panels b–d), model M2 matched exceptionally well the baseflow and flashiness indexes (R of 0.89 and 0.97 respectively, whereas model M0 had R values close to zero).

So, what was the difference between models M2 and M1? Whereas model M1 assumed that streamflow variability was driven by topography and used 3 topography related HRUs, model M2 considered geology as the dominant control and used 4 geology related HRUs. These contrasting assumptions are schematised in Figure 5a,b respectively. Despite the larger number of HRUs, model M2 had fewer parameters than model M1 (11 versus 21, respectively). This reduction was obtained by tailoring the model structures to the specific HRUs by focusing on their perceived dominant processes, and also by linking some parameter values across all HRUs.
As the new model M2 was very different from model M1 advocated in the original submission, the entire set of conclusions about streamflow diversity in the Attert catchment had to change. This outcome was somewhat embarrassing for the Authors, who nonetheless rewrote a big chunk of their story and re-submitted their work, hoping that Reviewer 3 would now be satisfied.

2.4 | Review round 3: Controlled model comparisons

When the Authors received the reviews, they understood it was not a good sign, as they had been expecting a short acceptance email. The Editor did not hide some discomfort in sending the paper back to the Authors once again. This time, understandably, the Editor did not consult the first two reviewers. So there was no need to scroll down to see the comments of Reviewer 3. After a moralising preamble on how the message of the paper had progressively changed through the review process (as a result of adopting more stringent model evaluation approaches), Reviewer 3 launched their final attack.

It read: “Results show that model M2 is better than model M1. But why is that? M1 and M2 differ in multiple respects—not only in the spatial discretisation approach, but also in the model structures used to represent each HRU, and in the presence of parameter links across HRUs. Are both differences pertinent? And what is their relative impact? In other words, what is the raison d'être for improved model performance?”

The Authors, stoic in adversity, faced this new concern. Two additional models, with structures “intermediate” between models M1 and M2, were designed. The first new model, M1A, would differ from model M1 solely in the spatial discretisation approach, namely using geology rather than topography to define the HRUs. The second new model, M1B, would differ from model M2 solely in the parameter regularisation, namely by not linking parameter values across the HRUs. The resulting model comparison sequence is illustrated in Figure 5c.

Ultimately, controlled model comparisons enabled the study to attribute changes in predictive performance to specific model decisions (Figure 1c). In summary, model M1A showed the largest improvement of all model variations, indicating that in this study the geology based landscape discretisation was the key modelling decision. Model M1B also showed some improvement over M1A, lending...
credence to the idea of tailoring the model structure to each HRU. M2 achieved a very similar performance to M1B, indicating that parameter regularisation shaved off unnecessary model complexity. The overall study conclusion was that streamflow diversity in the Attert catchment was controlled by spatial variability in rainfall (runoff coefficient) and geology (baseflow and flashiness indices).

With their last strengths, the Authors re-submitted Revision 3 of their paper. Their perseverance was finally rewarded with the manuscript being accepted. The interested reader can find the complete study in Fenicia et al. (2016), where models M0, M1 and M2 in this paper correspond to models M-Uni, M-Top and M-Geo-3, and the intermediate models M1A and M1B correspond to models M-Geo-1 and M-Geo-2.

Note some minor changes to the analysis methods used in the original study by Fenicia et al. (2016). For example, here the model calibration employed a square-root streamflow transformation.
(following a recent recommendation of McInerney et al. (2017)). Luckily for the Authors this particular change in analysis methods did not affect the study conclusions!

3 | WHAT HAPPENED?

The Authors rejoiced that the paper was accepted, and also learned some important lessons. Three methodological instruments helped diagnose model deficiencies and provided guidance to overcome them (Figure 1):

1. A well-chosen null-hypothesis model provides a more meaningful test of the proposed model, and provides context when reporting model results. In our example, the inclusion of the benchmark model M0 eroded some of the perceived merits of model M1, and provided a reference for reporting performance measures such as the NS efficiency in relative rather than absolute terms (thus reducing subjective judgement).

2. Multi-site validation is critical for a model intended to produce spatially distributed predictions. More generally, validation tests should be constructed to challenge the model directly in its intended operational setup. In our example, model M1 performed poorly in space–time validation, demanding a re-examination of the model assumptions, and in turn leading to a substantially better new model M2.

3. Controlled model comparisons yield deeper insights into individual model hypotheses. Models M1 and M2 differed in too many respects to be able to identify the cause—or causes!—of their different behaviour. Comparisons with two “intermediate” models made it possible to pinpoint the impact of each model enhancement. Notably, such analysis promoted the important scientific goal of model parsimony, that is, only adopting model changes that have been demonstrated to materially improve model performance.

These instruments were not straightforward to apply, yet had a sobering impact on the modelling conclusions. While the Authors were, no doubt, at times frustrated by the exigence of Reviewer 3, ultimately they appreciated the demonstrably higher quality and robustness of the resulting analyses.

4 | SUMMARY AND CONCLUSIONS

As environmental research and operations become dominated by model based approaches, and as these models become more complex and abstract, the methods used for model design and evaluation need to become correspondingly more comprehensive and stringent. The ongoing trend of prioritising results and interpretation over methods goes against this need, and can foster a laissez faire scientific attitude, where models are misused as inexhaustible resources of (non-robust and potentially misleading) findings and recommendations. Detailed and careful reviews offer some protection, much as accomplished by Reviewer 3 in our example. However, the review process on its own, as harsh as it often seems, cannot be the sole guardian against methodological weaknesses. Indeed this protection is increasingly weak, as the growing publication rate is (perhaps inevitably) mirrored by a decline in review quality, which is increasingly limited to “generic” assessment rather than detailed scrutiny (Politi et al., 2021). Relegating methods to smaller font size, and giving them less prominence in scientific papers and presentations, can only exacerbate these concerns. A stronger awareness of good modelling practices and appropriate modelling protocols—and keeping methods forefront in scientific reporting—is needed to foster a healthier scientific attitude and reduce the risk of non-robust findings.

This commentary illustrates the impact of modelling methods on study findings, and how three methodological instruments that are known yet under-utilised in hydrological and environmental modelling protocols can be used to achieve more robust and insightful conclusions. We argue that studies where models are used as explanatory and/or predictive tools, as well as studies where “better” models are proposed, should make use of the three instruments here investigated, adapted to the specific applications. Studies should articulate a meaningful benchmark, set up a stringent validation targeted to the research question, and construct informative model comparisons to elucidate the reasons for the results. These choices are not straightforward, and therein ultimately lies the science and art of hydrological modelling. We hope that our commentary will stimulate and help hydrologists to implement these methodological instruments in their ongoing and future work.

ACKNOWLEDGEMENT

We thank Editor Jim Buttle and Michael Leonard for constructive comments that helped us improve the manuscript.

ORCID

Fabrizio Fenicia https://orcid.org/0000-0002-8065-6004
Dmitri Kavetski https://orcid.org/0000-0003-4966-9234

REFERENCES

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. Water Resources Research, 54(11), 8792–8812. https://doi.org/10.1029/2018WR022606

Andréassian, V., Perrin, C., Berthet, L., le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M. H., & Valéry, A. (2009). HESS Opinions “Crash tests for a standardized evaluation of hydrological models”. Hydrology and Earth System Sciences, 13(10), 1757–1764. https://doi.org/10.5194/hess-13-1757-2009

Baker, D. B., Richards, R. P., Loftis, T. T., & Kramer, J. W. (2004). A new flashiness index: Characteristics and applications to midwestern rivers and streams. Journal of the American Water Resources Association, 40 (2), 503–522. https://doi.org/10.1111/j.1752-1688.2004.tb01046.x

Baker, V. R. (2017). Debates—Hypothesis testing in hydrology: Pursuing certainty versus pursuing utility. Water Resources Research, 53(3), 1770–1778. https://doi.org/10.1002/2016WR020078

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. Hydrological Processes, 9(3-4), 251–290. https://doi.org/10.1002/hyp.3360090305
Burgess, S. J. (1984). Rainfall-runoff model validation: The need for unambiguous tests. IFAC Proceedings Volumes, 17(2), 3163–3168. https://doi.org/10.1016/S1474-6670(17)61464-2

Clark, M. P., Kaetvs, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resources Research, 47(9), W09301. https://doi.org/10.1029/2010wr009827

Cunge, J. A. (2003). Of data and models. Journal of Hydroinformatics, 5(2), 75–98. https://doi.org/10.2166/hydro.2003.0007

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research, 28(4), 1015–1031. https://doi.org/10.1029/91WR02985

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwatkiowskii, L., Lorenz, R., Maloney, E., Meehl, G. A., Williamson, M. S. (2019). Taking climate model evaluation to the next level. Nature Climate Change, 9(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

Fenicia, F., Kavetski, D., Savenije, H. H. G., & Pfister, L. (2016). From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. Water Resources Research, 52(2), 954–989. https://doi.org/10.1002/2015WR017398

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., & Gasco, E. D. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. Water Resources Research, 50(9), 7443–7469. https://doi.org/10.1002/2014wr115484

Kavetski, D., & Clark, M. P. (2011). Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing. Hydrological Processes, 25(4), 661–670. https://doi.org/10.1002/hyp.7899

Kavetski, D., Fenicia, F., Reichert, P., & Albert, C. (2018). Signature-domain calibration of hydrological models using approximate Bayesian computation: Theory and comparison to existing applications. Water Resources Research, 54(6), 4059–4083. https://doi.org/10.1002/2017WR020528

Klemes, V. (1986). Operational testing of hydrological simulation models. Hydrological Sciences Journal, 31(1), 13–24. https://doi.org/10.1080/02626668609491024

McDonnell, J. J. (2003). Where does water go when it rains? Moving beyond the variable source area concept of rainfall-runoff response. Hydrological Processes, 17(9), 1869–1875. https://doi.org/10.1002/hyp.5132

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resources Research, 53(3), 2199–2239. https://doi.org/10.1002/2016WR019168

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., & Saladon, P. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. Journal of Hydrology, 522, 697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024

Pfister, L., & Kirchner, J. W. (2017). Debates—hypothesis testing in hydrology: Theory and practice. Water Resources Research, 53(3), 1792–1798. https://doi.org/10.1002/2016WR020116

Politi, P., Majumdar, S. N., Politi, A., & Ruffo, S. (2021). Some considerations about reviewing and open-access in scientific publishing. arXiv: 2104.01794. Placeholer Text https://ui.adsabs.harvard.edu/abs/2021arXiv210401794P

Qin, Y., Kavetski, D., & Kuczera, G. (2018). A robust Gaussian–Newton algorithm for the optimization of hydrological models: Benchmarking against industry-standard algorithms. Water Resources Research, 54(11), 9637–9654. https://doi.org/10.1029/2017WR022489

Refsgaard, J. C. (1997). Parameterisation, calibration and validation of distributed hydrological models. Journal of Hydrology, 198(1), 69–97. https://doi.org/10.1016/S0022-1694(96)03329-X

Refsgaard, J. C., & Hansen, J. R. (2010). A good-looking catchment can turn into a modeller’s nightmare. Hydrological Sciences Journal, 55(6), 899–912. https://doi.org/10.1080/02626667.2010.505571

Rood, R. B. (2019). Validation of climate models: An essential practice. In Computer Simulation Validation (pp. 737–752). Springer.

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? Hydrological Processes, 21(15), 2073–2080. https://doi.org/10.1002/hyp.6825

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. Hydrological Processes, 32(8), 1120–1125. https://doi.org/10.1002/hyp.11476

Sorooshian, S., & Dracup, J. A. (1980). Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases. Water Resources Research, 16(2), 430–442. https://doi.org/10.1029/WR016i002p00430

---

**How to cite this article:** Fenicia, F., & Kavetski, D. (2021). Behind every robust result is a robust method: Perspectives from a case study and publication process in hydrological modelling. Hydrological Processes, 35(8), e14266. https://doi.org/10.1002/hyp.14266