HYDROLOGICAL PROCESSES
Hydrol. Process. **28**, 5897–5918 (2014)
Published online 5 November 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.10082

GLUE: 20 years on

Keith Beven[1,2,3]* and Andrew Binley[1]

[1] *Lancaster Environment Centre, Lancaster University, Lancaster, UK*
[2] *Department of Earth Sciences, Uppsala University, Uppsala, Sweden*
[3] *CATS, London School of Economics, London, UK*

Abstract:

This paper reviews the use of the Generalized Likelihood Uncertainty Estimation (GLUE) methodology in the 20 years since the paper by Beven and Binley in Hydrological Processes in (1992), which is now one of the most highly cited papers in hydrology. The original conception, the on-going controversy it has generated, the nature of different sources of uncertainty and the meaning of the GLUE prediction uncertainty bounds are discussed. The hydrological, rather than statistical, arguments about the nature of model and data errors and uncertainties that are the basis for GLUE are emphasized. The application of the Institute of Hydrology distributed model to the Gwy catchment at Plynlimon presented in the original paper is revisited, using a larger sample of models, a wider range of likelihood evaluations and new visualization techniques. It is concluded that there are good reasons to reject this model for that data set. This is a positive result in a research environment in that it requires improved models or data to be made available. In practice, there may be ethical issues of using outputs from models for which there is evidence for model rejection in decision making. Finally, some suggestions for what is needed in the next 20 years are provided. © 2013 The Authors. *Hydrological Processes* published by John Wiley & Sons, Ltd.

KEY WORDS uncertainty estimation; epistemic error; rainfall–runoff models; equifinality; Plynlimon

Received 19 April 2013; Accepted 30 September 2013

'*Unfortunately practice generally precedes theory, and it is the usual fate of mankind to get things done in some boggling way first, and find out afterward how they could have been done much more easily and perfectly.*'

Charles S Peirce, 1882

GLUE: THE ORIGINAL CONCEPTION

It is now 20 years since the original paper on Generalized Likelihood Uncertainty Estimation (GLUE[†]) by Beven and Binley (1992; hereafter BB92). The paper has now received over 1200 citations (as of December 2012) and been used in literally hundreds of applications. An analysis of the citations to the paper shows that interest was initially low, only much later did it become a highly cited paper as interest in uncertainty estimation in hydrological modelling increased. GLUE has also been the subject of significant criticism in that time, and some people remain convinced that it is a misguided framework for uncertainty estimation. In this paper, we review the origins of GLUE, the controversy surrounding GLUE, the range of applications, some recent developments and the possibility that it might become a respectable (in addition to being widely used) methodology.

The origins of GLUE lie in Monte Carlo experiments using Topmodel (Beven and Kirkby, 1979) carried out by Keith Beven when working at the University of Virginia starting around 1980. These were instigated by discussions with George Hornberger, then Chair of the Department of Environmental Science at University of Virginia, who, while on sabbatical in Australia and working with Bob Spear and Peter Young, had been using Monte Carlo experiments in analysing the sensitivity of models to their parameters (Hornberger and Spear, 1980, 1981; Spear and Hornberger, 1980; Spear *et al*., 1994). This Hornberger–Spear–Young (HSY) global sensitivity analysis method depends on making a decision between models that provide good fits to any observables available (behavioural models) and those that do not (non-behavioural models).

The first outcome of these early Monte Carlo experiments with rainfall–runoff models was to find that there were often very many different models that

*Correspondence to: Keith Beven, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ.
E-mail: k.beven@lancaster.ac.uk

[†]The acronym GLUE was produced while Keith Beven was still at the University of Virginia (until 1982) but did not appear in print until the Beven and Binley (1992) paper.

appeared to be equally behavioural judged by their error variance or Nash–Sutcliffe efficiency index values, measures that were commonly used in evaluating model performance at that time (Duan *et al*., 1992, also later came to a similar conclusion, and it was also evident in the set-theoretic water quality model calibration work of van Straten and Keesman (1991), Rose *et al*. (1991) and Klepper *et al*. (1991) (see also Spear, 1997). It should be remembered that hydrological modelling in the 1980s was still very much in the mode of finding the optimum model by the most efficient means. There was a rather common attitude that there should be 'the' model of a catchment, perhaps ultimately based on physical laws (Abbott *et al*., 1986a), but the best conceptual storage model might be useful in the meantime. There was not much in the way of uncertainty analysis of models; there was much more work on better optimization methods (as in Duan *et al*., 1992).

The Monte Carlo experiments suggested, however, that there was not a clear optimum but rather what came to be called an equifinality[‡] of model structures and parameter sets that seemed to give equally acceptable results (Beven, 1993, 2006, 2009a; Beven and Freer, 2001). In the context of optimization, the terms non-uniqueness, non-identifiable or ambiguity were used in the literature to reflect that this was considered to be a problem. During this period, also using a Monte Carlo framework, Andrew Binley examined the role of soil heterogeneity on a model hillslope response, using a 3D Richards' equation solution (Binley *et al*., 1989a). This study revealed that a single effective parameter for the hillslope (as assumed in many catchment models) might not be universally valid but rather state dependent (Binley *et al*., 1989b), also undermining the idea of finding an optimal model.

Another (not unexpected) outcome of these Monte Carlo experiments was that there was no clear differentiation between behavioural and non-behavioural models. There was instead generally a gradual transition from models that gave the best results possible to model that gave really rather poor results in fitting the available observations. Applications of the HSY sensitivity analysis method have consequently sometimes resorted to ranking models by some performance index (or magnitude of some output variable) and then taking the top X% as behavioural.

A further outcome was that the set of behavioural model predictions did not always match the observations. There could be many reasons for this: effectively all the different sources of uncertainty and error in the modelling process. Sources of uncertainty include the model structure, the estimates of effective parameter values, the input forcing and boundary condition data and the observations with which the model is being compared. These are also invoked as reasons why there seems to be some upper limit of performance for a set of models (even models with many fitting parameters) and why performance in 'validation' periods is often poorer than in calibration (Klemeš, 1986).

From this point, however, it was a relatively simple conceptual step to weight each of the behavioural models by some likelihood measure on the basis of calibration period performance and use the resulting set of predictions to form a likelihood weighted cumulative density function (CDF) as an expression of the uncertainty for any predicted variable of interest (Figure 1). Models designated as non-behavioural, for whatever reason, can be given a likelihood of zero and need not therefore be run in prediction. This is the basis for GLUE as expressed in the original BB92 paper setting out the method (see also Binley and Beven, 1991). It was a very different way of doing uncertainty estimation from the methods being used at the time of finding the optimum model on the basis of maximum likelihood, evaluating the Jacobian of the log-likelihood surface with respect to parameter variation around that point and using Gaussian statistical theory (this is before the Bayesian paradigm really became dominant in applications of statistical inference to environmental problems; the maximum likelihood approach is not nearly so computationally demanding given the resources available at the time). That is uncertainty estimation related to a point in the model space, and to the error characteristics associated with that maximum likelihood parameter set; in contrast, the GLUE method is a global method that (in most applications, but not necessarily) treats the complex error characteristics associated with each behavioural parameter set implicitly.

The BB92 paper had its origins in the analysis of distributed hydrological modelling of Beven (1989), which had originally been prepared as a comment on the papers by Abbott *et al*. (1986a,1986b) but which was reworked as a paper because the editors of the Journal of Hydrology at the time suggested it was too long to publish as a comment. As a paper, however, it had to do more than comment and made the suggestion that future work in this area should try to assess the uncertainty associated with the predictions of distributed models (refer also to Beven, 2001, 2002a,2002b). The paper of Binley *et al*. (1991) was the first attempt to do this using a distributed rainfall–runoff model. Recognizing the computational constraints of Monte

[‡]Equifinality in this sense first appears in the book on General Systems Theory by Ludwig von Bertalanffy (1968). It was first used in the context of hydrological modelling by Beven (1975) and in the paper of Beven (1993) to indicate that this was a generic problem rather than a problem of non-uniqueness or non-identifiability in finding the 'true' model of a catchment.

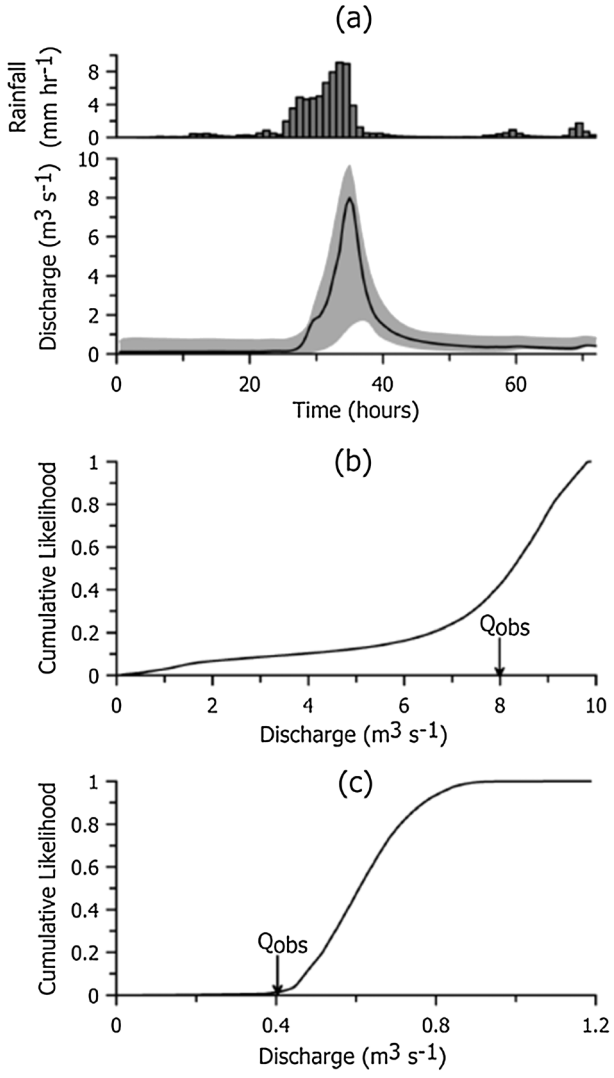

Figure 1. Example of Generalized Likelihood Uncertainty Estimation prediction bounds: (a) 5/95% limits for storm 1 in BB92, (b) cumulative likelihood for peak flow and (c) cumulative likelihood for flow at end of event. In (b) and (c), Q_{obs} indicates observed flow

Carlo simulations, they examined the method of Rosenblueth (1975) that requires only 2N + 1 simulations, where N is the number of parameters, in making an approximate estimate of prediction uncertainty. They concluded that the Rosenblueth sampling was only suitable as a first-order estimate. The Monte Carlo simulations in Binley *et al.* (1991), however, helped provide a framework for demonstrating GLUE in BB92. Binley *et al.* (1991) (and subsequently BB92) constrained their Monte Carlo sampling to 500 realizations even though they adopted a relatively simple distributed model [the Institute of Hydrology distributed model version 4 (IHDM4) of Beven *et al.*, 1987]. However, even to perform this level of computation at this time required the development of significant code enhancement in order to exploit a newly acquired 80 node transputer[§] parallel computer. Although this type of activity may be judged as routine nowadays, and even something that can be incorporated automatically by code compilers, in the 1980s, these studies were employing hardware and software that was extremely new to hydrological sciences and similar disciplines (although see also the earlier stochastic simulations of, for example, Freeze, 1975; Smith and Freeze, 1979; Smith and Hebbert, 1979).

BB92 set out the objective for GLUE to be generalized in the sense of using a range of potential likelihood measures and a range of ways of combining likelihood measures (not only Bayesian multiplication but also weighted addition, fuzzy union and fuzzy intersection). BB92 did include an attempt to make sampling more efficient (using a nearest neighbour technique to decide whether it was worth running a full simulation but with a random component analogous to the type of Metropolis–Hastings sampling that has become commonly used more recently, see below). It also included an assessment of the value of new data in inference using Shannon entropy and U-uncertainty measures.

Only very recently, our attention was drawn to the paper by Warwick and Cale (1988). That paper also drew on the HSY Monte Carlo method of sensitivity analysis. Model evaluation was based on user-specified limits of acceptability, similar to the set-theoretic model calibrations of Klepper *et al.* (1991) and Van Straten and Keesman (1991). Warwick and Cale (1988), however, added a weighting scheme when evaluating each model realization against observations, as in GLUE. In their case, however, the observations were synthetic, taken from the output of a model with the same structure as that being evaluated, so that there was a good chance of bracketing the synthetic observations. In that paper, they did introduce concepts of reliability and likelihood. Reliability was defined as the probability that a model would predict a system state to within the specified limits of acceptability; likelihood was defined as the probability of finding a model with a given reliability. They noted that the aim of a modelling exercise is to have a high likelihood of obtaining a highly reliable model. This is clearly easier for the synthetic case (refer also to Mantovan and Todini, 2006; Stedinger *et al.*, 2008).

[§]The transputer was a 1980s parallel computer, designed by David May at the University of Bristol and produced by Inmos, with chips designed to support pipes to other processors. The first floating point transputer, the T800, appeared in 1987. It was used here as TRAM daughter boards for PCs and programmed in a language called Occam.

THE GLUE CONTROVERSY

The range of options for model evaluation within the BB92 paper makes it clear that, given the multiple sources of uncertainty in the modelling process that are not well known, we did not think there was a single unique solution to the estimation of uncertainty. Any analysis would then be conditional of the judgments of the analyst appropriate to a particular problem.[¶] With hindsight, one regret in respect of the BB92 paper is that we did not also set out the use of a formal statistical likelihood within GLUE (even though this was performed not long after in the papers by Romanowicz *et al*., 1994; 1996 that were based on using explicit error models and formal Bayesian principles within GLUE). That might have avoided a lot of later misunderstanding and criticism of the methodology (that continues to this day, refer to Clark *et al*., 2011, 2012; Beven, 2012a).

BB92 comment that, '*We use the term likelihood here in a very general sense, as a fuzzy, belief, or possibilistic measure of how well the model conforms to the observed behaviour of the system, and not in the restricted sense of maximum likelihood theory which is developed under specific assumptions of zero mean, normally distributed errors……. Our experience with physically-based distributed hydrological models suggests that the errors associated with even optimal sets are neither zero mean nor normally distributed*' (p.281).

More recent applications of statistical inference to hydrological modelling have often been based on the use of formal likelihood functions but within a Bayesian framework (e.g. Kuczera *et al*., 2006; Vrugt *et al*., 2008, 2009a,2009b; Thyer *et al*., 2009; Renard *et al*., 2010; Schoups and Vrugt, 2010). This requires defining a formal model of the characteristics of the model residuals (or more generally, different sources of error) that then implies a particular form of likelihood function. It is now common within such an approach to include bias in the mean (or more complex 'model discrepancy' functions where structure is detected in residual series, Kennedy and O'Hagan, 2001). Autocorrelation in the residuals of hydrological models is common. Where this is strong, it can lead to wide uncertainty bounds when the model is used in simulation (e.g. Beven and Smith, 2013). The underlying assumption that the errors are, at base, essentially random in nature remains. Model predictions, and their associated error structures, are then weighted by their likelihood weights in forming a CDF of predicted variables. It can certainly be argued that this type of Bayesian inference is a special case of GLUE when the rather strong assumptions required in defining a formal likelihood function are justified (GLUE is indeed generalized in that sense). The error model then acts as an additional, non-hydrological, part of the model structure (as in Romanowicz *et al*., 1994).

This is, of course, controversial, and there are many hydrological modellers who have suggested quite the reverse that GLUE is just a poor approximation to formal Bayesian methods. In some cases, this view has been expressed very forcefully (e.g. Mantovan and Todini, 2006; Stedinger *et al*., 2008; Clark *et al*., 2011). The reason for this appears to be primarily that GLUE (in its use of informal likelihood measures) involves subjective decisions, and this is contrary to any aim of an objective hydrological science. This is despite the fact that Bayesian theory allows for subjectively chosen priors, and that in his original formulation, Bayes himself would have been accepting of subjective odds (or likelihoods) in evaluating hypotheses (e.g. Howson and Urbach, 1993). But, the priors become less important as more data are added to the inference, and a degree of objectivity can be claimed in verifying the assumptions made in formulating a likelihood by examination of the actual series of residual errors (although this is usually only performed for the maximum likelihood model, not the whole set of models with significant likelihood some of which could have quite different residual structures). If (but only if) the assumptions are verified, then the formal approach provides a means of estimating the (objective) probability of a future observation conditional on the model and its calibration data.

So, it may then seem perverse to lose this ideal of objectivity in GLUE if an informal likelihood measure is used (although we stress again that formal likelihoods *can* be used in GLUE if the strong assumptions can be justified). However, Beven *et al*. (2008) have shown how difficult it is to support this objective view, even for only small departures from the ideal case presented in Mantovan and Todini (2006). Real applications are not ideal in this sense (refer also to the discussions in Beven, 2006, 2010, 2012a; Beven and Smith, 2013). This makes the 'GLUE controversy' as much a matter of philosophical attitude to the treatment of different sources of uncertainty and error as it is an argument about whether one method is more appropriate than another (and a failure of a GLUE ensemble of models to bracket the observations can itself be informative, see below). In particular, we will argue that the hydrological consideration of error and uncertainty that can be incorporated into GLUE has some advantages over a purely statistical treatment, despite the apparent rigour and objectivity of the latter.

[¶]Jonty Rougier, a statistician at the University of Bristol, has suggested that because of this conditionality any assessment of uncertainty should be labelled with the name of the person or persons who agreed on the assumptions.

ALEATORY AND EPISTEMIC ERRORS

One reason for choosing *not* to use the formal statistical framework is that real applications may involve significant errors that result from a lack of knowledge (epistemic uncertainties) rather than simple random (aleatory) variability (for example, Helton and Burmaster, 1996; Allchin, 2004; Beven, 2009a; McMillan *et al*., 2010; Rougier and Beven, 2013; Rougier, 2013; Beven and Young, 2013). It is therefore somewhat surprising that it is suggested that modelling errors can be approximated by a predominantly aleatory structural model when we know that the input data to a model have non-stationary error characteristics and that these errors are then being processed through a complex nonlinear function (the model) with consequent non-stationary bias, heteroscedasticity and correlation. This view has been reinforced by studies of non-stationary data errors within the GLUE framework (e.g. Beven and Westerberg, 2011; Beven *et al*., 2011; Westerberg *et al*., 2011a,2011b; Beven and Smith, 2013). Ideally, of course, in any uncertainty estimation study, we would like to separate out the impacts of the different sources of error in the modelling process. This is, however, impossible, without very strong information about those different sources that, again for epistemic reasons, will not generally be available (for example, Beven, 2005, 2009a).

The important consequence of treating errors as aleatory when they are significantly epistemic is that the real information content of the calibration data is overestimated. This means that an (objective) likelihood function based on aleatory assumptions will overcondition the parameter inference (Beven *et al*., 2008; Beven and Smith, 2013) or inference about sources of uncertainty (e.g. Vrugt *et al*., 2008; Renard *et al*., 2010). Effectively, the likelihood surface is stretched too much. This is seen in the fact that the (objective) likelihoods for models with very similar error variances can be many orders of magnitude different if a large number of residual errors contribute to the likelihood function (as is the case with hydrological time series, see below). The resulting estimates of parameter variances will be correspondingly low. Taking account of autocorrelation in the residuals (expected for the reasons noted above) reduces this stretching, but the differences in likelihood between two similarly acceptable models can still be enormous. This is demonstrated later where different approaches to assessing the likelihood of a model are applied to the original example study of BB92. Stretching of the likelihood surface is one way of avoiding or greatly reducing equifinality of models and parameter sets but not because of any inherent differences in model performance, only because of the strong error structure assumptions and *even if* the best model found is not really fit for purpose.

It is, however, equally difficult to justify any particular subjective assumptions in choosing an informal likelihood measure (although refer to the discussion of Beven and Smith, 2013). Clearly, a simple measure proportional to the inverse error variance, inverse root-mean-square error or inverse mean absolute error, as proposed in BB92 will not stretch the surface so much (unless a near to perfect match to the data is obtained, unlikely in hydrological modelling) but perhaps is likely to *underestimate* the information content in a set of calibration data. How do we then achieve some (objective as possible) compromise that has an equally good but more realistic theoretical basis to formal likelihood functions? GLUE is already a formal methodology in that the choice of any likelihood measure must be made explicit in any application, such that it can be argued over and the analysis repeated if necessary, but it remains difficult to define a likelihood measure that properly reflects the effective information content in applications subject to epistemic errors. This is, of course, for good epistemic reasons!

In BB92, this was expressed as follows: '*The importance of an explicit definition of the likelihood function is then readily apparent as the calculated uncertainty limits will depend on the definition used. The modeller can, in consequence, manipulate the estimated uncertainty of his*[‖] *predictions by changing the likelihood function used. At first sight, this would appear to be unreasonable, but we would hope that more careful thought would show that this is not the case, provided that the likelihood definition used is explicit. After all, if the uncertainty limits are drawn too narrowly then a comparison with observations will suggest that the model structure is invalid. If they are drawn too widely, then it might be concluded that the model has little predictive ability. What we are aiming at is an estimate of uncertainty that is consistent with the limitations of the model(s) and data used and that allows a direct quantitative comparison between different model structures*' (p.285).

Our view of this has changed surprisingly little in 20 years (except that we might now reserve the term likelihood function for formal likelihoods and instead use likelihood measure in GLUE applications using informal likelihoods and limits of acceptability). We do now have a greater appreciation of the potential for model predictions to exhibit significant departures from the observations during some periods of a simulation. This was not apparent in the original event by event simulations of BB92 but we did say that, '*If it is accepted*

[‖]We would not, of course, wish to imply that hydrological modellers might be exclusively masculine, and it was not true then. In the UK, Cath Allen, Liz Morris, Ann Calver, Hazel Faulkner, Caroline Rogers, Alice Robson, Sue White and others had already made valuable contributions to hydrological and hydraulic modelling.

that a sufficiently wide range of parameter values (or even model structures) has been examined, and the deviation of the observations is greater than would be expected from measurement error, then this would suggest that the model structure(s) being used, or the imposed boundary conditions, should be rejected as inadequate to describe the system under study' (p.285). In many applications, there have been cases where *none* of the behavioural simulations provided predictions close to some observations to be considered as acceptably behavioural so that all the models tried could be rejected as unacceptable or non-behavioural (e.g. Page *et al*., 2007; Dean *et al*., 2009), even in some cases where global performance was actually rather good (Choi and Beven, 2007).

It must also not be forgotten that such failures might not be because the model structure is problematic but because the input and evaluation data are inconsistent during some parts of the record (e.g. Beven, 2005, 2010, Beven *et al*., 2011; Beven and Smith, 2013). All too often, data are provided and used *as if* error free when they are subject to significant (aleatory and epistemic) uncertainties. Discharges, in particular, should generally be treated as virtual rather than real observables (Beven *et al*., 2012b), whereas rainfall estimates over a catchment area can be poorly estimated for individual events by either raingauges or radar methods. Both models and data exhibit forms of epistemic error, as well as being subject to random variability (Beven, 2012a; Beven and Young, 2013). Epistemic error will generally be transitory, non-stationary and non-systematic. This explains, at least in part, the overestimation of the information content by assuming that errors are aleatory with (asymptotically) stationary distributions.

BAYES, GLUE AND THE PROBLEM OF INDUCTIVE INFERENCE

These issues are, in fact, a variant on Hume's problem of induction (e.g. Howson, 2003). How far can past historical data provide belief that we will observe similar occurrences in the future? Hume's argument was that past occurrences should not engender belief in future occurrences, surprises might always happen. The most recent popularization of Hume's problem is the 'black swans' concept of Taleb (2010). The suggestion that models calibrated to past historical data might be useful in informing us about the potential future behaviour of a catchment is a form of induction (Beven and Young, 2013). There are many examples, of course, where scientific theory has been used to predict future behaviour successfully. It is intrinsic to Popper's falsificationist approach to the scientific method where models that do not survive such tests should be rejected. It is difficult, however, to be strictly falsificationist when epistemic errors increase the possibility of rejecting a model that might be useful in simulation or forecasting, just because it has been evaluated using forcing data with errors (a Type II error). Epistemic error in the forcing data and observations could also lead to a model that would not useful in prediction not being rejected (a Type I error, Beven, 2009a, 2010). There has never been a successful philosophical explanation of why Hume's problem of induction is not correct. Howson (2003) argues that it is, in fact, correct, but its impact can be mitigated by Bayesian reasoning.

The variant to be considered here in hydrological reasoning about uncertainty is how far we should expect (high impact) surprises in future observations when we have epistemic (that is non-random but not necessarily systematic) errors in inputs, models, parameter values and observations. In hydrological systems, constrained by water and energy balances, we should expect some surprises, but we do not expect very great surprises if both inputs and outputs are estimated well. The constraints mean that catchment systems are not expected to respond in grossly chaotic ways (although such cases are known under extreme conditions; the volcanically induced jøkulhaups of Iceland; the catastrophic channel changes of the Yellow River in China; river network capture as a result of erosion in an extreme event; the sudden drops in river discharge after prolonged drought resulting from breakdowns in subsurface connectivity, none of which would normally be considered in a hydrological simulation model used in predicting hydrological impacts of future change). Such changes, and more general failure to estimate future boundary conditions, always provide a *post hoc* justification for model failure (refer to the groundwater modelling examples in Konikow and Bredehoeft, 1992), but their acknowledgement as deep uncertainties might also be important to the decision making process (Faulkner *et al*., 2007; Beven, 2011, 2012a; Ben-Haim, 2012).

Although we might not expect many great surprises in catchment behaviour under some normal conditions, we do expect deviations that may limit predictability (such as the timing error in predicting snowmelt 1 year out of four in Freer *et al*., 1996, or the event runoff coefficients of greater than 1 inferred from the observational data in Beven *et al*., 2011) in ways sufficient to suggest that the type of errors seen in calibration might be arbitrarily different to those that appear in future simulations (Kumar, 2011; Montanari and Koutsoyiannis, 2012; Beven and Young, 2013). Thus, can treating errors *as if* they are asymptotically convergent on some underlying distribution (as required in the use of a formal Bayesian likelihood) ever be an adequate assumption (refer to the

extended discussion in Beven, 2012a)? This is why such an assumption should be expected to overestimate the information content of a set of calibration data and consequently overstretch the likelihood surface.

It does not, however, resolve the question of how far should the likelihood surface be stretched, and how far prediction limits should allow for epistemic error. BB92 allowed some flexibility in this respect by defining likelihood measures with shaping factors '*to be chosen by the user*' (see also Beven and Freer, 2001) but without guidance as to what values those factors might take. This is clearly a matter of the relative importance of epistemic and aleatory uncertainties expected in the data for calibration and prediction periods, but this then first requires a means of separating such errors which is not possible without independent information about different sources of uncertainty (e.g. Beven, 2005, 2006, 2012b).

Howson (2003) suggests that a Bayesian approach can be useful in such problems involving induction by providing a deductive logic in changing modellers' beliefs that is consistent with the probability axioms. The criticisms of the overconditioning using statistical likelihood functions are *not* a criticism of a Bayesian approach to conditioning (see also the discussion of outliers in Kuczera *et al*., 2010). GLUE includes the possibility of using Bayes multiplication in conditioning (but is only one of the options for combining likelihoods in a more general learning process suggested in BB92). What we are suggesting is that the definition of appropriate likelihood measures needs to be revisited to more properly represent the information content of calibration data sets in the face of epistemic uncertainties, essentially as a form of engineering heuristic (Koen, 2003). We will return later to the question of how this might be achieved in considering future work in this area.

SO WHAT DO GLUE PREDICTION LIMITS REALLY MEAN?

The result of a GLUE analysis is an ensemble of behavioural models, each associated with a likelihood value. The likelihood values should be a reflection of the belief of the modeller in a particular model as a useful predictor for the future. This might include both prior beliefs and a modification of prior belief on the basis of performance in calibration as appropriate. Where calibration data are available, then each model is also associated with a series of residual errors. As noted in the previous texts, these residuals might have complex structure as a result of epistemic error. In the GLUE methodology, it is (nearly always) assumed that these residuals can be treated implicitly in prediction, although the empirical distributions of such errors can also be used (e.g. Beven and Smith, 2013). It is then assumed that the nature of the residuals is expected to be similar in prediction as in calibration (this is similar to the assumption in a statistical methodology that the hyperparameters of an error model determined in calibration will also hold in prediction). Thus, if a model is consistently underpredicting under certain circumstances in calibration, it is assumed that it will similarly underpredict under similar circumstances in prediction. If a model is consistently overpredicting under certain circumstances in calibration, it is expected that it will similarly overpredict under similar circumstances in prediction. This allows that the residual errors might have an arbitrarily complex structure but cannot allow for new forms of epistemic error in prediction (but neither can any statistical model). Alternatively, the errors can be represented explicitly, either by a parametric error model (as in Romanowicz *et al*., 1994, 1996) or non-parametrically given the distributions of errors determined in calibration (as in Beven and Smith, 2013).

Prediction limits are then determined in GLUE by forming the CDF of the likelihood weighted ensemble of simulations (including model errors if an explicit error structure model is used). Any required quantiles can then be taken from the CDF (for example, Figure 1). These will be quantiles of the simulated values and do not imply any expectation that future observations will be covered by the CDF, except implicitly to some similar level to that found in calibration. Past experience suggests that this can give useful coverage of predicted observations for cases where the ensemble of models is able to span most observations in calibration. One advantage of this approach is that because no model will predict negative discharges, the prediction limits never fall below zero (as is sometimes the case under Gausssian assumptions of symmetric error distributions with large error variances, albeit that error transformations can be used to mitigate this problem, e.g. Montanari and Brath, 2004; Montanari and Koutsoyiannis, 2012). A second advantage is that it can allow for non-stationarity in the distribution of simulated values under different hydrological conditions, including changes in both variance and form of the distribution (as shown in Figure 1b and c, see also Figure 2 of BB92 and Freer *et al*., 1996).

There are, however, applications where it is clear that the range of models tried cannot match particular observations in either calibration or validation. This could be because of model structural error, or, as noted earlier, it could be because of epistemic errors in the inputs (Beven and Westerberg, 2011; Beven *et al*., 2011; Beven and Smith, 2013). In either case, it is informative because it means that errors in the modelling process are not being hidden within a statistical error variance (or non-parametric distribution of errors). It suggests that

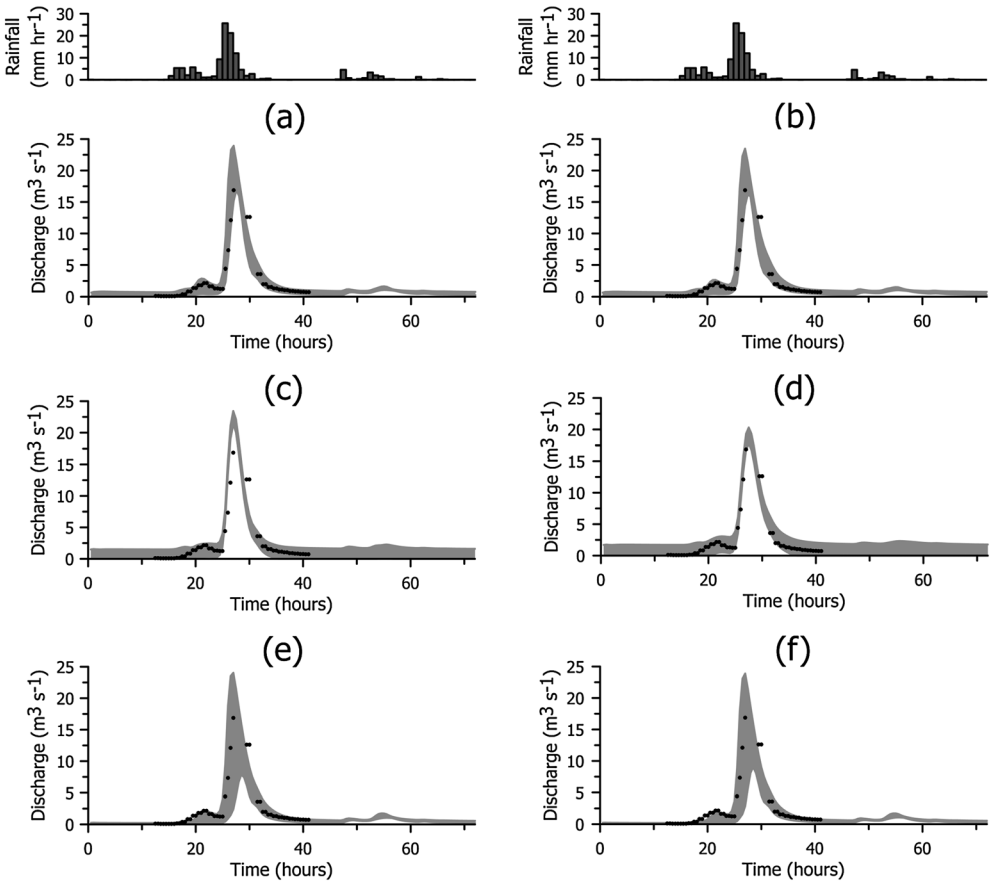

Figure 2. Example uncertainty limits for different likelihood measures. Limits are for storm 4 in BB92. Shaded area shows 5/95% limits; symbols show observed discharges. Left column shows limits on the basis of likelihoods derived from knowledge of storms 1–3; right column shows limits on the basis of knowledge of storms 1–4. (a) and (b) Residual variance-based measure, as in BB92. (c) and (d) Formal likelihood function (combined parameter and statistical error weighted by likelihood).. (e) and (f) Weighted least squares function assuming constant measurement error equal to 20% observed flow. (refer to Table III for likelihood function definitions)

either the model might be improved or that some of the observations might need further investigation as to whether they are disinformative for model inference.

The prediction bounds are, however, always conditional on the assumptions on which they are based: particularly the prior distribution of models run and the choice of likelihood measure (including any decision about differentiating behavioural and non-behavioural models). Given the possibility for epistemic error in the modelling process, these assumptions might be more or less 'objective' but must be made explicit. They are thus open to discussion, review and change if deemed inappropriate in a similar way to statistical error assumptions. Such review should be an important part of the modelling process but is often neglected, even where statistical assumptions are clearly not met. Feyen *et al*. (2007) is one example where such an evaluation revealed inappropriate error assumptions in a statistical likelihood, but this did not then lead to revision of the uncertainty estimation (see also the comment of Beven and Young, 2003).

RANGE OF APPLICATIONS AND EFFICIENCY OF SAMPLING

Over the last 20 years, the computing power available to hydrological scientists has increased dramatically. This has allowed the type of simple Monte Carlo sampling that GLUE requires to be applied to an ever wider range of models, even if it is still not possible to run sufficient samples for some more computationally demanding models. Analysis of the full range of GLUE applications (refer to the listings of the Electronic Appendix for this paper) reveals that the majority of the applications to date have been in rainfall–runoff modelling (as was the case study of BB92). There have also been significant numbers of applications in hydraulic modelling, water quality modelling, flood frequency estimation, urban and stormwater hydrology, soil and groundwater modelling, geophysics and ecology.

There have been only a small number of studies within the GLUE framework that have used an explicit error model (Romanowicz *et al*., 1994, 1996; Xiong and

O'Connor, 2008; Beven and Smith, 2013). Most studies have used an informal likelihood of some type, and most commonly, the efficiency measure of Nash and Sutcliffe (1970) has been used, with a threshold value to define the set of behavioural models. The efficiency measure has some important limitations as measure of performance for cases where the residuals can be assumed aleatory (Beran, 1999; Schaefli and Gupta, 2007; Smith *et al*., 2008; Gupta *et al*., 2009) but remains a widely used performance measure. It is important to remember, however, that GLUE is more general than using an efficiency measure and threshold with uniform prior parameter distributions. Other priors and other measures can be used. Some recent applications have returned to set membership evaluation of models, on the basis of limits of acceptability and fuzzy measures to define possibilities as a way of trying to allow for the complex characteristics of sources of epistemic error.

Although computer constraints on the application of GLUE have been relaxed (in comparison to the late 1980s), it remains an issue, either because of a model that is particularly slow to run so that it is still not possible to sample sufficient realizations or because of a high number of parameter dimensions. The most runs used in a GLUE application that we know of are the two billion runs in Iorgulescu *et al*. (2005, 2007), of which 216 were accepted as behavioural using a limits of acceptability approach. This was for a model that was just a few lines of code but which had 17 parameters. Two billion runs is then still a small sample compared with a discrete sampling strategy with ten values for each parameter. As stated earlier, in BB92, we were constrained to 500 realizations for each (relatively short) event and that was only possible in a reasonable time because we utilized an 80 node transputer system. More recently, GLUE calculations have been speeded up for certain models using highly parallel graphics processor cards (for example, Beven *et al*., 2012c)

It is possible to use adaptive sampling strategies to seek out areas of higher likelihood or possibility in the parameter space. It has already been noted that BB92 already used a strategy on the basis of nearest neighbour interpolation. Early on, Spear *et al*. (1994) suggested a space partitioning system as a way of improving the density of sampling behavioural models. Khu and Werner (2003) have proposed a method on the basis of genetic algorithm and artificial neural network techniques to map out the areas of high likelihood in the model space, whereas Blasone *et al*. (2008a,2008b) and McMillan and Clark (2009) have suggested combining GLUE and Markov chain Monte Carlo (MCMC) strategies to increase the efficiency of finding behavioural models. The DREAM algorithm could also be used in this context (e.g. Vrugt *et al*., 2009a; Laloy and Vrugt, 2012). Where strong information about prior parameter distributions is available, sampling strategies such as Latin hypercube or antithetic sampling can be used to reduce the number of runs required to represent that prior information (e.g. Avramidis and Wilson, 1996; Looms *et al*., 2008). With some of these techniques, it is not always clear just what density of sampling results and therefore whether the likelihood associated with a model should be modified to reflect sampling density (this is one advantage of either uniform sampling or methods that successfully achieve likelihood dependent sampling densities).

Efficiency, however, is not only a matter of the effectiveness of a search technique but also of the complexity of the response or likelihood surface. Model structures with thresholds, numerical artefacts (e.g. Kavetski and Clark, 2010), complex interactions between parameters, interactions between particular data errors and model performance and other factors can results in surfaces that are complex in shape. Sensitivity and covariation between parameters and the likelihood measure will also be complex and will not always be represented well by a simple covariance function. That means that the success and efficiency of a search technique might well depend on the initial sampling that is the basis for refining the search in successive iterations. If localized areas of high likelihood are not sampled in that initial (limited) sampling, then there is a possibility that they will never be sampled. That is why many sampling methodologies, including Markov chain Monte Carlo methods, and the BB92 nearest neighbour method include a probabilistic choice of making a model run, even if the parameter set is not necessarily predicted as being behavioural. This maintains a possibility of identifying areas of high likelihood that have not yet been sampled.

There remain many models that simply take too long to run or have too many parameter dimensions to allow adequate sampling of the model space. In some cases, the use of MCMC or other efficient sampling strategies within GLUE might help, especially when it is expected that the likelihood surface being sampled is smooth. However, it also seems likely that computer power available to the modeller will continue to increase faster than either modelling concepts or data quality. This will allow the application of GLUE type methods to a wider range of problems in the future.

ATTAINING RESPECTABILITY?

Despite the wide range of past applications of GLUE (refer to the Electronic Appendix to this paper), it seems that it is still not considered fully respectable. Criticism has focussed on the subjective assumptions required,

particularly in choosing a likelihood or way of combining likelihoods, which means that the resulting uncertainty estimation is conditional on those assumptions. There is no way of objectively verifying the probability of predicting a future observation (as in the case of evaluating a formal likelihood) because, as noted earlier, the GLUE prediction bounds do not generally have this meaning unless a valid explicit error model is used. But if formal likelihoods do overestimate the information content of the calibration data in real, non-ideal, examples, we would wish the choice of likelihood measure in GLUE to reflect the real information content in some way.

We note at this point that the choice of a Gaussian (L^2 norm) likelihood in statistical inference is, in itself, a subjective choice. Independently, Laplace (1774) had developed a form of analysis of errors, analogous to Bayes, but based on the absolute error (L^1 norm). There are equally other possibilities (Tarantola, 2006). In the 19th century, analytically tractability was all important, and the L^2 norm had many advantages in this respect but *any* of these norms can very easily be applied on modern computers. So, there is a choice that clearly should reflect belief in the information provided by a single residual, but what is not yet clear is what type of likelihood measure is most appropriate given the epistemic errors in the typical data sets used for inference in hydrological applications, and how that might be checked in simulation. The GLUE methodology is, however, general to all these different choices, from the most formal to the most informal when, if epistemic error is important, there will be no right answer (again, for good epistemic reasons).

Other disciplines have had to struggle with similar problems of information content and identifiability. Diggle and Gratton (1984) provide an early example of statistical inference for intractable error models. Later, the name approximate Bayesian computation (ABC) was given to a technique (actually somewhat analogous to a form of GLUE) developed for evaluating models in genetics for cases where a suitable formal Bayes likelihood function is difficult to define or evaluate (e.g. Tavaré *et al*., 1997; Beaumont *et al*., 2002). As in some GLUE applications, MCMC methods have been used to increase the efficiency of sampling a complex model space (Marjoram *et al*., 2003). It can be shown that, at least for certain problems, ABC can provide an asymptotic approximation to a formal Bayesian likelihood analysis (though that might be misleading for hydrologists where, as argued earlier in this paper, a classical formal Bayesian analysis might *not* be what is required). Toni *et al*. (2009) and Marin *et al*. (2011) provide reviews of ABC methods while interpretations of GLUE as a form of ABC have been presented by Nott *et al*. (2012) and Sadegh and Vrugt (2013).

But, there might be another way of gaining respectability and being more objective in applications of GLUE and that is to change the strategy of model evaluation to an approach that does not depend directly on model residuals but is the result of hydrological reasoning. This has been the subject of some recent developments in GLUE (see below).

REVISITING THE GWY: A COMPARISON OF LIKELIHOOD MEASURES WITHIN GLUE

To illustrate some of the topics that have been discussed in the previous texts, we have returned to the example application in BB92 where a number of storms were modelled for the small Gwy catchment in mid-Wales using the IHDM4 (Beven *et al*., 1987). IHDM4 is based on a 2D finite element solution of the Darcy-Richards equation for variably saturated flow in the subsurface. It does not explicitly represent macropores or other preferential flow processes but in BB92 was shown to produce reasonable catchment scale simulations. It may seem strange now that the study was limited to single storm simulations, but in the 1970s and 1980s, this was quite common, particularly in applications of distributed models (again in part for computational reasons). We can now reinterpret the exercise as a form of model fitting for an ungauged catchment where field campaigns are mounted to obtain rainfall and stream discharge data for a small number of events. This is one strategy to address the prediction of ungauged basins problem (Juston *et al*., 2009; Seibert and Beven, 2009; Blöschl *et al*., 2013). The characteristics of the storms used are given in Table I. The calibration parameters used were the same as in BB92 (Table II).

In this example, we have used the same storms as before, using GLUE to update the likelihood weights for different parameter sets as each new storm is added to the measurement set. As in BB92, we do not use the storms as they occurred but instead according to their number (e.g. storm 1 before storm 2). The reason for doing this was that in the original BB92 paper, we were interested to see how a model, calibrated on a series of similar sized

Table I. Storm characteristics for the Gwy catchment simulations

Storm	Date	Total rainfall (mm)	Peak flow ($m^3 s^{-1}$)
1	17–19 November 1981	80.5	8.0
2	27–29 January 1983	111.4	6.1
3	11–13 February 1976	107.3	8.5
4	5–7 August 1973	121.8	16.8*

*Estimated

Table II. Institute of Hydrology distributed model version 4 parameters and their ranges for the Gwy catchment simulations

Parameter	Description	Minimum	Maximum
K_s	Saturated hydraulic conductivity ($m\,h^{-1}$)	0.02	2.00
θ_s	Saturated moisture content ($m^3\,m^{-3}$)	0.15	0.60
φ_{in}	Initial soil moisture potential (m)	−0.40	−0.05
f	Overland flow roughness coefficient ($m^{0.5}\,h^{-1}$)	50.00	10 000.00

events, would perform for an event of different magnitude (storm 4 in this case, Table I). Each storm can also be used as a validation event before being incorporated into the calibration data set (e.g. left column in Figure 2). This time, we have incorporated a range of likelihood measures, including a statistical likelihood function (Table III). Note that for the formal likelihood, no model will be rejected as non-behavioural but where N_t is large all the model likelihoods will be very small and potentially subject to rounding error. Thus, as is usual practice, the calculations are made using the log likelihood. All models with likelihood values smaller than the maximum likelihood by 100 log units are neglected. The remaining likelihoods were back-transformed and rescaled to a cumulative of unity. A formal likelihood that includes a lag 1 autocorrelation component was also tried but made little difference to the results. In both cases, only a small number of models contribute significantly to the cumulative likelihood because of the stretching of the likelihood surface induced by the formal likelihood function.

As each storm is added into the conditioning process, the likelihoods are combined multiplicatively using Bayes equation in the following form:

$$L_p\left[M|y\right] = \frac{L_o[M]L_y\left[M|y\right]}{C} \tag{1}$$

where M indicates a model structure – parameter set combination, y is a set of observations with which the model outputs are compared, $L_o[M]$ is the prior likelihood for that model, $L_y[M|y]$ is the likelihood value arising from the evaluation, $L_p[M|y]$ is the posterior likelihood and C is a scaling constant such that the sum of the posterior likelihoods over all behavioural models is unity. We note again that GLUE can be Bayesian in this way but that it is not limited to Bayes equation in combining likelihoods. In BB92, we also suggested that weighted addition, fuzzy intersection and fuzzy union might also be used as ways of combining different likelihood measures that might come from evaluations on different periods of observations or quite different types of observations. One feature of this multiplicative combination is that if a model is non-behavioural on any evaluation ($L_y[M|y] = 0$), then the posterior likelihood for that model will be zero regardless of how well it has performed on earlier evaluations.

Figure 2 shows how the different likelihood measures have contrasting characteristics in their uncertainty limits, and their evolution as new data becomes available. As stated earlier, for our original analysis (in BB92), we were constrained by the number of model runs that could be performed. In revisiting this case study, we examined whether the 500 realizations originally used was an appropriate number. In terms of capturing the uncertainty limits (e.g. Figures 1 and 2), increasing the number of model runs using the BB92 efficiency based likelihood measure has little effect. However, if one is to consider resampling the parameter space, as demonstrated in BB92 using a relatively simple interpolation scheme, then such a small number of realizations (even for a four parameter study) could be inadequate.

Figure 3 shows dotty plots for two parameters for the BB92 efficiency based likelihood measured after adding the data from storm 1 to storm 4 into the inference process (c.f. Figure 7 in BB92). For 500 realizations (Figure 3a), a pattern is not evident when collapsed into 2D space. One may interpret this as an indication of multiple local

Table III. Example likelihood measures

Likelihood measure	Equation	Notes
Sum of squares of residuals	$L \propto \left(\sigma_\epsilon^2\right)^{-N}$	σ_ϵ^2 = variance of residuals. As in BB92, $N = 1$ is used.
Formal likelihood	$L \propto \left(2\pi\sigma_\epsilon^2\right)^{-N_t/2} exp\left[-\frac{1}{2\pi\sigma_\epsilon^2}\left(\sum_{t=1}^{N_t}\epsilon_t^2\right)\right]$	N_t = number of observations. ϵ_t is the residual at observation t.
Weighted least squares	$L \propto \sum_{t=1}^{N_t}\left(\epsilon_t/\varepsilon_t\right)^2$	ε_t is the measurement error of observation, fixed at 20% of flow at time t.

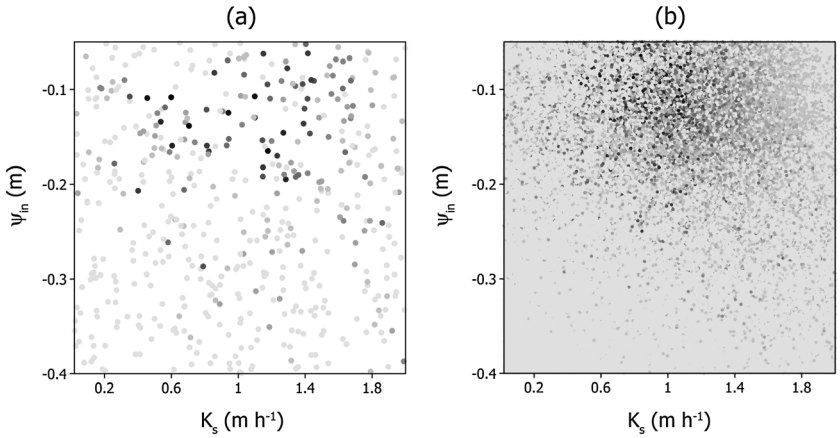

Figure 3. Dotty plots showing likelihood measures in 2D parameter space: (a) 500 realizations (as in BB92) and (b) 500 000 realizations. Likelihood measure computed as in BB92. Example shown is for storm 4, computed using observations from storms 1 to 4. Symbol colour indicates magnitude of likelihood measure (white, low; black, high)

optima. However, if we perform the same exercise for 500 000 model runs (Figure 3b), a clearer pattern emerges. Resampling based on only 500 realizations (for this four parameter case) may, therefore, be inefficient or may misguide the parameter search. Note, however, that in this case, there is little difference in the uncertainty limits estimated for the two sets of samples.

As the IHDM4 model study involves only four parameters, we can explore, visually, the parameter space further. Figure 4 reveals the variation in likelihood measure within the 4D space, shown as an isosurface in two 3D parameter plots. Figures 4a and b show the isosurface of a given likelihood measure given data from storms 1 to 3. Figure 4c and d, in contrast, show how the isosurface changes as data from storm 4 is incorporated.

Another way of examining the effect of likelihood measure on parameter conditioning is in terms of parameter distributions as new information is added after each storm. Figure 5 shows, for each of the parameters, the modification of the prior uniform distribution after adding only storm 1 and then after adding storms 1–4. Figure 5a shows the posterior cumulative distributions for each parameter using the BB92 efficiency-based likelihood measure; Figure 5b shows the equivalent distributions for the formal likelihood function. The difference in the degree of conditioning is immediately obvious. Even after only storm 1 is added, the formal likelihood surface has been stretched to focus in on a highly constrained range for each parameter, to the extent that the area of higher likelihoods is not that well defined even with a sample of 500 000 runs (adaptive density dependent sampling might help in that respect but would be unlikely to greatly expand the range of the posterior distributions). This range also changes from storm 1 to storm 4 for the φ_{in} and f parameters. This might be expected for φ_{in} because this defines the initial condition for each event, but f is intended to be a characteristic of the catchment soils. Such jumps are possible within the formal Bayesian framework, because models are never given zero likelihood, only very low values. That means that, as new information is added, a model might reflect the changing nature of the errors by recovering to a higher likelihood (and vice versa). It can also, however, be interpreted as an indication of severe overconditioning because of the formal likelihood assuming that the information in the series of residuals is the result of an aleatory process.

A limit of acceptability approach to model evaluation

In the Manifesto for the Equifinality Thesis, Beven (2006) suggested that a more hydrologically rigorous approach to model evaluation that takes proper account of observational data errors and is not based only on model residual errors, might be based on specifying limits of acceptability for individual observations with which model outputs would be compared (refer also to Beven, 2012a,2012b). This approach has since been used, for example, by Dean *et al*. (2009), Blazkova and Beven (2009), Liu *et al*. (2009), Krueger *et al*. (2009), McMillan *et al*., 2010, and Westerberg *et al*. (2011b). There had also been earlier forms of this approach within GLUE based on fuzzy measures (for example, in Blazkova and Beven, 2004; Freer *et al*., 2004; Page *et al*., 2003, 2004, 2007; Pappenberger *et al*., 2005, 2007).

Within this framework, behavioural models are those that satisfy the limits of acceptability for each observation. Ideally, the limits of acceptability should reflect the observational error of the variable being compared, together with the effects of input error and commensurability errors resulting from time or space scale differences between observed and predicted values. They might also reflect what is needed for a model to be fit-for-

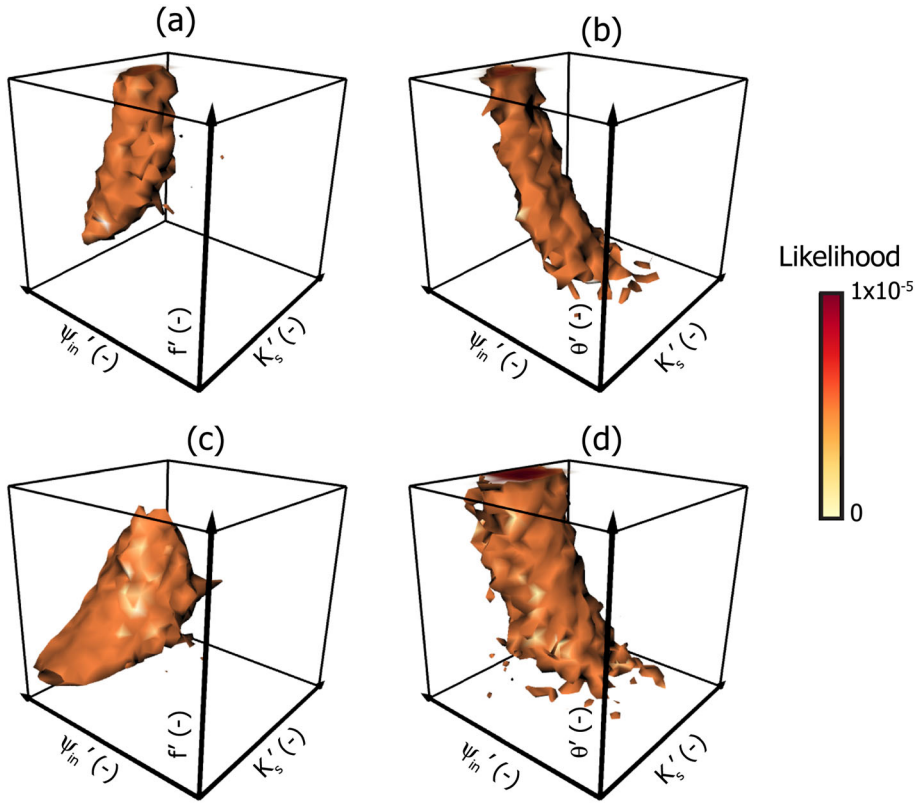

Figure 4. Isosurface of interpolated likelihood function for 500 000 realizations of two storm runs. Isosurface shown for central likelihood values [$>5\times10^{-6}$, equivalent to a residual variance of 1.38 $(m^3 s^{-1})^2$]. Likelihood measure computed as in BB92. (a) and (b) show variation isosurfaces computed on the basis of storm 1. (c) and (d) show isosurfaces computed on the basis of storm 1–4. The parameter values are normalized to the respective range for plotting purposes using the sampling ranges shown in Figure 5

purpose for a particular decision making process. Ideally, the limits should be set independently of any model structure and prior to making any model runs, although this is clearly difficult in allowing for the effects of input error. In extreme cases, hydrological inconsistencies between input and output data for specific events might mean that certain events are disinformative in respect of model evaluation (Beven *et al*., 2011; Beven and Smith, 2013). Setting limits of acceptability before running the model on the basis of best available hydrological knowledge might be considered more objective than the analogous use of a maximum absolute residual, which was also one of the measures proposed in BB92.

This approach has also been used here with the original Gwy application from BB92. Hudson and Gilman (1993) suggest that errors in stream gauging might be of the order of 3% for flows contained within the gauging structures, and errors in estimating catchment average rainfalls might be of the order of 4%. The latter estimate, however, makes use of an extensive network of ground level monthly storage gauges so that the uncertainty associated with individual storms might be significantly higher. An Institute of Hydrology report of Newson (1976) suggests that catchment average rainfalls for the site might be estimated to within 5%. An earlier report from Clarke *et al*. (1973) suggests that the coefficient of variation for hourly rainfalls on the basis of the recording gauges (albeit estimated only for wet spells in 1 month) was greater, of the order of 50%. The estimate for discharge uncertainties could increase dramatically when the structures were overtopped or by-passed during extreme events, but all the events considered here were within the capacity of the Gwy structure. Marc and Robinson (2007) also suggest that changes in the accuracy of flow gaugings for the nearby Tanllwyth subcatchment might have been sufficient to have had an effect on longer-term water balance estimates.

Here, lacking adequate knowledge of the input errors associated with each event, we have not made any attempt to account explicitly for input errors. Instead, we have specified limits of acceptability on the basis of ±10% and ±20% for the observed discharges, to make allowance for the potential effects of input error. These limits might be seen as generous, but in contrast to the earlier global evaluations of the IHDM model in this application, all 500 000 simulations fail to meet the limits of acceptability in storm 1 and all the other storms as well. Figure 6 shows the results for the best 100 runs in the form of a cumulative

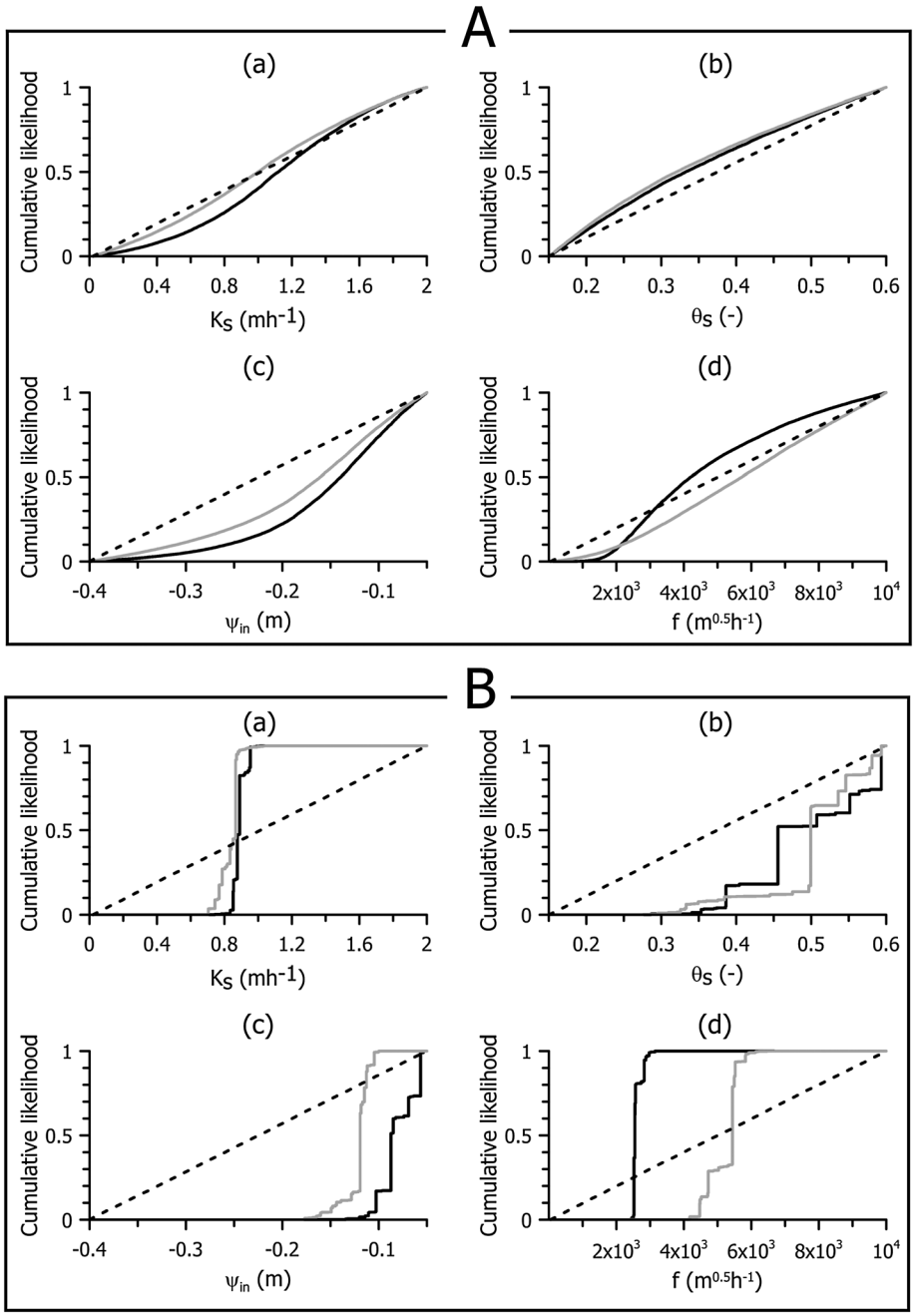

Figure 5. (a) Change in efficiency-based likelihood function with event for the four parameters. (b) Change in formal likelihood function with event for the four parameters. Dashed lines show prior likelihood; grey lines show likelihood after storm 1; solid black lines show likelihood after storm 4

distribution of model residuals normalized for the limits of acceptability. The very best model achieves only 82% compliance with the limits of acceptability (normalized misfit <1) over all the observed discharges in storm 1.

Clearly, despite the rather relaxed limits, the limits of acceptability evaluation is more demanding than the earlier global evaluations in this case. This could be in part because of the approximate initial conditions affecting the low flow simulation of each storm, or an underestimate of the effect of input error, or IHDM model structural error, based as it is on a purely Darcy-Richards subsurface flow process representation. There is also an issue about whether the use of a percentage error is appropriate to define whether a model is fit for purpose. This makes some allowance for the potential heteroscedasticity of discharge observation errors at higher flows, but even 20% limits of acceptability at low flows might be a very small error.

Thus, the analysis was revisited using the same 20% limits but with restricted to a minimum misfit from the

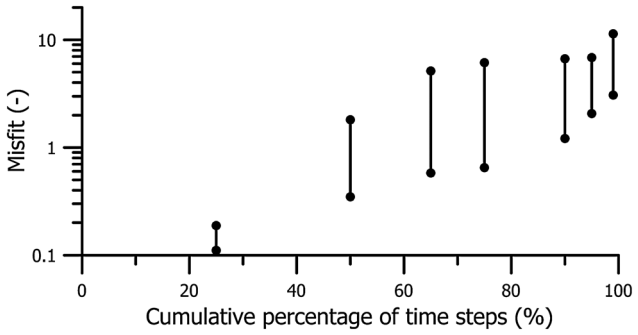

Figure 6. Results of limits of acceptability analysis for the best 100 (out of 500 000) runs of Institute of Hydrology distributed model version 4 for storm 1. Best here is defined by the number of time steps for which the model satisfies the limits of acceptability. Misfit is specified as a normalized absolute scale where unity represents the upper or lower 20% limit around each observation. The vertical bars indicate the range of misfit for the best 100 realizations, for a given limit of acceptability (expressed as a percentage)

observed discharge. Even allowing for minimum limits of acceptability of up to $2\,m^3\,s^{-1}$ did not result in any behavioural simulations for any storm. This is an indication that, although the global likelihood measures used earlier result in relatively constrained uncertainty bounds, they have the effect of averaging over the detailed time step error characteristics of the individual models, which do not satisfy the relaxed limits of acceptability. So, this raises the question of whether the limits of acceptability is still too demanding or whether the IHDM4 model and the data used to drive it are not fit-for-purpose in this case, relying on error and model realization averaging to achieve the relative success of the global likelihood measures. That is a decision that might depend on the aim of a particular application of the model but can be considered to be an objective reason for rejection of the model and/or the data being used in this application. Such a rejection is perhaps not unexpected (with the IHDM4 model assumptions that the Richards equation applies in a soil profile of uniform depth and conductivity and idealized initial conditions prior to each storm) and can be a good thing in the learning process of doing better hydrology. The global likelihood measures, however, do not result in such a rejection. Indeed, it is worth noting that the formal Bayes likelihood will *never* reject a model, only produce low likelihoods, that might then be rescaled to appear significant because of the role of the scaling constant C in [1]. As a result in Figure 5, only a small number of models have a significant impact on the posterior parameter distributions

RESPONDING TO MODEL REJECTION

Rejection of all the models tried is a positive result in that it shows that some improvement is required to either the model structure being used or the data that force the model or are used in evaluation. An analysis of the failures can then suggest where improvements might be required (e.g. Choi and Beven, 2006) or when the data are hydrologically inconsistent (Beven *et al*., 2011). Within the limits of acceptability framework, an analysis can also provide information about the most critical observations in inducing model failure (e.g. Blazkova and Beven, 2009). In the research sphere therefore, model rejection should not be the end point of a study but should lead to further development and improvement in knowledge.

Rejection of all the models tried is not, however, a very useful result in a practical application when some decision needs to be made on the basis of risk measures dependent on model simulations. The practicing hydrologist may not then have either time or resources to effect some model revision or re-evaluate the available observations. This has not been an issue in the past when there has been more emphasis on finding the best model (or set of behavioural models) with (or without!) some estimate of uncertainty. But, if there is reason to reject all the models tried, then it raises an ethical issue about how far a model is fit for purpose in the decision making process. We could, of course, relax the criteria of rejection (which will also generally increase the range of uncertainty, given more chance of simulating future surprises), but unless there are good hydrological arguments for doing so, we should surely still be wary of using predictions that may not be fit for purpose.

RECENT DEVELOPMENTS AND ISSUES FOR FUTURE RESEARCH

If epistemic uncertainties are important in hydrological modelling (as we believe that they are), then how can we try to account for such errors in model evaluations? We should not, after all, expect a model to make predictions of better quality than the data that has been used in its calibration or the data with which it will be compared in evaluation, but we should expect that epistemic uncertainties will, as the result of a lack of knowledge, be difficult to quantify. This then suggests making some assessment of the quality of the data *before* running the model, including the elimination of any data that appear to be inconsistent. This includes, for example, storms with apparent runoff coefficients greater than 1 (for whatever epistemic reason). No model would be able to reproduce such an output if it imposes a closed water balance (as most hydrological models do). Analysis of rainfall–runoff data, however, has shown that such cases are not uncommon, even for some quite large rainfall volumes (e.g. the case study of Beven *et al*., 2011; Beven and Smith, 2013).

So, data quality should be checked carefully, but even after such checking, there will be some uncertainties about both input and evaluation data. Assessing the measurement uncertainties associated with both input and evaluation data independently of any simulation model structure should, in principle, be possible. In practice, two difficulties arise, both associated with epistemic issues.

The first is that we may not have the knowledge with which to assess the uncertainties of either inputs or evaluation data. This may be because of the limitations of the available measurement techniques or number of measurements in providing what the model needs or an evaluation variable that is commensurable with a model predicted variable (Beven, 2006, 2009a). Examples are the uncertainties associated with rainfall over a catchment, as they vary from event to event. Interpolation of raingauge observations requires a model (both of factors, such as wind speed, affecting the raingauge catch that are commonly ignored and of interpolation in space for which there are numerous different techniques available, e.g. Shaw *et al*., 2010). Estimation of rainfall intensities from radar reflectivity requires a model. Both of these interpretative models might require different parameters for different events (or even sub-event time steps), which may themselves be subject to epistemic uncertainties, especially when limited data are available to estimate the interpolation characteristics (McMillan *et al*., 2012).

The use of an observed water level to infer discharge in a channel requires a model of the rating curve, which might involve epistemic uncertainty in extrapolating to flood levels beyond the range of the rating curve measurements. There are also commensurability issues about relating point measurements of soil water or water table levels in observation wells to model predicted variables at catchment or discretization element scales. Hydrological modellers appreciate these issues much more now than 20 years ago, but we cannot say that we have really made too much progress in understanding how they influence the real information content of calibration data.

The second difficulty is that even when we could make some sort of assessment of input errors, the impact of those errors on prediction uncertainties depends on processing through a particular model structure and parameter set. The inverse of this problem is seen in some recent studies that try to identify input errors as part of a Bayesian identification methodology (e.g. the BATEA studies of Thyer *et al*., 2009 and the DREAM studies of Vrugt *et al*., 2009b but see also Beven, 2009b, in respect of the latter). These attempt to identify rainfall multipliers for individual events that give good predictions conditional on a chosen model structure. But, there is clearly potential for interaction between these identified rainfall multipliers and any model structural errors. These effects cannot be separated (Beven, 2005, 2006). The result is that the parameters of the model are apparently very well estimated – equifinality has been effectively eliminated but only by transferring uncertainty to the inputs in a way that compensates for model structural error and which cannot be easily extrapolated to prediction events.

The concept of not expecting a model to perform better than the input and evaluation data should, however, hold. Given an independent estimate of input error, we can then use a forward propagation of that error through any given combination of model structure and parameter set for comparison with any evaluation data. The fact that we are not used to estimating such errors independently of a model run should not preclude the development of objective techniques to do so. It is worth noting that these techniques might not be simply statistical. A hydraulic extrapolation of a discharge rating curve that takes account of the changing cross-section and roughness in overbank flow might have greater value than a statistical regression extrapolation. Similarly, simple geostatistical interpolation of rainfall fields (e.g. McMillan and Clark, 2009) might not be informative if the field is not second order stationary and if variograms are non-stationary and poorly estimated from small numbers of gauges. It might still be a step in the right direction, but this is not just a statistical problem.

Consider for the moment a situation within which it is required to test one or more rainfall–runoff model structures. The same steps are required in the analysis as in the original BB92 case.

1. An estimation of the prior likelihoods (as probabilities, possibilities or measures of belief) for model structures and parameters.
2. A method of sampling the model space to find behavioural models and reject non-behavioural models.
3. A method of defining a likelihood weight for each behavioural model that can then be used in predicting the CDF of output variables (including for an explicit error model if used).

Steps 2 and 3 now require some modification, however, because step 2 now no longer requires the assessment of performance of a deterministic run of a model structure and parameter set combination, but the assessment of a model run subject to some input errors and comparison with uncertain evaluation variables (resulting from measurement, interpretation model and commensurability errors as described above). Effectively this requires an assessment of the consistency of two uncertain variables at multiple time steps. It has some similarities with the method of phi-shadowing of Smith (2001), although he looks primarily at the problem of

getting the initial conditions for a forecast to shadow the evolution of the system within observational uncertainties, and not the problem of continuously uncertain forcing data.

And therein lies another dimensionality problem, because whether a model parameter set can be expected to provide good predictions given a sequence of uncertain inputs depends on the particular realization of those inputs. This also arises where the inputs for a model are themselves generated by a model with uncertain parameters. This was the case in the GLUE flood frequency application of Blazkova and Beven (2009). In their study, a small number of models from a sample of 600 000 runs were found that were consistently within specified limits of acceptability for criteria on the basis of uncertain discharges (as expressed in flood frequency and flow duration curves) and snow accumulation quantiles. Checking those models for different realizations of the inputs, however, suggested that there was a low probability of that parameter set being behavioural across all input realizations. Likelihood of a particular parameter set depended heavily on input realization; a parameter set could change from being behavioural to non-behavioural (there is also a lesson there for model applications that consider the inputs to be known perfectly). This also has implications for making Type I and Type II errors in testing models as hypotheses when, in fact, Type III errors are important in the modelling process (refer to the discussion in Beven, 2010, 2012a).

However, even assuming that input uncertainties can be defined before running any of the models (hopefully in an objective way from the evidence available), a full evaluation of the effects will then require many realizations to be run with each model parameter set greatly increasing the computational burden. How important this is will depend on how sensitive are the results to input uncertainty relative to other uncertainties. The range of multipliers identified by application of the Bayesian BATEA and DREAM methods (even allowing for the fact that model structure and parameter uncertainties are being compensated by rainfall multiplier uncertainties) suggests that input uncertainties cannot be considered negligible.

But, this is a problem of epistemic uncertainty. The reason why little research has been performed about the uncertainties in rainfall input fields in hydrological modelling is not that we do not understand that it is a problem but that there is not generally adequate knowledge on which to base a quantification of the uncertainty. There are then three possible responses, all of which are questions for future research. The first is to collect more detailed data, and there are projects currently addressing this at least in rainfall dominated regimes (Anagnostou *et al*., 2004; McMillan *et al*., 2011). The second would be to speculate about the likely uncertainties (at least for catchment average rainfalls) on the basis of experience of detailed studies elsewhere. The third is to treat the input uncertainties implicitly, as is currently the case in nearly all hydrological modelling studies. The third option will often be attractive for reasons of resources and proportionality to the problem under study.

There remains the question of defining an appropriate likelihood measure (although experience with some of the studies that have used the limits of acceptability form of evaluation within GLUE suggests that this does not make a great difference to prediction uncertainties once the set of behavioural models has been selected). Beven (2006) outlines different functional forms that might be used to define likelihood weights for each available residual; more recent studies have shown how residuals can be normalized to be comparable across measurement types and changing limits (e.g. Liu *et al*., 2009; Blazkova and Beven, 2009; and the normalized misfit used in Figure 6 in the previous texts).

This leaves a final question of how to combine the likelihood values for different observations. Should they be combined multiplicatively, as in Bayes equation, or in some other way? This is actually an interesting question because it really depends on what the underlying value of different observations might be in conditioning the model inference and again this may involve differences of philosophy and choice of technique. Beven *et al*. (2008), for example, raise the issue of whether a new period of similar calibration data should be used to provide strong conditioning (as happens in the stretching of the likelihood surface in Bayes) or whether the new period really adds a significant amount of information to the inference (for example, in reducing the possibility of a Type I error of accepting a model that is not ultimately a good simulator of the catchment). A new period that was quite different to a previous calibration period might provide much more information in constraining Type I and Type II errors *provided that* we can be equally confident in the quality of that data.

So, a further interesting research question is the evaluation of the information content in different types and periods of calibration data, *independent of* the model (s) being used. This clearly involves issues of data quality, of the range of behaviours in a period and of the commensurability of observations with expected model predicted variables. Although there are definitions of information criteria available (Akaike, Bayes, Deviance, Young, etc.), they all apply *post hoc* to a model application, dependent on the residual series and resulting parameter covariances and conditional on an assumption that the model is correct. Evaluation of information content prior to a model run should affect the type of

relative likelihood measure used in a model application but requires a different approach to the traditional criteria (Beven and Smith, 2013, for one approach).

WHAT SHOULD DEFINE SUCCESS IN UNCERTAINTY ESTIMATION?

This would appear to be a simple question. In the traditional statistical approach to uncertainty estimation, success would be defined by correctly characterizing the probability of predicting a new observation in a posterior analysis of a conditional validation period (as in a split-record evaluation test in hydrology). Thus, the 95% prediction bounds should contain 95% of observations, and a quantile–quantile plot should show no strong deviations from normality (or some other distributional assumption). Such a test implies, of course, that the statistical model is aleatory and homogeneous between calibration and prediction periods.

It is clear, however, that in real applications, the importance of epistemic errors makes it difficult to meet such criteria. It is rather common in hydrological modelling to find that model performance in prediction is not as good as in calibration. This implies that the residual characteristics are not homogeneous between calibration and validation periods but are non-stationary (at least in the short term, even if they have long-term stationarity when integrated over a sufficiently long sample of epistemic errors, e.g. Montanari and Koutsoyiannis, 2012). It also implies that there may be elements of surprise in conditional validation and future simulation periods, when the epistemic errors might be quite different to those seen in calibration.

As noted earlier, within the GLUE framework most past applications that treat the residual characteristics implicitly do not guarantee that the ensemble predictions will match the probabilities of new observations, so there is no equivalent quantitative measure of success. Simulations by the set of behavioural models can only span the observations if those models will both overpredict and underpredict all observations and experience shows that this is not always the case. Where it is the case, success is often of a similar level to statistical inference (and without the possibility of prediction limits crossing zero that will happen using a simple additive error model with large error variances in statistical inference). A failure to encapsulate future observations can, however, occur for good epistemic reasons so that such failures might contain valuable information about errors in the model or forcing data. Thus, applying GLUE without a statistical error model is more likely to fail in the sense of bracketing new observations but in doing so will concentrate attention on whether the model and data are fit-for-purpose or need to be improved (as in the limits of acceptability evaluation of the application of IHDM4 to the Gwy catchment above).

THE NEXT 20 YEARS

There is no doubt that the original BB92 paper has had the effect of stimulating a great deal of discussion about the sources of uncertainty in hydrological modelling and how best to deal with them (this is surely one reason why it has been cited so often). The debate between GLUE and formal statistical approaches is still on-going (Clark *et al*., 2011, 2012; Beven *et al*., 2012a), with no sign of real resolution because there is no right answer to the problem of epistemic uncertainties. If we could know enough about the nature of the sources of uncertainties, we could devise ways of dealing with them. Without that knowledge, every approach will be an approximation.

Reviewing progress over the last 20 years, we would finish by suggesting some critical issues that need to be addressed by the community in future. We take it as read that the future will see both better model structures and better observational data (at least in research applications) that will reduce the epistemic uncertainty in both process representations and model evaluations, but there remain some critical issues that require further research.

Of these, perhaps the most critical is evaluating the real information content of hydrological data series and the related issue of reducing the epistemic errors in input and output data. This requires as much hydrological reasoning about evaluating data as statistical theory. We see this as one of the advantages of using the GLUE framework, which can focus attention down to the hydrological significance of single events and observations (see the discussion in Beven and Smith, 2013).

A second related issue is the design of model evaluation strategies (and likelihood measures) that allow for the epistemic error generic to hydrological data series and that allow for model rejection when not fit-for-purpose, rather than compensation by an error model under the assumption that sources of uncertainty can be treated as if only aleatory in nature. This is crucial in shaping the likelihood surface in any model application and therefore the potential for improving the efficiency of defining the shape of that surface using advanced sampling strategies. We note again the essential difference between GLUE and formal statistical approaches in this respect as the implicit handling of errors in GLUE and the explicit error model of the statistical approach (although remember that there is no reason why an explicit error model cannot be included in GLUE, including empirical nonparametric distributions of errors; it simply becomes an additional non-hydrological model component).

The advantage of using an explicit error model is that the variance can expand to make it more likely to bracket a future observation in prediction (with a probabilistic interpretation in the ideal case of purely aleatory error). The advantage of the implicit treatment of error in the GLUE approach is that it is clearer when a model fails, either for model structural or data error reasons. This is important for the future of hydrological modelling, because we only really learn by rejecting models or theories (while making sure that we are not making the error of rejecting a model only because of error in the observational data of course!). It is also important for the ethics of hydrological modelling in practical applications. If there is evidence that a model should be rejected, we should be wary of using predictions from such a model in decision making, even if those predictions are associated with an estimate of uncertainty.

ACKNOWLEDGEMENTS

Many people have contributed to GLUE applications over the last 20 years, and we have learned something from every application. Bruno Ambroise, Giuseppe Aronica, Daniela Balin, Kathy Bashford, Paul Bates, Sarka Blazkova, Rich Brazier, Kev Buckley, Wouter Buytaert, Hyung Tae Choi, Luc Feyen, Stuart Franks, Jim Freer, Francesc Gallart, Barry Hankin, Ion Iorgulescu, Johan Birk Jensen, Christophe Joerin, John Juston, Rob Lamb, Dave Leedal, Yangli Liu, Steve Mitchell, Jeff Neal, Mo Xingguo, Trevor Page, Florian Pappenberger, Pep Piñol, Renata Romanowicz, Jan Seibert, Daniel Sempere, Paul Smith, Søren Thorndahl, Raul Vazquez, Ida Westerberg, Philip Younger and Danrong Zhang have all contributed in this way.

This work is a contribution to the CREDIBLE consortium funded by the UK Natural Environment Research Council (Grant NE/J017299/1). Thanks to the two referees whose suggestions helped to improve the final version of this paper.

REFERENCES

Abbott MB, Bathurst JC, Cunge JA, O'Connell PE, Rasmussen J. 1986a. An introduction to the European Hydrological System – Système Hydrologique Européen (SHE): 1. History and philosophy of a physically based, distributed modelling system. *Journal of Hydrology* **87**: 45–59.

Abbott MB, Bathurst JC, Cunge JA, O'Connell PE, Rasmussen J. 1986b. An introduction to the European Hydrological System – Système Hydrologique Européen, (SHE): 2. Structure of a physically based, distributed modelling system. *Journal of Hydrology* **87**: 61–77.

Allchin D. 2004. Error types. *Perspectives on Science* **9**: 38–59.

Anagnostou EN, Anagnostou, MN, Krajewski WF, Kruger A, Miriovsky BJ. 2004. High-resolution rainfall estimation from X-band polarimetric radar measurements. *Journal of Hydrometeorology* **5**(1): 110–128.

Avramidis AN, Wilson JR. 1996. Integrated variance reduction strategies for simulation. *Operations Research* **44**(2): 327–346.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

Ben-Haim Y. 2012. Why risk analysis is difficult, and some thoughts on how to proceed. *Risk Analysis* **32**: 1638–1646. DOI: 10.1111/j.1539-6924.2012.01859.x

Beran M. 1999. Hydrograph prediction – how much skill? *Hydrology and Earth System Sciences Discussions* **3**(2): 305–307.

Beven KJ. 1975. *A deterministic spatially distributed model of catchment hydrology*, unpublished PhD thesis. University of East Anglia, Norwich, UK.

Beven KJ. 1989. Changing ideas in hydrology: the case of physically based models. *Journal of Hydrology* **105**: 157–172.

Beven KJ. 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. *Advanced in Water Resource* **16**: 41–51.

Beven KJ. 2001. Dalton Medal Lecture: how far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences* **5**(1): 1–12.

Beven KJ. 2002a. Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. *Hydrological Processes* **16**(2): 189–206.

Beven KJ. 2002b. Towards a coherent philosophy for environmental modelling. *Proceeding of the Royal Society of London* **458**: 2465–2484.

Beven KJ. 2005. On the concept of model structural error. *Water Science and Technology* **52**(6): 165–175.

Beven KJ. 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* **320**: 18–36.

Beven KJ. 2009a. *Environmental Modelling: An Uncertain Future?* Routledge: London.

Beven KJ. 2009b. Comment on 'Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?' by Jasper A. Vrugt, Cajo J. F. ter Braak, Hoshin V. Gupta and Bruce A. Robinson. *Stochastic Environmental Research and Risk Assessment* **23**: 1059–1060. DOI: 10.1007/s00477-008-0283-x

Beven KJ. 2010. Preferential flows and travel time distributions: defining adequate hypothesis tests for hydrological process models. *Hydrological Processes* **24**: 1537–1547.

Beven KJ. 2011. I believe in climate change but how precautionary do we need to be in planning for the future? *Hydrological Processes* **25**: 1517–1520. DOI: 10.1002/hyp.7939

Beven KJ. 2012a. Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience, Académie de Sciences*, Paris. DOI: 10.1016/j.crte.2012.01.005

Beven KJ. 2012b. *Rainfall–Runoff Models: The Primer*, 2nd edn. Wiley-Blackwell: Chichester.

Beven KJ. 2013. So how much of your error is epistemic? Lessons from Japan and Italy. *Hydrological Processes*, **27**(11): 1677–1680. DOI: 10.1002/hyp.9648.

Beven KJ, Binley AM. 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes* **6**: 279–298.

Beven KJ, Freer J. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems. *Journal of Hydrology* **249**: 11–29.

Beven KJ, Kirkby MJ. 1979. A physically based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin* **24**(1): 43–69.

Beven KJ, Smith PJ. 2013. Concepts of information content and likelihood in parameter calibration for hydrological simulation models ASCE. *Journal Hydrologic Engineering.*, in press.

Beven KJ, Westerberg I. 2011. On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrological Processes* **25**: 1676–1680. DOI: 10.1002/hyp.7963

Beven KJ, Young PC. 2003. Comment on Bayesian recursive parameter estimation for hydrologic models by M Thiemann, M Trosset, H Gupta and S sorooshian. *Water Resources Research* **39**(5): W01116. DOI: 10.1029/2001WR001183

Beven KJ, Young PC. 2013. A guide to good practice in modeling semantics for authors and referees. *Water Resources Research* **49**: DOI: 10.1002/wrcr.20393

Beven KJ, Calver A, Morris EM. 1987. The Institute of Hydrology distributed model. *Institute of Hydrology Report* No.98, Wallingford, UK.

Beven KJ, Smith PJ, Freer J. 2008. So just why would a modeller choose to be incoherent? *Journal of Hydrology* **354**: 15–32.

Beven K, Smith PJ, Wood A. 2011. On the colour and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences* **15**: 3123–3133. DOI: 10.5194/hess-15-3123-2011

Beven KJ, Smith PJ, Westerberg I, Freer JE. 2012a. Comment on Clark et al., pursuing the method of multiple working hypotheses for hydrological modeling, W09301, 2011. *Water Resources Research* **48**: W11801. DOI: 10.1029/2012WR012282

Beven KJ, Buytaert W, Smith LA. 2012b. On virtual observatories and modeled realities (or why discharge must be treated as a virtual variable). *Hydrological Processes*. DOI: 10.1002/hyp.9261

Beven KJ, Leedal DT, Hunter, N, Lamb R. 2012c. Communicating uncertainty in flood risk mapping, in Proceedings, FloodRisk2012.

Binley AM, Beven KJ. 1991. Physically based modelling of catchment hydrology: a likelihood approach to reducing predictive uncertainty. In *Computer Modelling in the Environmental Sciences*, Farmer DG, Rycroft MJ (eds). Clarendon Press: Oxford; 75–88.

Binley AM, Elgy J, Beven KJ. 1989a. A physically based model of heterogeneous hillslopes I. Runoff production. *Water Resources Research* **25**(6): 1219–1226.

Binley AM, Beven KJ, Elgy J. 1989b. A physically based model of heterogeneous hillslopes II. Effective hydraulic conductivities. *Water Resources Research* **25**(6): 1227–1233.

Binley AM, Beven KJ, Calver A, Watts L. 1991. Changing responses in hydrology: assessing the uncertainty in physically based predictions. *Water Resources Research* **27**(6): 1253–1262.

Blasone R-S, Madsen H, Rosbjerg D. 2008a. Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling. *Journal of Hydrology* **353**(1): 18–32.

Blasone RS, Vrugt JA, Madsen H, Rosbjerg D, Robinson BA, Zyvoloski GA. 2008b. Generalized Likelihood Uncertainty Estimation (GLUE) using adaptive Markov chain Monte Carlo sampling. *Advances in Water Resources* **31**(4): 630–648.

Blazkova S, Beven KJ. 2004. Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic. *Journal of Hydrology* **292**: 153–172.

Blazkova S, KJ Beven. 2009. A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research* **45**: W00B16. DOI: 10.1029/2007WR006726

Blöschl G, Sivapalan M, Wagener T, Viglione A, Savenije H. 2013. *Runoff Prediction in Ungauged Basins: Synthesis across Processes*. Places and Scales, Cambridge University Press: Cambridge.

Choi HT, Beven KJ. 2007. Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in distributed rainfall–runoff modelling within GLUE framework. *Journal of Hydrology* **332**(3-4): 316–336.

Clark M, Kavetski, D, Fenicia F. 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, **47**(9), DOI: 10.1029/2010WR009827

Clark M, Kavetski D, Fenicia F. 2012. Reply to K. Beven et al., comment on Clark et al., pursuing the method of multiple working hypotheses for hydrological modeling, W09301, 2011. *Water Resources Research*, **48**(11), DOI: 10.1029/2012WR012547

Clarke RT, Leese MN, Newson AJ. 1973. Analysis of data from Plynlimon raingauge networks, April 1971-March 1973. Institute of Hydrology Report No. 27, Wallingford, UK.

Dean S, Freer JE, Beven KJ, Wade AJ, Butterfield D. 2009. Uncertainty assessment of a process-based integrated catchment model of phosphorus (INCA-P). *Stochastic Environmental Research and Risk Assessment* **23**: 991–1010. DOI: 10.1007/s00477-008-0273-z

Diggle PJ, Gratton RJ. 1984. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society, Series B* **46**: 193–227. 4746, 4763

Duan Q, Sorooshian S, Gupta V. 1992. Effective and efficient global optimisation for conceptual rainfall–runoff models. *Water Resources Research* **28**: 1015–1031.

Faulkner H, Parker D, Green C, Beven K. 2007. Developing a translational discourse to communicate uncertainty in flood risk between science and the practitioner. *Ambio* **16**(7): 692–703.

Feyen L, Vrugt JA, O'Nuallain B, van der Knijff J, De Roo A. 2007. Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the LISFLOOD model. *Journal of Hydrology* **332**: 276–289.

Freer J, Beven KJ, Ambroise B. 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. *Water Resources Research* **32**(7): 2161–2173.

Freer J, McMillan H, McDonnell JJ, Beven KJ. 2004. Constraining Dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology* **291**: 254–277.

Freeze RA. 1975. A stochastic-conceptual analysis of one-dimensional groundwater flow in non-uniform homogeneous media. *Water Resources Research* **11**(5): 725–741.

Gupta HV, Kling H, Yilmaz KK, Martinez GF. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology* **377**(1): 80–91.

Helton JC, Burmaster DE. 1996. Treatment of aleatory and epistemic uncertainty in performance assessments for complex systems. *Reliability Engineering and System Safety* **54**(2–3): 91–258.

Hornberger GM, Spear RC. 1980. Eutrophication in Peel Inlet – I. The problem-defining behavior and a mathematical model for the phosphorus scenario. *Water Research* **14**(1): 29–42.

Hornberger GM, Spear RC. 1981. An approach to the preliminary analysis of environmental systems. *Journal Environmental Management* **12**: 7–18.

Howson C. 2003. *Hume's problem: Induction and the Justification of Belief*. Oxford Uinversity Press: Oxford.

Howson C, Urbach P. 1993. *Scientific Reasoning: The Bayesian Approach*, 2nd edn. Open Court: Chicago.

Hudson JA, Gilman, K. 1993. Long-term variability in the water balance of the Plynlimon catchments. *Journal of Hydrology* **143**: 355–380.

Iorgulescu I, Beven KJ, Musy A. 2005. Data-based modelling of runoff and chemical tracer concentrations in the Haute–Menthue (Switzerland) research catchment. *Hydrological Processes* **19**: 2557–2574.

Iorgulescu I, Beven KJ, Musy A. 2007. Flow, mixing, and displacement in using a data-based hydrochemical model to predict conservative tracer data. *Water Resources Research* **43**: W03401. DOI: 10.1029/2005WR004019

Juston J, Seibert J, Johansson PO. 2009. Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. *Hydrological Processes* **23**(21): 3093–3109.

Kavetski D, Clark MP. 2010. Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. *Water Resources Research* **46**(10), DOI: 10.1029/2009WR008896

Kennedy MC, O'Hagan A. 2001. Bayesian calibration of mathematical models. *Journal of the Royal Statistical Society* **D63**(3): 425–450.

Khu S-T, Werner MGF. 2003. Reduction of Monte-Carlo simulation runs for uncertainty estimation in hydrological modelling. *Hydrology and Earth System Sciences* **7**(5): 680–692.

Klemeš V. 1986. Operational testing of hydrologic simulation models. *Journal of Hydrology* **31**: 13–24.

Klepper O, Scholten H, Van Kamer JD. 1991. Prediction uncertainty in an ecological model of the Oosterschelde Estuary. *Journal of Forecasting* **10**(1-2): 191–209.

Koen BV. 2003. *Discussion of the Method: Conducting the Engineer's Approach to Problem Solving*. OUP: Oxford; 260 pp.

Konikow LF, Bredehoeft JD. 1992. Groundwater models cannot be validated. *Advances in Water Resources* **15**: 75–83.

Krueger T, Quinton JN, Freer J, Macleod CJ, Bilotta GS, Brazier RE, Haygarth PM. 2009. Uncertainties in data and models to describe event dynamics of agricultural sediment and phosphorus transfer. *Journal of Environmental Quality* **38**(3): 1137–1148.

Kuczera G, Kavetski D, Franks S, Thyer M. 2006. Towards a Bayesian total error analysis of conceptual rainfall–runoff models: characterising model error using storm-dependent parameters. *Journal of Hydrology* **331**: 161–177.

Kuczera G, Renard B, Thyer M, Kavetski D. 2010. There are no hydrological monsters, just models and observations with large uncertainties! *Hydrological Sciences Journal* **55**(6): 980–991.

Kumar P. 2011. Typology of hydrologic predictability. *Water Resources Research* **47**: W00H05, DOI: 10.1029/2010WR009769

Laloy E, Vrugt JA. 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. *Water Resources Research* **48**(1), DOI: 10.1029/2011WR010608

Laplace PS. 1774. Mémoire sur la probilitié des causes par les événements. *Mémoires de l'Academie de Science de Paris* **6**: 621–656.

Liu Y, Freer JE, Beven KJ, Matgen P. 2009. Towards a limits of acceptability approach to the calibration of hydrological models: extending observation error. *Journal of Hydrology* **367**: 93–103. DOI: 10.1016/j.jhydrol.2009.01.016

Looms MC, Binley A, Jensen KH, Nielsen L, Hansen TM. 2008. Identifying unsaturated hydraulic parameters using an integrated data fusion approach on cross-borehole geophysical data. *Vadose Zone Journal* **7**: 238–248.

Mantovan P, Todini E. 2006. Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. *Journal of Hydrology* **330**: 368–381.

Marc V, Robinson M. 2007. The long-term water balance (1972–2004) of upland forestry and grassland at Plynlimon, mid-Wales. *Hydrology and Earth System Sciences* **11**(1): 44–60.

Marin J-M, Pudlo P, Robert CP, Ryder R. 2011. Approximate Bayesian computational methods. *Statistics and Computing* **22**(6): 1167–1180. DOI: 10.1007/s11222-011-9288-2

Marjoram P, Molitor J, Plagnol V, Tavaré S. 2003. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 15,324–15,328.

McMillan H, Clark M. 2009. Rainfall–runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme. *Water Resources Research* **45**: W04418. DOI: 10.1029/2008WR007288

McMillan H, Freer J, Pappenberger F, Krueger T, Clark M. 2010. Impacts of uncertain river flow data on rainfall–runoff model calibration and discharge predictions. *Hydrological Processes* **24**(10): 1270–1284.

McMillan H, Jackson B, Clark M, Kavetski D, Woods R. 2011. Rainfall uncertainty in hydrological modelling: an evaluation of multiplicative error models. *Journal of Hydrology* **400**(1): 83–94.

McMillan H, Krueger K, Freer J. 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*. DOI: 10.1002/hyp.9384

Montanari A, Brath A. 2004. A stochastic approach for assessing the uncertainty of rainfall–runoff simulations. *Water Resources Research* **40**: W01106. DOI: 10.1029/2003WR002540.

Montanari A, Koutsoyiannis D. 2012. A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research* **48**: W09555. DOI: 10.1029/2011WR011412

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models: 1. A discussion of principles. *Journal of Hydrology* **10**: 282–290.

Newson AJ. 1976, Some aspects of the rainfall of Plynlimon, mid-Wales. Institute of Hydrology Report No. 34, Wallingford, UK.

Nott DJ, Marshall L, Brown J. 2012. Generalized Likelihood Uncertainty Estimation (GLUE) and approximate Bayesian computation: what's the connection? *Water Resources Research* **48**: W12602. DOI: 10.1029/2011WR011128

Page T, Beven KJ, Freer J, Jenkins A. 2003. Investigating the uncertainty in predicting responses to atmospheric deposition using the Model of Acidification of Groundwater in Catchments (MAGIC) within a Generalised Likelihood Uncertainty Estimation (GLUE) framework. *Water, Air, Soil Pollution* **142**: 71–94.

Page T, Beven KJ, Whyatt D. 2004. Predictive capability in estimating changes in water quality: long-term responses to atmospheric deposition. *Water Soil and Air Pollution* **151**: 215–244.

Page T, Beven KJ, Freer J. 2007. Modelling the chloride signal at the Plynlimon catchments, Wales using a modified dynamic TOPMODEL. *Hydrological Processes* **21**: 292–307.

Pappenberger F, Beven K, Horritt M, Blazkova S. 2005. Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations. *Journal of Hydrology* **302**: 46–69.

Pappenberger F, Frodsham K, Beven KJ, Romanovicz R, Matgen P. 2007. Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrology and Earth System Sciences* **11**(2): 739–752.

Renard B, Kavetski D, Kuczera G, Thyer M, Franks SW. 2010. Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. *Water Resources Research* **46**(5), DOI: 10.1029/2009WR008328

Romanowicz R, Beven KJ, Tawn J. 1994. Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach. In *Statistics for the Environment II. Water Related Issues*, Barnett V, Turkman KF (eds). Wiley: Chichester; 297–317.

Romanowicz R, Beven KJ, Tawn J. 1996. Bayesian calibration of flood inundation models. In *Floodplain Processes*. Anderson MG, Walling DE, Bates PD (eds). Wiley: Chichester; 333–360.

Rose KA, Smith EP, Gardner RH, Brenkert AL, Bartell SM. 1991. Parameter sensitivities, Monte Carlo filtering, and model forecasting under uncertainty. *Journal of Forecasting* **10**(1-2): 117–133.

Rosenblueth E. 1975. Point estimates for probability moments. *Proceedings of the National Academy of Sciences* **72**(10): 3812–3814.

Rougier J. 2013. Quantifying Hazard Losses. In *Risk and uncertainty assessment for natural hazards*, Rougier J, Sparks S and Hill L (eds). Cambridge University Press: Cambridge, UK; 19–39.

Rougier J, Beven KJ. 2013. Model limitations: the sources and implications of epistemic uncertainty. In *Risk and uncertainty assessment for natural hazards*, Rougier J, Sparks S, and Hill L (eds). Cambridge University Press: Cambridge, UK; 40–63.

Sadegh M, Vrugt JA. 2013. Approximate Bayesian computation in hydrologic modeling: equifinality of formal and informal approaches. *Hydrology and Earth System Sciences Discussions* **10**(4739–4797): 2013.

Schaefli B, Gupta HV. 2007. Do Nash values have value? *Hydrological Processes* **21**(15): 2075–2080.

Schoups G, Vrugt JA. 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research* **46**(10): W10531. DOI: 10.1029/2009WR008933

Seibert J, Beven K. 2009. Gauging the ungauged basin: how many discharge measurements are needed? *Hydrology and Earth System Sciences* **13**: 883–892.

Shaw EM, Beven KJ, Chappell NA, Lamb R. 2010. *Hydrology in Practice*, 4th edn. Spon: London.

Smith LA. 2001. Disentangling uncertainty and error: on the predictability of nonlinear systems. Birkhauser.

Smith L, Freeze RA. 1979. Stochastic analysis of steady state groundwater flow in a bounded domain, 2, two-dimensional simulations. *Water Resources Research* **15**(6): 1543–1559.

Smith RE, Hebbert RHN. 1979. A Monte Carlo analysis of the hydrologic effects of spatial variability of infiltration. *Water Resources Research* **15**(2): 419–429.

Smith PJ, Tawn J, Beven KJ. 2008. Informal likelihood measures in model assessment: theoretic development and investigation. *Advances in Water Resources* **31**(2008): 1087–1100.

Spear RC. 1997. Large simulation models: calibration, uniqueness and goodness of fit. *Environmental Modelling & Software* **12**(2): 219–228.

Spear RC, Hornberger GM. 1980. Eutrophication in Peel Inlet – II. Identification of critical uncertainties via generalized sensitivity analysis. *Water Research* **14**(1): 43–49.

Spear RC, Grieb TM, Shang N. 1994. Parameter uncertainty and interaction in complex environmental models. *Water Resources Research* **30**: 3159–3170.

Stedinger JR, Vogel RM, Lee SU, Batchelder R. 2008. Appraisal of the Generalized Likelihood Uncertainty Estimation (GLUE) method. *Water Resources Research* **44**: W00B06. DOI: 10.1029/2008WR006822

Taleb NN. 2010. *The Black Swan*, 2nd edn. Penguin: London.

Tarantola A. 2006. Popper, Bayes and the inverse problem. *Nature Physics* **2**: 492–494.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* **145**(2): 505–518.

Thyer M, Renard B. Kavetski D, Kuczera G, Franks SW, Srikanthan S. 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling a case study using Bayesian total error analysis. *Water Resources Research* **45**. DOI: 10.1029/2008WR006825

Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6**: 187–202.

Van Straten G, Keesman KJ. 1991. Uncertainty propagation and speculation in projective forecasts of environmental change. *Journal Forecasting* **10**: 163–190.

Von Bertalanffy L. 1968. *General Systems Theory*. George Braziller: New York.

Vrugt JA ter Braak, CJF, Clark MP, Hyman JM, Robinson BA. 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research* **44**(12). DOI: 10.1029/2007WR006720

Vrugt, JA, ter Braak CJF, Diks CGH, Higdon D, Robinson BA, Hyman JM. 2009a. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation* **10**(3): 273–290.

Vrugt JA, Ter Braak, CJ, Gupta HV, Robinson BA. 2009b. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment* **23**(7): 1011–1026.

Warwick JJ, Cale WG. 1988. Estimating model reliability using data with uncertainty. *Ecological Modelling* **41**: 169–181.

Westerberg I, Guerrero J-L, Seibert J, Beven KJ, Halldin S. 2011a. Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes* **25**: 603–613. DOI: 10.1002/hyp.7848

Westerberg IK, Guerrero J-L, Younger PM, Beven KJ, Seibert J, Halldin S, Freer JE, Xu, C-Y. 2011b. Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences* **15**: 2205–2227. DOI: 10.5194/hess-15-2205-2011

Xiong L, O'Connor KM. 2008. An empirical method to improve the prediction limits of the GLUE methodology in rainfall–runoff modeling. *Journal of Hydrology* **349**(1–2): 115–24.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

