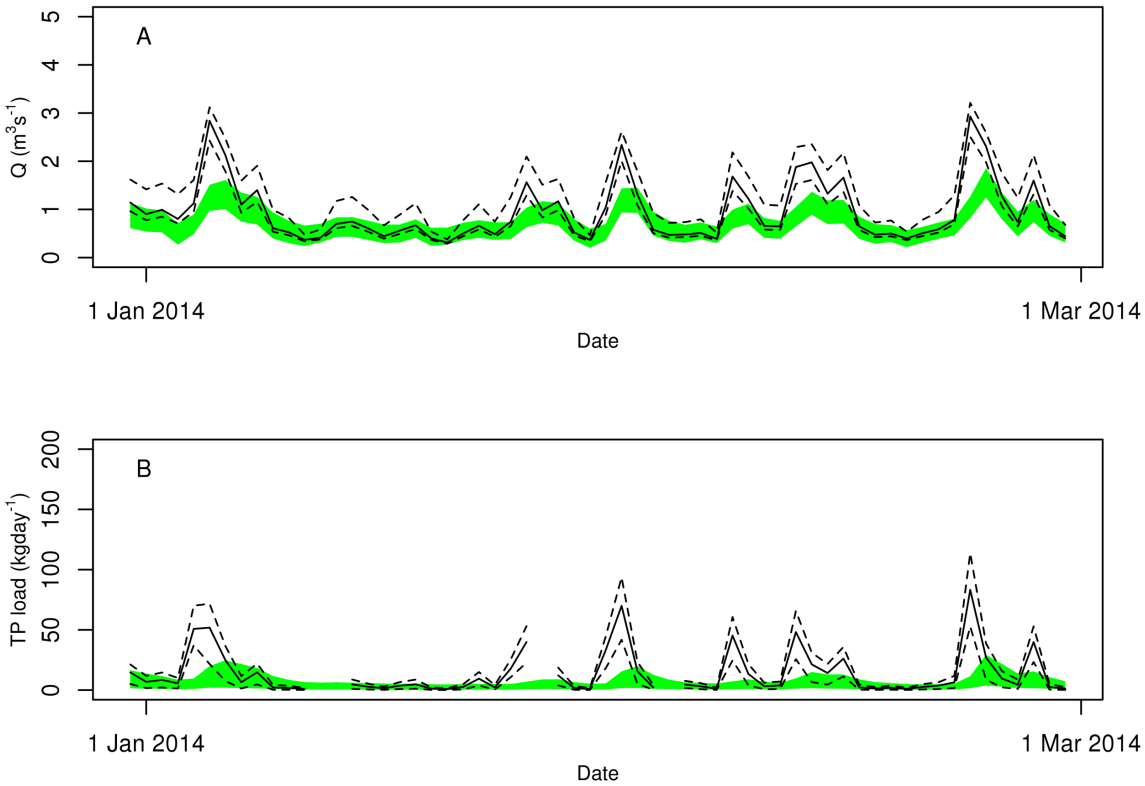
On hypothesis testing in hydrology: why falsification of models is still a really good idea.

Authors: **Keith J Beven*, Lancaster University**
k.beven@lancaster.ac.uk
Orchid: 0000-0001-7465-3934
no conflicts of interest

Abstract

This opinion piece argues that in respect of testing models as hypotheses about how catchments function, there is no existing methodology that adequately deals with the potential for epistemic uncertainties about data and hydrological processes in the modelling processes. A rejectionist framework is suggested as a way ahead, wherein assessments of uncertainties in the input and evaluation data are used to define limits of acceptability prior to any model simulations being made. The limits of acceptability might also depend on the purpose of the modelling so that we can be more rigorous about whether a model is actually fit-for-purpose. Different model structures and parameter sets can be evaluated in this framework, albeit that subjective elements necessarily remain, given the epistemic nature of the uncertainties in the modelling process. One of the most effective ways of reducing the impacts of epistemic uncertainties, and allow more rigorous hypothesis testing, would be to commission better observational methods. Model rejection is a good thing in that it requires us to be better, resulting in advancement of the science.

Graphical/Visual Abstract and Caption



An example of model invalidation in an application of the SWAT model to the 12.5km^2 Newby Beck catchment in Cumbria. Black solid and dashed lines represent the daily estimates of discharges and total phosphorus loads, calculated from high frequency observations (15min for flow, 30 min for discharge), with 95% uncertainty bound estimates. Green band represents range of the 1001 best simulations from 5 million model runs, evaluated on both discharge and total phosphorus. The model was run at a daily time step. Retention of the 1001 simulations required extension of the limits of acceptability based on the observational uncertainties by a factor of 5.3 during the evaluation period. The figure represents the simulations for a separate validation period.

Introduction

Hydrology is an inexact science, subject to both random (aleatory) and knowledge (epistemic) uncertainties. As such there are important issues about how to test models as hypotheses about system functioning in all aspects of hydrology and water science. When we can assume that models errors have a simple aleatory structure, then the full power of statistical hypothesis testing is available. But this is not normally the case. It is more usual that epistemic uncertainties dominate model performance such that how to do hypothesis testing is a more open question. A recent series of papers in Water Resources Research addressed the problem of hypothesis testing in hydrology[18]. They concluded that testing models as hypotheses would be a good way of improving hydrological science but is difficult: because both observed and predicted variables are theory-laden quantities and subject to significant uncertainties; because model structures are often complex, but incomplete; and because parameters are often difficult to estimate given the data available (parameter inference is underdetermined, leading to equifinality of conceptual models and parameter sets)[41,44].

These papers also addressed the question of epistemic uncertainties by suggesting that where we lack hydrological knowledge then there is a need for more exploratory and experimental hydrology (what Baker[1] refers to as abductive inference) and that exploratory hydrology can be fun and rewarding and should be valued more highly by research funders[1,35,44]. Most of us (certainly of my generation) cut our hydrological teeth on exploratory hydrology in trying to understand what was happening in one particular catchment area, and we were consequently influenced by its own uniqueness of hydrological and landscape characteristics (see, in my case, Beven[3,4,5]). We soon realised that exploratory hydrology is already difficult, given that the field techniques we had available were not really adequate to investigate flow pathways and fluxes in detail, particularly in the subsurface and at scales large than small cores. In fact, that is still the case - it remains difficult to get good estimates of the rainfalls over even a small catchment area; more so for actual evapotranspiration over heterogeneous land use and hillslopes; and even more so for fluxes in subsurface flow pathways[41,44]. It is clear that there is still much to learn from exploratory hydrology but there remains a lack of discussion of how we might test models as hypotheses in the face of epistemic uncertainties.

Exploratory hydrology, exploratory modelling and models as hypotheses

As experimental hydrologists, we also soon realise that as catchment scale increases it becomes much more difficult to do active studies of processes that are representative of the wider range of conditions that come into play. We then often resort to inferences from observed hydrological responses at the scale of interest, and one way of doing so is to create models that reflect our small scale understanding from observation and experimental study as far as possible, though this might not properly represent the change in dominant processes at different scales[16,28]. That has also served the purpose of allowing quantitative predictions to be provided to decision makers who manage the water system to meet the needs of society. Water Resource Management, in all its aspects, has been a major driver for model development in hydrology, in addition to that drive to demonstrate *"that we do, after all, understand our science and its complex interrelated phenomena"* (in the words of Max Kohler).

But the complexity of hydrological systems means that models that reflect that understanding will have many components and parameters, even though there are many aspects about which we have

relatively poor understanding. The combination of a model structure and a parameter set can be considered as a hypothesis about how a hydrological system functions (I include here the additional components that depend on the simulation of water fluxes, including water quality, sediment transport, drainage systems, and other features that might be required for water resources management). If we can find an acceptable model of that system, either deterministic or stochastic, it can be used to make deductive simulations of behaviour for a variety of purposes. That is why we want our models to get the right results for the right reasons, so that they are fit-for-purpose when used to simulate required variables, not all of which might be readily observable. There has been a long discussion in hydrological modelling about how best to determine appropriate effective values of parameters to allow for the uniqueness of particular catchment areas when calibration data are often limited and uncertain[6,8], but less about what qualifies as fit-for-purpose for different types of purpose. As hydrological scientists, we do not want to be using models that are not fit-for-purpose. To do so would be to draw the wrong inferences about the future behaviour of hydrological systems and lead to less than robust decisions in management. This suggests, therefore, that testing models (or components of models) as hypotheses is a valuable part of doing science in hydrology, and that falsification of models is still a really good idea.

But that is not really how hydrology, as an example of an inexact science, seems to have worked. There are not many papers in the literature that explicitly reject hydrological models as hypotheses. This is certainly partly because of the positive bias of publication. Papers are much more likely to be accepted for publication if they conclude that a model gives adequate simulations of the observations, than reporting failures (even if, in some cases, this requires data assimilation to update the model states as the simulation proceeds). There are just a few examples of papers that report rejecting all the versions of a model tried [5,19,24,36,37,43], but personal experience suggests that such papers can be rather difficult to get past referees, especially where a referee is implicated in the development of a model.

More generally, poor model results do not get reported. They are considered rather as part of the development of a modelling study. They are improved by debugging the model code, changing the model assumptions, modifying parameter sets or "correcting" model boundary conditions. By such learning processes, we aim to gradually improve models as representations of hydrological systems, even if we do so without going through a specific hypothesis testing process (it might rather be called exploratory modelling, by analogy with exploratory field hydrology). The result, however, is that we have many competing hydrological models, with different assumptions, parameterisations and numerical solution schemes, that purport to do the same thing – modelling the rainfall-runoff process, modelling water tables, modelling isotope and nutrient concentrations and other water quality variables, modelling erosion and sediment transport, and so on. We also have modelling systems that provide many options of different process components, most recently, the SUMMA system[21]. This implies a need to test different model structures as hypotheses about how a catchment works, in addition to the estimation of parameter values within those structures. It might be the case that different hydrological processes and regimes in different catchments will require different modelling assumptions, but why has there been so little real testing of those models as hypotheses or, even more importantly, as fit-for-purpose?

One reason is that they have all been considered acceptable in some sense, at least by the authors and referees on the papers in which they, and their simulation results, are described. Since hydrology

is an inexact science, we (as modellers and referees ourselves) expect that the degree to which we can simulate any available observations will be necessarily limited. There will always be some residual error, whether that be due to error and uncertainty in the observational data itself, necessary approximations in the model assumptions, or error and uncertainty in the forcing data for the model. And it is clear that the resulting errors and uncertainties cannot be treated in simple statistical terms given their epistemic nature (see, for example,[6,8,12]). This makes both establishing appropriate likelihood measures and defining appropriate methods of hypothesis testing particularly challenging[10].

Hypothesis testing and fit-for-purpose

But some form of more rigorous evaluation, that allows the possibility of model falsification as not fit-for-purpose, is surely required. That might depend, of course, on what the particular purpose might be. Purpose will govern the types of model structures chosen to be evaluated and the way in which they might be evaluated given the data available. For some purposes we will be more interested in flood peaks, in others recession behaviour, in others flow pathways, in others how water fluxes relate to residence and travel times and water quality. It might also be appropriate to use different model structures and parameter sets for different model scales. So what methodologies are available for testing models as hypotheses in this context?

For some, hypothesis testing implies a statistical analysis (for example the analysis of changes in flow duration curves of Kroll et al.[33], and the tests of flood frequency distributions using information criteria in Haddad and Rahman[26]). This generally requires making assumptions about the structure of the model residuals conditional on the model being true, and that the sources of uncertainty are fundamentally aleatory in nature. This is not really a good basis for considering whether models are fit-for-purpose, especially when we suspect that there will be important epistemic uncertainties involved. Statistical likelihood functions do not allow for giving hypotheses a likelihood of zero, only for comparisons between likelihoods. Likelihoods might be very very small (and range over tens or hundreds of orders of magnitude) but are never zero. Model rejection in that context requires some additional subjective judgements, either in assuming a prior likelihood of zero for some model configurations in a Bayesian framework, defining some tolerance level in Approximate Bayesian Computation, or in deciding on the choice of one model over another using some information criterion or Bayes ratios. There is no real mechanism for falsification in such a framework (except again by some qualitative judgement by the modeller that the results are not yet good enough to write the paper).

A more attractive approach, still based in probability theory, is the information theory approach advocated by Nearing and Gupta[38,39]. This makes a comparison between the performance of a model and the information content that can be extracted directly from the available data using purely data-based or machine learning algorithms. The basis of comparison is an entropy measure as calculated from the cumulative distribution of the variables being predicted and the equivalent model outputs. This approach has some attractive features, in particular that it does not require any assessment of sources of uncertainty in the modelling process, but works directly with the data as recorded. It also does not require any explicit assumptions about the structure of the modelling residuals. Nearing et al.[39] argue that it is far more valuable to consider the information provided by a model than the uncertainty associated with the predictions, and we should require that any process model should

provide more information than can be gleaned from the data itself. Thus any process model that results in an entropy greater than the data-based model could be rejected (though it is important to compare like with like: in an early application of this approach Gong et al.[25] tested a hydrological simulation model against a one-step ahead data-based forecasting model; unsurprisingly the simulation model did not perform as well!). Some interesting recent studies have concerned "model benchmarking" in evaluating land surface parameterisations in climate models[2,27,40] and in a multi-site application of the VIC rainfall-runoff model (Newman et al.[42]).

The advantages of this approach would, however, also seem to contain the seeds of some important limitations. Since the entropy measure is based only on the cumulative distribution of the variable of interest, any information about timing errors, either within an event or in the overprediction and underprediction of different events, is not taken into account. It might also be the case that if there are consistent epistemic uncertainties in the forcing data and evaluation observations, then not all events might be informative in evaluating model performance[15], in that the data conflict with basic concepts that underlie the model. A data-based model could compensate for consistent biases in the data, in ways that a process model constrained for example by mass and energy balance cannot.

Of course, demonstrating that a purely data-based model can extract more information from the data than a process based model is in itself a valuable learning tool. It suggests that we could do better. There are other issues with this information based approach including the possibility of different data-based models being more or less successful for different data periods (a form of equifinality of data-based models); testing for the possibility of over-fitting of data-based models when uncertainties are epistemic; and whether a difference in entropy measures should be considered significant if we accept that there are uncertainties in the data. It has also been suggested that treating data as crisp values might not result in the best data-based models[32].

The information based assessment of models is, however, one way of asking the question of just how good should we expect a model to be, given the information contained in a data set? Posing that question a little differently, we could also ask just how bad does the performance of a model have to be for it to be rejected as not fit-for-purpose given what we understand about uncertainties in hydrological data sets? This type of rejectionist framework has always been available in the Generalised Likelihood Uncertainty Estimation (GLUE) methodology, initially using a decision about a threshold for one or more performance measures (and widely criticised for the subjective nature of that choice) and later in the use of limits of acceptability based on what is known about uncertainties in the observational data[6,11,17,34]. Within this framework we can decide on when a model (structure and parameter set) as hypothesis should be considered as acceptable or rejected using what we know, or can speculate, about the nature of errors in the observational data and about what is required to make a difference to a decision in the purpose of a model application. We can also decide to make new observations or new types of observations with a view to being more rigorous in deciding whether a model is giving the right results for the right reasons (at least where this is feasible given the available observational techniques and resources). It is important, as noted by Beven[6,8], that those limits of acceptability (and any assumptions on which they are based, including the identification of disinformative periods of data) should be defined before running the models to be evaluated.

A suitable framework is available, therefore but, as already noted, the application of any form of model hypothesis testing in hydrology is relatively rare. Why is this? Is there some concern that more models might not be considered as acceptable (a recent study that showed that all SWAT models tried in an application simulating discharge and nutrient concentrations in a UK catchment could be rejected has been proving difficult to get published)? Is it simply the expectation of limited accuracy of models in the inexact sciences so that if the results look qualitatively reasonable then it is not necessary to look more closely, since all models will be expected to be wrong in some details?

But that is then saying that our standards need not be too high. Should we not have the ambition of being a little more rigorous than that? This question has, of course, been raised before, notably by Vit Klemeš in his papers on model testing. It seems that little has changed in the three decades since Klemeš[31] demanded: "*What are the grounds for credibility of a given hydrological simulation model? In current practice, it is usually the goodness of fit of the model output to the historic record in a calibration period, combined with an assumption that conditions under which the model will be used will be similar to those under calibration. This may be reasonable in the simplest cases of the "filling-in missing data" problem but certainly not if the express purpose of the model is to simulate records for conditions different from those corresponding to the calibration record. Here we have to do with the problem of general model transposability which has long been recognized as the major aim and the most difficult aspect of hydrological simulation models. Despite this fact, very little effort has been expended on the testing of this most important aspect*" (p.15).

Hypothesis testing and data uncertainties

Given that we should not expect a model to be better than the data that is used to force it or evaluate it, we need to make a careful assessment of such data uncertainties. By analogy with Type I and Type II errors in statistical decision making, there are two types of errors that we wish to avoid in evaluating a model[6,8]. We do not want to reject a model that would be useful in prediction just because of data uncertainties; and we do not want to accept a model that would be misleading in prediction just because of data uncertainties. Of these two types of error the former is more important since once a model is rejected it will (generally) not be considered further. In the latter case, we would expect that further evaluation would show that a model is not actually fit for purpose.

So any form of hypothesis testing in hydrology needs to take proper account of data uncertainties. But, as noted earlier, such uncertainties for both forcing and evaluation data will be usually dominated by epistemic rather than aleatory error. Analysis of such errors requires assumptions to be made about the characteristics of different sources of uncertainty, and clearly it is possible to be wrong in making such assumptions, for good epistemic reasons. That does not, however, imply that it is not worth the effort. The very fact of having to decide about assumptions already makes the process more rigorous, in that those assumptions then define an audit trail for the analysis, an audit trail that can then be evaluated by potential users of the model outputs for an application[14].

There remain issues to be resolved about the types of assumptions that might be made. If we consider the case of a distributed rainfall-runoff model that requires rainfall and evapotranspiration forcing data, based on local raingauge and eddy correlation latent heat observations, and that will be evaluated using soil moisture, water table and discharge data then probably the only variable that is easy to assess for error and uncertainty is the stream discharge (and even then for extreme high and low flows there will normally be significant epistemic uncertainty in the rating curve). For the input

data we will not be too sure about how accurate and representative the raingauge and eddy correlation data are for different types of events over the catchment, even if there are multiple site observations. Constructing plausible realisations for such errors (as opposed to simple stochastic models of point variables) has not been properly addressed, and would seem to be quite a difficult problem; again for good epistemic reasons (for example, are there any constraints on "outlier" errors for different event types that might result in events that are disinformative in model evaluation[10,15]). For the internal state data we will not be too sure about how the point soil moisture and water table data might relate to the equivalent variables at the discrete element scale in the distributed model (the commensurability problem). It is also unclear how measured values of catchment characteristics might relate to the effective values of model parameters (also a form of commensurability problem). In addition, any errors in the forcing data will get processed through the nonlinear dynamics of the (approximate) model structure to produce model errors of complex and non-stationary structure[10]. Note that does not mean that we need to work outside a probabilistic assessment of uncertainty, only that it is difficult to define likelihoods and probabilities that reflect the epistemic nature of the uncertainties involved. We can probably generally conclude, however, that epistemic uncertainties create problems for estimating likely occurrences in any formal framework (including, as noted earlier, for information based testing using entropy measures) and will necessarily involve some subjective choices that will affect any consequent estimates of probabilities.

So how to proceed? One way is through the limits of acceptability approach. This is equivalent to a form of fuzzy reasoning, with the limits acting as constraints in the sense of the General Theory of Uncertainty of Zadeh[48,49] and given an axiomatic basis in the General Information Theory of Klir[32]. It was one of the options suggested in the original GLUE paper of Beven and Binley[see11] and in the set theoretic approach of Keesman and van Straten[28,46], and Rose et al.[45]. In imposing limits as constraints we can try to assess the observational error in the predicted variables and use that as the basis for model evaluation. Limits can be imposed on individual observations, or on summary statistics of those observations. If the limits of acceptability are normalised to a common scale[6,15,17], different limits of acceptability evaluations against different observations can be considered in a common framework.

Whether the model predictions lie within limits determined in this way, however, will depend on the error and uncertainty associated with the forcing data. The limits will need to be extended to allow for the uncertainty in the forcing data, and as noted above, this adjustment might need to be non-stationary in nature. Because of the difficulty in constructing input error realisations when the errors are epistemic in nature, we cannot easily determine the magnitude of the adjustment (this could be the subject of research in basins where there are very good spatial observations of hydrological forcing data). We can, however, assess what critical adjustment of the limits would be required for a model run to be considered acceptable. Given some definition of the limits based on the evaluation data, this is easily calculated for every model run on a normalised scale.

So what would we then expect as hydrologists? If a model performs within limits of two times the assessed observation error would we consider that model to be acceptable. Probably yes. What about 5 times? Or 10 times? Would a model that cannot simulate within 10 times the limits of acceptability based on the evaluation observational data be considered as useful in prediction? Perhaps not, unless we had reason to suspect that the forcing data could produce errors of such a magnitude (what would cause us to suspect that degree of effect?). Any decision about an appropriate limit would necessarily depend on how good the forcing data are, of course, but if the

forcing data (combined with any model structural errors) are sufficiently in error that the simulation cannot fall within 10 times the base limits of acceptability, should that model be considered as useful in prediction or fit-for-purpose? Should the debate about hypotheses testing be about defining standards of fitness-for-purpose in different circumstances, in the same way as statisticians allow for standards in allowing for Type I and Type II errors? This might also be imposed as a further fuzzy constraint within Zadeh's Generalised Theory of Uncertainty, which allows for natural language variables. Could fitness-for-purpose be handled in such a framework?

Reducing data uncertainties

Statistical theory allows for the reduction in Type I and Type II errors as more sample become available (though the uncertainties associated with individual samples is often ignored, or assumed to be taken from a simple common statistical distribution). Perhaps the most effective way of improving hypothesis testing in the inexact sciences would be to decrease the uncertainties associated with the forcing and evaluation data. Many of the advances in hydrology in the last 50 years have been initiated by the availability of a new type of measurement. I have already suggested that the hydrological community should be much more proactive about commissioning new experimental methods[9], in a similar manner to commissioning a satellite such as SWOT[47]. This is a long, long, process, but would surely benefit our science. It would be an interesting discussion about what the community should, in fact, commission. This would be limited by the current state of technology, but would also require some hypotheses about what new variables it would be most important to observe, or what existing observables, including rainfalls and discharge, it might be most important to improve. Commensurability issues also require consideration of the scale of observations required (with due consideration to the physical and technological constraints). It would already be an advance for hypothesis testing if we could be sure that the observations used to drive and evaluate a hydrological did, themselves, satisfy the water balance and energy balance equations over the area of interest.

Advancing the science.

We currently have a situation in hydrology where a wide variety of models are used to do essentially the same types of predictions and future projections of river flows and other variables of interest. Perhaps the majority of papers published in water resources journals now involve model predictions and projections of some type. Where model intercomparisons have been done, different models give different results and it is often the case that the rankings of models in terms of performance will vary with the period of data used, site or type of application. This would seem to be a very unsatisfactory situation for the advancement of the science, especially when we expect that when true predictions are made, they will turn out to be at best highly uncertain and at worst quite wrong. It is a situation that cries out for more rigorous testing of models as hypotheses, while recognising the uncertainties associated with the data. But defining what might be considered as rigorous requires a research programme based on the best data sets available, and preferably data sets where both hydrological and tracer response information are available, so that better testing of whether a model is getting the right result for the right reasons is possible[22,23,29]. There is an implication that, given rigorous hypothesis testing we should surely be much more willing to falsify some of the models that are currently available and widely used. This is, after all, a good thing so that the science will progress in the future, by rejecting what has been inadequate in the past.

Conclusion

This paper has discussed some of the issues involved in testing models as hypotheses about catchment functioning in the face of epistemic uncertainties in both data and process representations. A framework for hypothesis testing, in terms of defining limits of acceptability before making model runs, is suggested. Past discussion[e.g.10,23,20,36] suggests that hydrologists might find it difficult to agree on such a framework, depending on how far epistemic uncertainty is seen as an issue. It is, however, a framework that might be refined as we learn more about the nature of the observational and commensurability uncertainties for both forcing and evaluation data, and about the value of different types of evidence about the system response that could be used in model evaluation. Eventually this might lead towards more rigour in testing models as hypotheses. Such a framework focuses attention on the quality of forcing and evaluation data used in model testing, resulting in a suggestion that the community should be more pro-active in commissioning better observational methods. That might be the most effective way of reducing the impacts of epistemic uncertainties. And if we cannot falsify models in this way, then what does that imply about the uncertainties in the hydrological data that we use, and the decisions that depend on both data and model outcomes?

Sidebar title: Hypothesis testing in hydrology

[Please include sidebars in the body of the text where appropriate]

Acknowledgments

This paper was stimulated by the series of WRR Debates papers on Hypothesis Testing in Hydrology and has benefited greatly from comments provided by Grey Nearing, Rich Vogel, Gordon Grant, Guillaume Thirel, Stuart Lane and three anonymous referees. This version takes some of those comments into account, albeit that there are fundamental differences in philosophy between some of us. Thanks are due to Michael Hollaway who carried out the simulations and prepared the figure for the graphical abstract. This work is a contribution to NERC Grant NE/R004722/1.

References

1. Baker, V. R., 2017, Debates—Hypothesis testing in hydrology: Pursuing certainty versus pursuing uberty, Water Resour. Res., 53, doi: 10.1002/2016WR020078.

2. Best, M.J., Abramowitz, G., Johnson, H.R., Pitman, A.J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P.A., Dong, J. and Ek, M., 2015. The plumbing of land surface models: benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), pp.1425-1442.

3. Beven, K.J., 1978, 'The hydrological response of headwater and sideslope areas'. *Hydological Sciences Bulletin*, 23(4), 419-437.

4. Beven, K J, 2000, Uniqueness of place and process representations in hydrological modelling, *Hydrology and Earth System Sciences*, 4(2), 203-213.

5. Beven, K J, 2001, Dalton Medal Lecture: How far can we go in distributed hydrological modelling?, *Hydrology and Earth System Sciences*, 5(1), 1-12.

6. Beven, K J, 2006, A manifesto for the equifinality thesis, *J. Hydrology*, 320, 18-36.

7. Beven, K J, 2008, On doing better hydrological science, *Hydrological Processes (HPToday)*, 22: 3549-3553.

8. Beven, K J, 2012, Causal models as multiple working hypotheses about environmental processes, *Comptes Rendus Geoscience, Académie des Sciences*, Paris, 344: 77–88, doi:10.1016/j.crte.2012.01.005 .

9. Beven, K J, 2016a, Advice to a young hydrologist, *Hydrological Processes,* 30, 3578–3582*;* DOI: 10.1002/hyp.10879.

10. Beven, K J., 2016b, EGU Leonardo Lecture: Facets of Hydrology - epistemic error, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrol. Sci. J.* 61(9):1652-1665, DOI: 10.1080/02626667.2015.1031761

11. Beven, K. J., and Binley, A. M., 2014, GLUE, 20 years on. *Hydrol. Process.* 28(24):5897-5918, DOI: 10.1002/hyp.10082.

12. Beven, K J, Buytaert, W and Smith, L. A., 2012a, On virtual observatories and modeled realities (or why discharge must be treated as a virtual variable), *Hydrol. Process.,* DOI: 10.1002/hyp.9261

13. Beven, K., P.J. Smith, I. Westerberg, and J. Freer, 2012b, Comment on "Pursuing the method of multiple working hypotheses for hydrological modeling" by M. P. Clark et al., Water Resour. Res., 48, W11801, doi:10.1029/2012WR012282.

14. Beven, K. J., Leedal, D. T., McCarthy, S., 2014, Framework for assessing uncertainty in fluvial flood risk mapping, CIRIA report C721, 2014, at http://www.ciria.org/Resources/Free_publications/fluvial_flood_risk_mapping.aspx

15. Beven, K. J., and Smith, P. J., 2015, Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models, *ASCE J. Hydrol. Eng*., DOI: 10.1061/(ASCE)HE.1943-5584.0000991.

16. Beven K J and Wood, E F, 1993, Flow routing and the hydrological response of channel networks, in Beven K J and Kirkby, M J, *Channel Network Hydrology,* Wiley: Chichester, 99-128.

17. Blazkova, S., and Beven, K J, 2009, A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, *Water Resour. Res*., 45, W00B16, doi:10.1029/2007WR006726.

18. Blöschl, G., 2017, Debates—Hypothesis testing in hydrology: Introduction, *Water Resour. Res*., 53, 1767–1769, doi:10.1002/2017WR020584.

19. Choi, H T and Beven, K J, 2007, Multi-period and Multi-criteria Model Conditioning to Reduce Prediction Uncertainty in Distributed Rainfall-Runoff Modelling within GLUE framework, *J. Hydrology,* 332 (3-4): 316-336

20. Clark, M. P., D.Kavetski, and F.Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, **47**, W09301, doi:10.1029/2010WR009827.

21. Clark, M.P., Nijssen, B., Lundquist, J.D., Kavetski, D., Rupp, D.E., Woods, R.A., Freer, J.E., Gutmann, E.D., Wood, A.W., Gochis, D.J. and Rasmussen, R.M., 2015. A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research*, *51*(4), pp.2515-2542.

22. Davies, J, Beven, K J, Nyberg, L and Rodhe, A, 2011, A discrete particle representation of hillslope hydrology: hypothesis testing in reproducing a tracer experiment at Gårdsjön, Sweden, *Hydrological Processes,* 25: 3602–3612. doi: 10.1002/hyp.8085.

23. Davies, J., K. J. Beven, A. Rodhe, L. Nyberg and K. Bishop, 2013, Integrated modelling of flow and residence times at the catchment scale with multiple interacting pathways, *Water Resour. Res.* 49(8): 4738-4750 DOI: 10.1002/wrcr.20377

24. Dean, S, J. E. Freer, K. J. Beven, A. J. Wade and D. Butterfield, 2009, Uncertainty Assessment of a Process-Based Integrated Catchment Model of Phosphorus (INCA-P), *Stoch. Environ. Res. Risk Assess*. 23:991–1010, DOI 10.1007/s00477-008-0273-z

25. Gong, W., Gupta, H.V., Yang, D., Sricharan, K. and Hero, A.O., 2013. Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water resources research*, *49*(4), pp.2253-2273.

26. Haddad, K. and Rahman, A., 2011. Selection of the best fit flood frequency distribution and parameter estimation procedure: a case study for Tasmania in Australia. *Stochastic Environmental Research and Risk Assessment*, *25*(3), pp.415-428.

27. Haughton, N., Abramowitz, G., Pitman, A.J., Or, D., Best, M.J., Johnson, H.R., Balsamo, G., Boone, A., Cuntz, M., Decharme, B. and Dirmeyer, P.A., 2016. The plumbing of land surface models: Is poor performance a result of methodology or data quality?. *Journal of Hydrometeorology*, *17*(6), pp.1705-1723.

28. Keesman, K. and Straten, G., 1990. Set membership approach to identification and prediction of lake eutrophication. *Water Resources Research*, *26*(11), pp.2643-2652.

29. Kirchner, J.W., 2006. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), W03S04, DOI: 10.1029/2205WR004362

30. Kirkby, M J., 1976, Tests of the random network model and its application to basin hydrology, Esrth Surface Processes, 1: 197-212

31. Klemeš, V. 1986, Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31:1, 13-24,doi: 10.1080/02626668609491024

32. Klir, G.J., 2006, Uncertainty and Information, Wiley: Hoboken, NJ

33. Kroll, C.N., Croteau, K.E. and Vogel, R.M., 2015. Hypothesis tests for hydrologic alteration. *Journal of Hydrology*, *530*, pp.117-126.

34. Liu, Y, Freer, JE, Beven, KJ and Matgen, P, 2009, Towards a limits of acceptability approach to the calibration of hydrological models: extending observation error, *J. Hydrol.*, 367:93-103, doi:10.1016/j.jhydrol.2009.01.016.

35. 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.

36. Mitchell, S, Freer, J and Beven, KJ, 2009, Multiple sources of predictive uncertainty in modeled estimates of net ecosystem CO2 exchange, *Ecol. Model.* 220: 3259–3270, doi:10.1016/j.ecolmodel.2009.08.021

37. Mitchell, S, Beven, K J, Freer, J and Law, B,, 2011, Processes influencing model-data mismatch in drought-stressed, fire-disturbed, eddy flux sites. *JGR-Biosciences,* 116: doi:10.1029/2009JG001146

38. Nearing, G.S. and Gupta, H.V., 2015. The quantity and quality of information in hydrologic models. *Water Resources Research*, *51*(1), pp.524-538.

39. Nearing, G.S., Tian, Y., Gupta, H.V., Clark, M.P., Harrison, K.W. and Weijs, S.V., 2016a. A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, *61*(9), pp.1666-1678.

40. Nearing, G.S., Mocko, D.M., Peters-Lidard, C.D., Kumar, S.V. and Xia, Y., 2016b. Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *Journal of Hydrometeorology*, *17*(3), pp.745-759.

41. Neuweiler, I. and R. Helmig, 2017, Debates—Hypothesis testing in hydrology: A subsurface perspective, *Water Resour. Res.*, 53, doi: 10.1002/2016WR020047.

42. Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B. and Nearing, G., 2017. Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, (2017).

43. Page, T., Beven, K.J. and Freer, J., 2007, Modelling the Chloride Signal at the Plynlimon Catchments, Wales Using a Modified Dynamic TOPMODEL. *Hydrological Processes,* 21*,* 292-307.

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

45. Rose, K.A., Smith, E.P., Gardner, R.H., Brenkert, A.L. and Bartell, S.M., 1991. Parameter sensitivities, Monte Carlo filtering, and model forecasting under uncertainty. *Journal of Forecasting*, *10*(1-2), pp.117-133.

46. Van Straten, G.T. and Keesman, K.J., 1991. Uncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example. *Journal of Forecasting*, *10*(1-2), pp.163-190.

47. Yoon, Y., Durand, M., Merry, C.J., Clark, E.A., Andreadis, K.M. and Alsdorf, D.E., 2012. Estimating river bathymetry from data assimilation of synthetic SWOT measurements. *J. Hydrology*, 464, pp.363-375.

48. Zadeh, L.A., 2005, Toward a generalized theory of uncertainty (GTU)—an outline, *Information Sciences*, 172: 1–40.

49. Zadeh, L. A., 2006, Generalized theory of uncertainty (GTU)—principal concepts and ideas, *Computational Statistics & Data Analysis*, 51: 15 – 46

Further Reading