Leonardo Lecture
Facets of Uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication

Keith Beven
Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK and
Department of Earth Sciences, Uppsala University, Uppsala, SE 75263, Sweden
k.beven@lancaster.ac.uk

Abstract
This paper presents a discussion of some of the issues associated with the multiple sources of uncertainty and non-stationarity in the analysis and modelling of hydrological systems. Different forms of aleatory, epistemic, sematic and ontological uncertainty are defined. The potential for epistemic uncertainties to induce disinformation in calibration data and arbitrary non-stationarities in model error characteristics, and surprises in predicting the future, are discussed in the context of other forms of non-stationary. It is suggested that a condition tree is used to be explicit about the assumptions that underlie any assessment of uncertainty. This also provides an audit trail for providing evidence to decision makers.

Introduction
I first started carrying out Monte Carlo experiments with hydrological models in 1980, while working at the University of Virginia. This was not a new approach at that time, but the computing facilities available (a CDC600 "mainframe" computer at UVa) made it feasible for the types of hydrological model being used then. Adopting a Monte Carlo approach was a response to a personal "gut feeling" that traditional statistical approaches (at that time an analysis of uncertainty around the maximum likelihood model) were not sufficient to deal with the complex sources of uncertainty in the hydrological modelling process. Over time, we have learned much more about how to discuss facets of uncertainty in terms of aleatory, epistemic, ontological, linguistic, and other types of uncertainty (for one set of definitions see Table 1). Our perceptual model of uncertainty is now much more sophisticated but I will argue that this has not resulted in analogous progress in uncertainty quantification and, more particularly, uncertainty reduction. As one referee on this paper suggested, it can be argued that the classification of uncertainties is not really necessary: there are only epistemic uncertainties (arising from lack of knowledge) because we simply do not know enough about hydrological systems and their inputs and outputs. It is then a matter of choice as to how to treat those uncertainties, including formal probabilistic and statistical frameworks.

What is clear is that such epistemic uncertainties will limit the inferences that can be made about hydrological systems. In particular, we are often dependent on the uncertainties associated with past observations (see, for example, Figure 1) and have not really done a great deal about reducing hydrological data uncertainties into the past. Some observational uncertainties can certainly be treated as random variability or aleatory, but can also be subject to arbitrary

uncertainties. Here, I use the word arbitrary to distinguish epistemic uncertainties that do not have simple structure or stationary statistical characteristics on the time scales used for model calibration and evaluation. This time scale qualification is important in this context since the only information we will have about the impact of different sources of uncertainties on model outputs will be contained in the sequences of model residuals within some limited period of time. It is easy to show that stochastic models based on purely aleatory variability can exhibit apparent short period irregularity or non-stationarity (see for example Kousoyiannis, 2010; Montanari and Koutsoyiannis, 2012). However, there is then the question of how to identify the characteristics of long period variability from shorter periods of model residuals that might contain the type of arbitrary characteristics defined above. It has been shown that some arbitrary uncertainties of this type might be *disinformative* to the model calibration process (Beven et al., 2011; Beven and Westerberg, 2011; Beven and Smith, 2014; Kauffeldt et al., 2013; Figure 1), even if they might be informative in other senses (such as in identifying inconsistences in hydrological observations, Beven and Smith, 2014).

A disinformative event in this context is one for which the observational data are inconsistent with the fundamental principles (or *capacities* in the sense of Cartwright, 1999) that might be applied to hydrological systems and models. Most hydrological simulation models (as opposed to forecasting models, see Young and Beven, 2013) impose a principle of mass balance. We expect catchment systems to also satisfy mass balance (and energy balance, and momentum balance, see Reggiani et al., 1999). The observational data, however, might not. Figure 1 is a good example of this, with far more output as discharge from the catchment than the recorded inputs for that event. While there are some circumstances, such as a rain-on-snow event where this could be realistic scenario, clearly no model that is constrained by mass balance would be able to reproduce such an event, suggesting that the residuals would induce bias in any model inference. It also suggests that we should take a much closer look at the data to be used in model calibration and evaluation *before* running a model (including the neglect of potential snowmelt inputs).

The implication of allowing that some model residuals might be affected by this type of arbitrary epistemic uncertainty is that commonly used probabilistic or statistical approaches to uncertainty estimation do not take enough account of the epistemic nature of uncertainty in the modelling process. It is not just a matter of finding an appropriate statistical distribution or, alternatively, some non-parametric probabilistic structure for the model residuals (e.g. Schoups and Vrugt, 2010; Sikorska et al., 2014), especially when the sample of possible arbitrary uncertainties (or surprises) might be small. It will be suggested in what follows that we need to be more pro-active about methods for uncertainty identification and reduction. This might help to resolve some of the differences between current approaches.

Defining Types of Uncertainty (and why the differences are important)

Past analysis in a variety of modelling domains in the environmental sciences has distinguished a variety of types of uncertainties and errors, including aleatory uncertainty, epistemic uncertainty, semantic or linguistic uncertainty and ontological uncertainty (e.g. Beven and Binley, 1992; McBratney, 1992; Regan et al., 2002; Ascough et al., 2008; Beven, 2009; Beven et al., 2014; Raadgever et al., 2011; Beven and Young, 2013). Table 1 lists one such classification relevant to the application of hydrological models. In particular, the definition of aleatory uncertainty is constrained to the case of of stationary statistical variation (noting that this might involve a structural statistical model but with stationary parameters), for which the full power of statistical theory and inference is appropriate. Epistemic uncertainties, on the other hand, have been broken down into those associated with model forcing data and observations of system response, and those associated with the representation of the system dynamics. As in Figure 1, the observational data might sometimes be hydrologically inconsistent, and might lead to disinformation being fed into the model inference process (Beven et al., 2011; Beven and Smith, 2014). Any of these might be sources of the rather arbitrary nature of errors in the forcing data and resulting model residual variability noted above.

Many aspects of the modelling process involve multiple sources of uncertainty, and without making very strong assumptions about the nature of these different sources it is not possible to separate the effects of the different uncertainties (Beven, 2005). Attempts to separate the error associated with rainfall inputs to a catchment, for example, result in some large changes to event inputs and a strong interaction with model structural error (e.g. Vrugt et al., 2008; Kuczera et al., 2010; Renard et al., 2010). The very fact that there are epistemic uncertainties arising from lack of knowledge about how to represent the response, about the forcing data, and about the observed responses, reinforces this problem. If we knew what type of assumptions to make then the errors would no longer be epistemic in nature.

Defining a method of uncertainty estimation (and why there is so much controversy about how to do so)

Uncertainty estimation has been the subject of considerable debate in the hydrological literature. There are those who consider that formal statistics is the only way to have an objective estimate of uncertainty in terms of probabilities (e.g. Mantovan and Todini, 2006; Stedinger et al., 2008) or that the only way to deal with the unpredictable is as probabilistic variation (Montanari, 2007; Montanari and Koutsoyiannis, 2012). There are those who have argued that treating all uncertainties as aleatory random variables will lead to overconfidence in model identification, so that more informal likelihood measures or limits of acceptability might be justified (e.g. within the GLUE framework of Beven, 2006, 2012; Beven and Binley, 1992, 2013; Freer et al., 2004; Smith et al., 2008; and within Approximate Bayesian Computation by Nott et al. 2012; and Sadegh and Vrugt, 2013, 2014). There are those who recognise the complex structure of hydrological model errors but who use transformations of different types to fit within a formal statistical framework (e.g. Montanari and

147 Brath, 2005). Some of these opinions have been explored in a number of
148 commentaries and opinion pieces (Beven, 2006a,b, 2008, 2012; Hamilton, 2007;
149 Montanari, 2007; Hall et al., 2008; Sivakumar, 2008; Todini and Mantovan, 2008)
150 as well as in more technical papers.
151
152 There is, of course, no right answer – precisely because there are multiple
153 sources of epistemic uncertainty, including model structural uncertainty, that are
154 impossible to separate. There are also different frameworks for assessing
155 uncertainties and different ways of formulating likleihoods. If we had knowledge
156 of the true nature of the sources of uncertainty then they would not be epistemic
157 and we might then be more confident about using formal statistical theory to
158 deal with all the sources of unpredictability. As noted earlier, some epistemic
159 uncertainties should be reducible by further experimentation or observation, so
160 that there is an expectation that we might move towards more aleatory residual
161 error in the future. In hydrology, however, this still seems a long way off,
162 particularly with respect to the hydrological properties of the subsurface. And
163 if, of course, there is no right answer then this leaves plenty of scope for different
164 philosophical and technical approaches for uncertainty estimation – or, put
165 another way, how to define an uncertainty estimation methodology involves
166 ontological uncertainties (Table 1). In this situation there is a lot of uncertainty
167 about uncertainty estimation, and this is likely to be the case for the foreseeable
168 future. This has the consequence that communication of the *meaning* of
169 different estimates of uncertainty can be difficult. This should not, however, be
170 an excuse for not being quite clear about the assumptions that are made in
171 producing a particular uncertainty estimate (Faulkner et al., 2007; Beven and
172 Alcock, 2012; see later).
173
174 **Defining non-stationarity (in catchments and model residuals)**
175
176 Many people think that the only important distinction in the modelling process is
177 between variables that are predictable and uncertainties that are not. Model
178 residuals might have components of both: some identifiable predictable
179 structure as well as some unpredictable variability. The structure indicates
180 some aspect of the system dynamics (or boundary condition and evaluation
181 data) that is not being captured by the model. It is often represented as a
182 deterministic function. In the very simplest case, a stationary mean bias; in more
183 complex cases the function might indicate some structured variability in time or
184 space, such as a trend or seasonal component. The unpredictable component, on
185 the other hand, is usually treated as if the variability is purely aleatory on the
186 basis that if something is not predictable then it should be considered within a
187 probabilistic framework (e.g. Montanari, 2007) albeit that, as already noted, the
188 nature of that variability might have some long time scale properties
189 (Koutsoyiannis, 2010; Montanari and Koutsoyiannis, 2012).
190
191 This is important because it has implications for evaluating models as
192 hypotheses in the face of epistemic errors (or long time scale aleatory errors).
193 Hypothesis testing has traditionally been the realm of statistical inference and
194 probability, including the recent application of Bayesian statistical theory to
195 hydrological modelling (e.g. Clark et al., 2011). Purely empirically, probability

and statistics can, of course, describe anything from observations to model residuals regardless of the actual sources of uncertainty as an expression of our reasonable expectations (Cox, 1946). However, for any particular set of data, the resulting probabilities are conditional on the sample being considered. This is one reason why we try to abstract the empirical to a functional distributional form or the type of empirical non-parametric distributions used by Sikorska et al. (2014) or Beven and Smith (2014).

For simple cases where the empirical sample is random and stationary in its characteristics (after taking account of any well defined structure) then there is a body of theory to suggest what we should expect in terms of variability in statistical characteristics as a function of sample size. There is also then a formal relationship between the statistical characteristics and a likelihood function that can be used in model evaluation. The simplest case is when the statistics of the sample have zero mean bias, constant variance, are independent and can be summarized as a Gaussian distribution. More complex likelihood functions could take account of bias, heteroscedasticity, autocorrelation and other assumptions about the distribution. Even these more complex cases, however, are what I have called ideal cases in the past (e.g. Beven, 2002; 2006a). Fundamentally, they assume all variability in model residuals is aleatory in nature.

But real problems are not ideal in this sense, as illustrated above they are subject to arbitrary epistemic errors. It is then debatable as to whether it is appropriate to treat the errors *as if* they are aleatory. The reason is that the effective information content of any observations (or model residuals) will be reduced by epistemic uncertainties relative to the ideal case. Why is this? It is because the stationary parameter assumption of the aleatory component gives the possibility of future surprise a very low likelihood. Yet evaluating the performance of hydrological models in real applications often reveals surprises that are clearly not aleatory in this way, including occasional surprises of gross under or over predictions. This makes it difficult to define a formal statistical model of the residual structure and consequently, if the methods of estimating likelihoods in formal statistics are not valid, makes hypothesis testing of models more difficult (e.g. Beven, 2010; Beven et al., 2012).

Consider the situation where the estimates of rainfall over a catchment might be of variable quality during a series of events in a model calibration period. The error in the estimates is not aleatory or distributional in nature because the distribution of events is not expected to be stationary (except possibly over very long periods of time but that is not really of interest for the period of calibration data that might be available). This is the context in which we can describe the variability as rather arbitrary i.e. we do not really know whether the rainfall uncertainties conform to any statistical distribution or if the errors in a calibration period are a good guide to the errors in the prediction period that we are actually interested in. The same could be true, of course, for aleatory errors with long-term properties (see examples in Koutsoyiannis, 2010; Montanari and Koutsoyiannis, 2012; Koutsoyiannis and Montanari, 2015). The underlying stochastic process might then be stationary but it might be difficult to identify

the properties of that process from a short-term sample with apparently non-stationary statistics. These are then both forms of epistemic uncertainty. In both cases we lack knowledge about the arbitrary nature of events or the stochastic process. We could in principle, of course, constrain that uncertainty by better observational methods, or longer data series - though that is not very useful when we only have access to calibration data collected in the past, even if we might hope to have improved data into the future.

An interesting example in this respect is the post-audit analyses of a number of groundwater modelling studies presented in Konikow and Bredehoeft (1992) and Anderson and Woessner (1992). Model predictions of future aquifer behavior were compared with what actually happened as the future evolved. In most studies the models failed to predict the future that actually happened. In some cases this was because, with hindsight, the original model turned out to be rather poor; in other cases it was because the future boundary conditions for the simulations had not been well predicted. In hindcasting with the correct boundary conditions the predictions were much better. Hindcasting is not all that useful, however. Where modelling is used to inform decision making (as in these groundwater cases) it is predictions of the future that are required. In these studies therefore, error characteristics were not stationary and the future turned out to hold epistemic surprises (either that the calibrated model was poor, or that the changes in boundary conditions were not those expected).

These examples involve a number of forms of non-stationarity. These are summarized in Table 2. In Class 1 we place the classical definition of non-stationarity discussed by Koutsoyiannis and Montanari (2015) in the context of stochastic process theory. They, in fact, consider that this is the *only* legitimate use of the word non-stationarity in being consistent with its technical definition. In doing so, they are assuming that once any deterministic structure has been taken into account, all forms of epistemic error can be represented by a stationary stochastic model. The parameters of that model will, under the ergodic hypothesis, converge to the true values of the stochastic process as more and more observations are collected. That might, in the case of a complex stochastic process (or even some simple fractal processes) take a very large sample, but that does not negate the principle. Indeed, for a deterministic dynamical system, a stochastic representation will have stationary properties only if it is ergodic. If non-stationarity is assumed, then the system will not have ergodic properties and, Koutsoyiannis and Montanari (2015) suggest, inference will be impossible. This view means either we are back to treating all epistemic uncertainty as aleatory and stationary, once any deterministic structure has been removed, or we are simply left with unpredictability as a result of lack of knowledge.

This view has the backing of formal stochastic theory but I think there are two issues with it. The first is the difference between what might hold in the ergodic case and the limit sample of behaviours we have in calibrating models in practical applications. The example of a stationary stochastic process giving rise to apparently non-stationary behavior and statistics used to illustrate Koutsoyiannis and Montanari (2015) illustrates this nicely. If we have access

only to a limited part of the full record, we might see periods of different statistical characteristics, or periods that include jumps. Real hydrological data might certainly be of this form, but the identification of the true stochastic process would not be possible without very long series (this is true for any fractal type behavior). The fact that we know that the changing statistics are produced by a stationary process in such a hypothetical example, does not negate the fact that the statistics are changing and we should be wary of using an oversimplified error model (see discussion of Figure 2 below).

Secondly, the dynamics of a nonlinear catchment model will introduce changes in the statistical properties of residuals both in the way it processes errors in the inputs and as a result of model structural error that cannot be compensated by a simple deterministic non-stationarity. From a purely hydrological point of view we expect that model residuals should have rather different characteristics on the rising limb to those around the peak to those on the falling limb in terms of bias, changing variance, and changing autocorrelation. The problem will be greater for the type of arbitrary event to event epistemic input (or model structure) error discussed above. The error in that event will also have an effect on setting up the antecedent conditions for the following event, and in some catchments, for some time into the future. The statistics of the error will be changing. Again therefore we should be wary of using an oversimplified error model. It is possible that again there may be a complex stochastic model that would describe all the potential changes in error statistics, but it is doubtful if it would be identifiable given the small sample of potential errors in a calibration period. It is notable that, even given a long period of calibration data, Sikorska et al. (2014) did not attempt to identify an underlying stochastic model of the residuals, but instead used a non-parametric probabilistic approach (in the reasonable expectation tradition of Coxian probability, Cox, 1946), to represent the changing variability of the modelling uncertainties under different circumstances (see also Beven and Smith, 2014). There is a difficulty with any non-parametric method, however, of how to deal with potential uncertainties in the future that are outside the range of those seen in the past.

Why is it important to make these distinctions? It is because it has an impact on what we should expect in testing a model as a hypothesis of how a catchment functions, and in particular whether it should be considered to be fit for purpose. For example, catchments change over time (Non-stationarity Class 2) but models are often fitted with parameters that are assumed constant in time (and often space). Why is this considered acceptable practice? Perhaps, because there is an implicit expectation that this type of non-stationarity will be dominated by uncertainty in the boundary conditions used to drive a model (including the potential for Non-stationarity Class 3). There may, of course, be some clues as to whether these non-stationarities are important if there is some identifiable structure in the model residuals that could be included as a deterministic component in Non-stationarity Class 1. But we might only see the net effect of all these non-stationarities in the changing properties of the unpredictable errors (Non-stationary Class 4). But these are rarely investigated. In practical applications, statistical model inference is normally carried out *as if* all sources of error were aleatory with simple stationary properties. This assumption allows

the full power of statistical inference to be applied to model calibration but would seem to be an unrealistic assumption for hydrological and other environmental models.

Defining likelihood (and the implications for information content and hypothesis testing).

The advantage of taking a formal statistical approach to model calibration is that there is a formal link between the structure of a set of model residuals and the appropriate likelihood function. If, and only if, the assumptions about the structure of the errors are valid, then there is an additional advantage that there is a theoretical estimate of the probability of predicting a new observation. These advantages are undermined by the non-stationarities that arise from epistemic error that will generally reduce the information content (or introduce more disinformation) in the inference process than would be the case if all errors were simply aleatory with stationary parameters. So treating all sources of error as if aleatory will result in over-conditioning (and less protection against surprise in prediction). There is evidence for this in the very tight posterior parameter distributions that often arise in Bayesian calibrations of rainfall – runoff models. The likelihood surface is made very peaky such that models with very similar error variance can have tens or even hundreds of orders of magnitude difference in likelihood (Figure 2). That really does not seem realistic to me, and did not when I first started evaluating likelihoods of multiple runs in the 1980s. The origins of the GLUE methodology lie there.

So one way ahead here might be to find more realistic likelihood functions that reflect the reduced information content for these non-ideal cases and are robust to epistemic error. The question then is how to properly reflect the real information in a set of data when the variations are clearly not aleatory and when the summary statistics might be significantly period dependent. Again, whether the long-term properties are stationary or not is not really relevant, we want to protect against surprise in prediction (as far as is possible for an epistemic problem). In the rainfall-runoff modelling case it has been suggested that the use of summary statistics for model evaluation, such as the flow duration curve, might be more robust to error in this sense (e.g. Westerberg et al., 2011b; Vrugt and Sadegh, 2013).

Beven et al. (2011) and Beven and Smith (2014) show how, for the relatively flashy South Tyne catchment in northern England (322 km^2), it is possible to differentiate obviously disinformative events from informative events in model calibration within the GLUE methodology. They take an event-based approach to model evaluation that tries to reflect the relative information content expected for informative and disinformative events. They suggest that factors that will increase the relative information content of an event include: the relative accuracy of estimation of the inputs driving the model; the relative accuracy of observations with which model outputs will be compared (including commensurability issues); and the unusualness of an event (extremes, rarity of

initial conditions,….). Factors that will decrease the relative information content of an event include: repetition (multiple examples of similar conditions); inconsistency of the input and output data; the relative uncertainty of observations (e.g. highly uncertain overbank flood discharges would reduce information content of an extreme event, discharges for catchments with ill-defined rating curves might be less informative than in catchments with well defined curves); and also a preceding disinformative / less informative event over the dynamic response time scale of the catchment.

The approach depends on classifying events prior to running the model into different classes based on rainfall volume and antecedent conditions. Outlier events can be identified and examined to see if they are disinformative in terms of their runoff coefficients or other characteristics. Limits of acceptability are established for model performance in each class of informative events and a likelihood measure is based on average model performance in each class. The information content for informative events following disinformative events is weighted less highly.

Models that do not meet the limits of acceptability are rejected (given zero likelihood) in the GLUE methodology and do not therefore contribute to the set of models to be used in prediction. This is one way of testing models as hypotheses. Epistemic error also plays a role here in that we would not want to make false negative (Type II) errors in rejecting a model that might be useful in prediction because it has been forced with poor input data. This is more serious than a false positive error in that if a poor model is not initially rejected we can hope that future evaluations would reveal its limitations. Statistical inference deals with this problem by never giving a zero likelihood, only very very small likelihoods to models that do not perform well (as seen in the orders of magnitude change in Figure 2). This also means, however, that no model is ever rejected and hypothesis testing has to depend on some other subjective criterion, such as some informal limits on the Bayes ratios for competing models. One implication for this is that if no model is rejected, there is no guarantee that the best model found is fit for purpose. This must also be assessed separately.

For the South Tyne catchment it turns out that using a standard data set, as collected by the Environment Agency, there were a large number of disinformative events as distinguished by unrealistically high or low runoff coefficients. Excluding these events from the model calibration results in different posterior distributions of the model parameters (see Figure 3). It also allows the characteristics of informative and disinformative events to be considered separately.

When it comes to prediction, however, we do not know *a priori* whether the next event will be informative or disinformative. This can only be evaluated post-hoc, once the future has evolved (in model testing, of course, the "future" considered is some "validation" data set). This may involve non-stationarities of error characteristics that have not been seen in the calibration period. Beven and Smith (2014) allowed for this by evaluating the error characteristics for informative and disinformative events separately and treating each new event as

441 if it might be either informative or disinformative (Figure 4). It was shown to
442 help in spanning the observations for events later shown to be disinformative,
443 but clearly cannot deal with every surprise that might occur in prediction,
444 particularly when the system itself is non-stationary.
445
446 **Defining model rejection in hypothesis testing (and why uncertainty**
447 **estimation is not the end point of a study)**
448
449 In the case of the modelling study of the South Tyne catchment, some models
450 were found that satisfied the limits of acceptability. This is not always the case;
451 in other studies no models have satisfied all the criteria of acceptability imposed
452 (see, for example, the attempts at "blind validation" of the SHE model by Parkin
453 et al. 1996, and Bathurst et al., 2004; and the studies of Pappenberger et al.,
454 2007; Page et al., 2007; Choi and Beven, 2009; Dean et al., 2009; and Mitchell et
455 al., 2011, within the GLUE framework using a variety of different models).
456
457 In terms of the science this is, of course, a good thing in that if all the models are
458 rejected then improvements must be made to either the data or the model
459 structures and parameter sets within those structures being used. That is how
460 real progress is made. But the possibility of epistemic errors in the data used to
461 force a model might make it difficult to make an assessment of how constrained
462 any limits of acceptability should be. We know that all models are
463 approximations and so such limits should be set to reflect the expectation of how
464 well a model should be able to perform. This is a balance. We should not expect
465 a model to predict to a greater accuracy than the assessed errors in the input and
466 evaluation data. If it does we might suspect that it has been over-fitted to
467 accommodate some of the particular realisation of error in the calibration data.
468
469 But we also do not want to make that Type II false negative error of rejecting a
470 model that would be useful in prediction, just because of epistemic errors and
471 disinformation in the forcing or evaluation data. This suggests that if we do
472 reject all the models tried as not fit for purpose we should look first at the data
473 where the model is failing and assess the potential for error in that data,
474 especially if the failures are consistent across a large number of models. In
475 rainfall-runoff modelling this is rarely done, but hydrological modellers are
476 beginning to become more aware of the issues (e.g. Krueger et al., 2009;
477 McMillan et al., 2010, 2012; Westerberg et al., 2011a; Kauffeldt et al., 2013). We
478 also have to be careful that we have searched the model space adequately to
479 ensure that no models have been missed. This can be difficult with high
480 numbers of parameters, when the areas of acceptable models in the model space
481 might be quite local. Iorgulescu et al. (2005) for example made 2 billion runs of a
482 model in a 17 parameter space of which 216 were found to satisfy (rather
483 constrained) limits of acceptability. Blazkova and Beven (2009) made 600000
484 runs of a continuous simulation flood frequency model and found that only 37
485 satisfied all the limits of acceptability. They also demonstrated that whether this
486 was the case depended on the stochastic realisation of the inputs used.
487 Improved efficiency of sampling within this type of rejectionist strategy might
488 then be valuable (e.g. the $DREAM_{ABC}$ code of Sadegh and Vrugt, 2014).
489

But where all the models tried consistently fail, and we do not have any reason for suggesting that the failure is due to disinformative data, then it suggests that a better model is needed. This might lead to new hypotheses about how the system is functioning, or new ways of representing some processes (see also Gupta and Nearing, 2014). Model rejection is not a failure, it is an opportunity to improve either the model or data or both. Finding a better model will not provide total protection against future epistemic surprises but would, we hope, be a step in the right direction. How big a step is possible, however, will also depend on reducing uncertainty in the forcing and evaluation data.

Communicating uncertainty to users of model predictions

There are two main reasons for incorporating uncertainty estimation into a study. One is for scientific purposes, to improve understanding of the problem, and carry out hypothesis testing more rigorously. The second is because taking account of the uncertainty in model predictions might make a difference to a decision that is made in a practical application, for example, whether the planning process can take account of uncertainty in the predicted extent of flooding for the statutory design return period. For this second purpose it is necessary to communicate the *meaning* of the model predictions, and their associated uncertainties, to decision makers (e.g. Faulkner et al., 2007).

But, as we have seen, there can be no right answer to the estimation of uncertainty. Every estimate is conditional on the assumptions that are made and in most applications there are many assumptions that must be made (see, for example, Beven et al., 2014). In this case it might be useful to the communication process if the users, or particular groups of users, are introduced to the nature of those assumptions. In fact, it will generally facilitate the communication process if the users can be involved in making decisions about the relevant assumptions whenever possible. The collection of assumptions that underlie any particular application can be considered to be a form of "Condition Tree" (Beven and Alcock, 2012; Beven et al., 2014). At each level of the condition tree the assumptions must be made explicit, forming a form of audit trail for the analysis. It has even been suggested[1] that every uncertainty assessment should be labelled with the names of those who produced it (and, by extension, perhaps those who agreed the assumptions on which it is based).

Can we talk of confidence rather than uncertainty in model simulations?

Decisions about hydrological systems are made under uncertainty, and often severe uncertainty, all the time. Decision and policy makers are, however, far more interested in *evidence* than uncertainty. Evidence-based framing has become the norm in many areas of environmental policy (e.g. Boyd, 2013). In the UK, the Government has considered standards for evidence (Intellectual Property Office, 2011) and the Environment Agency has an Evidence Directorate and produces documents summarising the evidence that underpins its corporate strategy. Clearly such an Agency wants to have confidence in the evidence used

[1] For example by Jonty Rougier at Bristol University

in such policy framing. Confidence should be inversely related to error and uncertainty, but is often assessed without reference to quantifying uncertainty in either data or model results.

An example case study is the benchmarking exercise carried out to test 2D flood routing models (Environment Agency, 2013). Nineteen models were tested on 12 different test cases, ranging from dam break to fluvial and urban flooding. All the test cases were hypothetical with specified roughness parameters, even if in some of the cases the geometry was based on real areas. Some had some observations available from laboratory test cases. Thus, confidence in this case represents agreement amongst models. It was shown that not all models were appropriate for all test cases, particularly those involving supercritical flow, and that some models that used simplified forms of the St. Venant equations while faster to run had more limited applicability. Differences between models depended on model implementation and numerics, so that acceptability of a model in terms of agreement with other models was essentially a subjective judgment.

There is an implicit assumption in assessing confidence in this way that in real applications to less than ideal datasets, the models that agree can be calibrated to give satisfactory simulations for mapping and planning purposes. While the report did recommend that future comparisons should also aim to assess the value of models in assessing uncertainty in the predictions, the impacts of epistemic uncertainty in defining the input, roughness parameters, and details of the geometry of the flow domain would seem to be more important than the differences between models in which we have confidence after such testing (see Beven et al., 2014). In real applications confidence can only be assessed by comparison with observed data, while allowing for uncertainties in inputs. Even then, there is evidence that effective values of roughness parameters might change with the magnitude of an event, so that confidence in calibration might not carry over to more extreme events (Romanowicz and Beven, 2003). Yet, for planning purposes, the Environment Agency is interested in mapping the extent of floods with annual exceedance probabilities (AEP) of 0.01 and 0.001. It is, of course, rather rare to have observations for floods within this range of AEP, more often we need to extrapolate to such levels.

It is possible to assess the uncertainty associated with such predictions and to visualise that uncertainty either as probability maps (e.g. Leedal et al., 2010; Neal et al., 2012; Beven et al., 2014) or as different line styles depending on the uncertainty in flood extent in different areas (Wicks et al., 2008). In some areas, where the flood fills the valley floor, the uncertainty in flood extent might be small, but the uncertainty in water depth, with its implications for damage calculations, might be important. In other, low slope, areas the uncertainty in extent might be significant. The advantage of doing both estimates is that confidence can be given a scale, even if as in the Intergovernmental Panel on Climate Change (IPCC) that scale is expressed in words rather than probability. In fact, the IPCC distinguishes a scale of confidence (from “very low” to “very high”) from a scale of likelihood (from “exceptionally unlikely” to “virtually certain” based on a probability scale) (IPCC, 2010). Confidence indicates how

convergent the estimates of past and future change are at the current time; likelihood the degree of belief in particular future outcomes. Thus the summary of the outcomes from IPCC5 states that "ocean warming dominates the increase in energy stored in the climate system, accounting for more than 90% of the energy accumulated between 1971 and 2010 (*high confidence*). It is *virtually certain* that the upper ocean (0–700 m) warmed from 1971 to 2010, and it *likely* warmed between the 1870s and 1971. It is *very likely* that the Arctic sea ice cover will continue to shrink and thin and that Northern Hemisphere spring snow cover will decrease during the 21st century as global mean surface temperature rises." (IPCC, 2013).

Now the IPCC will not assign any probability estimates to any of the model runs that contribute to their conclusions. They are described as projections, subject to both model limitations and conditional on scenarios of future greenhouse gas emissions. The future scenarios, and hence any probability statements, are necessarily incomplete. This has not, however, stopped the presentation of future projections in probabilistic terms in other contexts, such as those derived from an ensemble of regional model runs in the UK Climate Projections (UKCP09, see http://ukclimateprojections.defra.gov.uk). The outcomes from UKCP09 are being used to assess impacts on UK hydrology (e.g. Bell et al., 2012; Kay and Jones, 2012; Cloke et al., 2010) but there is sufficient epistemic uncertainty associated with both the input scenarios and the climate model implementations to be concerned about expressions of confidence or likelihood in these cases, when the probabilities may be incomplete and we should be aware of the potential for the future to surprise (Beven, 2011; Wilby and Dessai, 2010). Incomplete probabilities are inconsistent with a risk-based decision theoretic approach based on the exceedance probabilities of risk, although it might be possible to assess a range of exceedance curves under different assumptions about future scenarios (Rougier and Beven, 2013).

We are often in this situation. Hence the need to agree assumptions and methodologies with potential users of model outcomes as discussed in the last section. Consequently any expressions of confidence or likelihood are conditional on the assumptions, a conditionality that depends not only on what has been included, but also what might have been left out of an analysis. There will of course be epistemic uncertainties that are "unknown unknowns". Those we do not have to worry about until, for whatever reason, they are recognized as issues and become "known unknowns". More important are factors that are already "known unknowns", but which are not included in the analysis because of lack of knowledge or lack of computing power or some other reason. Confidence and likelihood need to reflect the sensitivity of potential decisions to such factors since they are not easily quantified in uncertainty estimation.

An uncertain future?

So while quantitative uncertainty estimation is valuable in assessing the range of potential outcomes consistent with an (agreed) set of assumptions, it will generally be the case that difficult to handle epistemic uncertainties will mean that the assessment is incomplete (for good epistemic reasons). Future

surprises come from that incompleteness (e.g. Beven, 2013). Assessments of evidence, and expressions of confidence and likelihood should reflect the potential for surprise and robust decisions need to be insensitive to both the assessed uncertainty and the potential for surprise (erring on the side of caution, risk aversion or being precautionary). From a modeller's perspective this has the advantage that it will reduce the possibility of a future post-audit analysis showing that the model predictions were wrong, even if why that is the case might be obvious with hindsight (it is quite possible that this will be the case with the current generation of climate models as future improvements start to reduce the errors in predicting historical precipitation, for example).

From a decision maker's perspective, the issues are more problematic. If, even with a detailed (and expensive) assessment of uncertainty, there remains a potential for surprise then just how risk averse or precautionary is it necessary to be in order to make robust decisions about the future. The answer is probably that we often cannot afford to be sufficiently robust in adapting to change; it will just be too expensive. The costs and benefits of protecting against different future extremes can be assessed, even if the probability of that extreme might be difficult to estimate. In that situation, the controlling factor is likely to be the available budget (Beven, 2011). That should not, of course, take away from the responsibility for ensuring that the science that underlies the evidence is as robust as possible, and communicated properly, even if those uncertainties are high and we cannot be very confident about future likelihoods in providing evidence to decision makers.

Acknowledgements

This work is a contribution to the CREDIBLE consortium funded by the UK Natural Environment Research Council (Grant NE/J017299/1). It is a written version of the Leonardo Lecture given at the Facets of Uncertainty meeting in Kos, Greece, in October 2013 with financial support from EGU. I am grateful to Paul Smith for carrying out the model runs on which the figures are based. I am also extremely grateful to the Steven Weijs, Grey Nearing and an anonymous referee who despite disagreeing with a lot of the paper made the effort to provide very detailed comments. The paper is much improved as a consequence of their efforts, albeit that we do not agree about much that is fundamental to the issues raised.

References

Anderson, M. P., and Woessner, W.W., 1992. The role of the post audit in model validation, *Advances in Water Resources*, 15, 167-173.

Ascough, J. C., Maier, H. R., Ravalico, J. K., and Strudley, M. W., 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling*, *219*(3), 383-399.

Bathurst, J. C., Ewen, J., Parkin, G., O'Connell, P. E., and Cooper, J. D., 2004. Validation of catchment models for predicting land-use and climate change impacts. 3. Blind validation for internal and outlet responses. *J. Hydrol.*, 287(1), 74-94.

Bell, V. A., Kay, A. L., Cole, S. J., Jones, R. G., Moore, R. J., and Reynard, N. S., 2012. How might climate change affect river flows across the Thames Basin? An area-wide analysis using the UKCP09 Regional Climate Model ensemble. *J. Hydrol.*, 442, 89-104.

685 Beven, K.J., 2002. Towards a coherent philosophy for environmental modelling, *Proc. Roy. Soc.*
686 *Lond. A,* 458, 2465-2484.
687 Beven, K.J., 2005. On the concept of model structural error, *Water Science and Technology*, 52(6),
688 165-175.
689 Beven, K.J., 2006a. A manifesto for the equifinality thesis, *J. Hydrol.*, 320, 18-36.
690 Beven K.J., 2006b. On undermining the science? *Hydrol. Process.*, 20, 3141–3146.
691 Beven, K. J., 2008. On doing better hydrological science, *Hydrol. Process.*, 22, 3549–3553.
692 Beven, K.J., 2009. *Environmental Modelling: An Uncertain Future?* Routledge: London.
693 Beven, K.J., 2010. Preferential flows and travel time distributions: defining adequate hypothesis
694 tests for hydrological process models, *Hydrol. Process.*, 24, 1537-1547.
695 Beven, K.J., 2011. I believe in climate change but how precautionary do we need to be in planning
696 for the future?, *Hydrol. Process.*, **25**, 1517–1520, DOI: 10.1002/hyp.7939.
697 Beven, K .J, 2012. Causal models as multiple working hypotheses about environmental processes,
698 *Comptes Rendus Geoscience, Académie de Sciences*, Paris, DOI:10.1016/j.crte.2012.01.005.
699 Beven, K.J., 2013. So how much of your error is epistemic? Lessons from Japan and Italy. *Hydrol.*
700 *Process.,* 27(11): 1677–1680, DOI: 10.1002/hyp.9648.
701 Beven, K.J. and A.M. Binley, 1992. The future of distributed models: model calibration and
702 uncertainty prediction, *Hydrol. Process.*, 6, 279-298.
703 Beven, K.J., Leedal, D. T., and McCarthy, S., 2014. Framework for assessing uncertainty in fluvial
704 flood risk mapping, CIRIA report C721 2014, available at
705 http://www.ciria.org/Resources/Free_publications/fluvial_flood_risk_mapping.aspx)
706 Beven, K.J., Smith, P. J., and Wood, A., 2011. On the colour and spin of epistemic error (and what
707 we might do about it), *Hydrol. Earth Syst. Sci.,* 15, 3123-3133, DOI: 10.5194/hess-15-
708 3123-2011.
709 Beven K.J., Westerberg, I., 2011. On red herrings and real herrings: disinformation and
710 information in hydrological inference, *Hydrol. Process.*, 25(10), 1676-1680, DOI:
711 10.1002/hyp.7963.
712 Beven, K.J. and Alcock, R., 2012, Modelling everything everywhere: a new approach to decision
713 making for water management under uncertainty, *Freshwater Biology,* 56,
714 DOI:10.1111/j.1365-2427.2011.02592.x.
715 Beven, K.J., P. Smith, I. Westerberg, and J. Freer, 2012, Comment on ''Pursuing the method of
716 multiple working hypotheses for hydrological modeling'' by P. Clark et al., *Water Resour.*
717 *Res.*, 48, W11801, DOI:10.1029/2012WR012282.
718 Beven, K.J., and P. Young, 2013, A guide to good practice in modeling semantics for authors and
719 referees, *Water Resour. Res.*, 49, DOI:10.1002/wrcr.20393..
720 Beven, K.J. and Binley, A. M., 2013, GLUE twenty years on, *Hydrol. Process.*, . 28(24):5897-5918,
721 DOI: 10.1002/hyp.10082
722 Beven, K.J., and Smith, P. J., 2014, Concepts of Information Content and Likelihood in Parameter
723 Calibration for Hydrological Simulation Models, *ASCE J. Hydrol. Eng.,* DOI:
724 10.1061/(ASCE)HE.1943-5584.0000991.
725 Blazkova, S., Beven, K.J, 2009. A limits of acceptability approach to model evaluation and
726 uncertainty estimation in flood frequency estimation by continuous simulation: Skalka
727 catchment, Czech Republic. *Water Resour. Res.*, 45: W00B16,
728 DOI:10.1029/2007WR006726
729 Brazier, R. E., Beven, K. J., Freer, J. and Rowan, J. S., 2000, Equifinality and uncertainty in
730 physically-based soil erosion models: application of the GLUE methodology to WEPP, the
731 Water Erosion Prediction Project – for sites in the UK and USA, *Earth Surf. Process.*
732 *Landf.*, 25, 825-845.
733 Boyd, I. 2003, Making science count in government, *eLife*, 2:e01061. DOI: 10.7554/eLife.01061
734 Cartwright, N. 1999, *The Dappled World: a Study of the Boundaries of Science*. Cambridge
735 University Press: Cambridge, UK.
736 Choi, H. T. and Beven, K. J., 2007, Multi-period and Multi-criteria Model Conditioning to Reduce
737 Prediction Uncertainty in Distributed Rainfall-Runoff Modelling within GLUE framework,
738 *J. Hydrol.*, 332 (3-4), 316-336.
739 Clark, M.P., D. Kavetski, and F. Fenicia, 2011, Pursuing the method of multiple working
740 hypotheses for hydrological modeling, *Water Resour. Res.*, 47, W09301,
741 DOI:10.1029/2010WR009827
742 Cloke, H. L., C. Jeffers, F. Wetterhall, T. Byrne, J. Lowe, and F. Pappenberger, 2010, Climate impacts
743 on river flow: projections for the Medway catchment, UK, with UKCP09 and CATCHMOD,

744 *Hydrol. Process.*, 24, 3476–3489.
745 Cox, R. T., 1946, Probability ,frequency and reasonable expectation, *American Journal of Physics*,
746 14: 1-13, DOI: 10.1119/1.1990764.
747 Dean, S, J. E. Freer, K. J. Beven, A. J. Wade and D. Butterfield, 2009, Uncertainty Assessment of a
748 Process-Based Integrated Catchment Model of Phosphorus (INCA-P), *Stoch. Environ. Res.*
749 *Risk Assess.* 23, 991–1010, DOI: 10.1007/s00477-008-0273-z.
750 Environment Agency, 2013, Benchmarking the latest generation of 2D hydraulic modelling
751 packages, *Report SC120002*, Environmental Agency: Bristol.
752 Faulkner, H., Parker, D., Green, C., Beven, K. J., 2007. Developing a translational discourse to
753 communicate uncertainty in flood risk between science and the practitioner, *Ambio*,
754 16(7), 692-703.
755 Freer, J.E., McMillan, H., McDonnell, J.J., and Beven, K.J., 2004. Constraining dynamic TOPMODEL
756 responses for imprecise water table information using fuzzy rule based performance
757 measures, *J. Hydrol.*, 291, 254-277, DOIi: 10.1016/j.jhydrol.2003.12.037.
758 Gupta, H. V., and G. S. Nearing, 2014. Debates—The future of hydrological sciences: A (common)
759 path forward? Using models and data to learn: A systems theoretic perspective on the
760 future of hydrological science, *Water Resour. Res.*, 50, 5351–5359,
761 DOI:10.1002/2013WR015096.
762 Hall J, O'Connell E, Ewen J. 2007. On not undermining the science: discussion of invited
763 commentary by Keith Beven. *Hydrol. Process.*, 21(7), 985–988.
764 Hamilton S. 2007. Just say NO to equifinality. *Hydrol. Process.*, 21(14), 1979–1980.
765 Intellectual Property Office, 2011. Good evidence for policy, UK Government (available at
766 http://www.ipo.gov.uk/consult-2011-copyright-evidence.pdf).
767 Iorgulescu, I, Beven, K J and Musy, A, 2005, Data-based modelling of runoff and chemical tracer
768 concentrations in the Haute-Menthue (Switzerland) research catchment, *Hydrol. Process.*,
769 19, 2557-2574.
770 IPCC, 2010. Guidance note for lead authors of the 5th IPCC Assessment on Consistent Treatment
771 of Uncertainties, available at http://www.ipcc.ch/pdf/supporting-material/uncertainty-
772 guidance-note.pdf.
773 IPCC, 2013. Headline statements from the summary for policymakers, available at:
774 http://www.ipcc.ch/news_and_events/docs/ar5/ar5_wg1_headlines.pdf .
775 Jain, A. and Dubes, R. 1998. *Algorithms for clustering data*. Prentice Hall.
776 Kauffeldt, A., S. Halldin, A. Rodhe, C.-Y. Xu, and I. K. Westerberg, 2013. Disinformative data in
777 large-scale hydrological modelling, *Hydrol. Earth Syst. Sci.,* 17, 2845-2857.
778 Kay, A. L., and Jones, R. G., 2012. Comparison of the use of alternative UKCP09 products for
779 modelling the impacts of climate change on flood frequency. *Climatic Change*, 114(2),
780 211-230.
781 Konikow, L. F. and Bredehoeft, J. D., 1992, Groundwater models cannot be validated? *Advances in*
782 *Water Resources,* 15, 75-83.
783 Koutsoyiannis, D., 2010. HESS Opinions" A random walk on water". *Hydrol. Earth Syst. Sci.*, 14(3),
784 585-601.
785 Koutsoyiannis, D. and Montari, A., 2015, Negligent killing of scientific concepts: the stationarity
786 case, *Hydrol. Sci. J.,* in press.
787 Krueger, T., Quinton, J. N., Freer, J., Macleod, C. J., Bilotta, G. S., Brazier, R. E., Butler, P. and
788 Haygarth, P. M., 2009. Uncertainties in data and models to describe event dynamics of
789 agricultural sediment and phosphorus transfer. *J. Environ. Qual.*, 38(3), 1137-1148.
790 Kuczera, G., Renard, B., Thyer, M. and Kavetski, D., 2010. 'There are no hydrological monsters, just
791 models and observations with large uncertainties!', *Hydrol. Sci. J.*, 55(6), 980-991.
792 Leedal, D. T., J. Neal, K. J. Beven, P. Young and P. Bates, 2010. Visualization approaches for
793 communicating real-time flood forecasting level and inundation information*, J. Flood Risk*
794 *Management,* 3, 140-150.
795 Liu, Y-L., Freer, J., Beven, K., Matgen, P., 2009. Towards a limits of acceptability approach to the
796 calibration of hydrological models: Extending observation error. *J. Hydrol.*, 367(1-2), 93-
797 103.
798 Mantovan P, Todini E., 2006. Hydrological forecasting uncertainty assessment: incoherence of the
799 GLUE methodology. *J. Hydrol.*, 330, 368–381.
800 McBratney, A. B., 1992. On variation, uncertainty and informatics in environmental soil
801 management. *Soil Research*, 30(6), 913-935.
802 McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M., 2010. Impacts of uncertain

river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrol. Process.*, 24(10), 1270-1284.

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

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

Montanari A. 2007. What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology. *Hydrol. Process.*, 21, 841–845.

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

Neal. J., C. Keef, P. Bates, K. J. Beven and D. T. Leedal, 2013. Probabilistic flood risk mapping including spatial dependence, *Hydrol. Process.*, 27: 1349-1363, DOI: 10.1002/hyp.9572.

Nott, D. J., L. Marshall, and J. Brown, 2012. Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resour. Res.*, 48, DOI:10.1029/2011WR011128.

Pappenberger, F., Beven, K.J., Frodsham, K., Romanovicz, R. and Matgen, P., 2007. Grasping the unavoidable subjectivity in calibration of flood inundation models: a vulnerability weighted approach. *J. Hydrol.*, 333, 275-287.

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

Parkin, G., O'Donnell, G., Ewen, J., Bathurst, J. C., O'Connell, P. E., and Lavabre, J., 1996. Validation of catchment models for predicting land-use and climate change impacts. 2. Case study for a Mediterranean catchment. *J. Hydrol.*, *175*(1), 595-613.

Raadgever, G. T., Dieperink, C., Driessen, P. P. J., Smit, A. A. H., and Van Rijswick, H. F. M. W., 2011. Uncertainty management strategies: Lessons from the regional implementation of the Water Framework Directive in the Netherlands. *Environmental Science & Policy*, *14*(1), 64-75.

Regan, H. M., Colyvan, M., & Burgman, M. A., 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, *12*(2), 618-628.

Reggiani P., Hassanizadeh S. M., Sivapalan M. and Gray W. G. 1999, A unifying framework for watershed thermodynamics: Constitutive relationships. *Advances in Water Resources*, 23, 15–39.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M. & Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. *Water Resour. Res.*, 46, W05521, 22pp, DOI:10.1029/2009WR008328.

Romanowicz, R. and Beven, K. J., 2003. Bayesian estimation of flood inundation probabilities as conditioned on event inundation maps, *Water Resour. Res.,* 39(3), W01073, 10.1029/2001WR001056.

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

Sadegh, M., and J. A. Vrugt, 2013. Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation, *Hydrol. Earth Syst. Sci.,* 17, 4831–4850, DOI:10.5194/hess-17-4831-2013.

Sadegh, M., and J. A. Vrugt, 2014. Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM(ABC), *Water Resour. Res.*, 50, 6767–6787, DOI:10.1002/2014WR015386.

Schoups, G., and J. A. Vrugt, 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors, *Water Resour. Res.*, 46, W10531, DOI:10.1029/2009WR008933.

Sikorska, A. E., Montanari, A and Koutsoyiannis, D., 2014. Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques, *ASCE J. Hydrol. Eng.* DOI: 10.1061/(ASCE)HE.1943-5584.0000926.

Sivakumar, B. 2008. Undermining the science or undermining Nature? *Hydrol. Process.*, 22, 893–897.

Smith, P. J., Tawn, J., Beven, K. J., 2008. Informal likelihood measures in model assessment: theoretic development and Investigation. *Advances in Water Resources,* 31, 1087–1100,

862 DOI:10.1016/j.advwatres.2008.04.012.
863 Stedinger, J. R., Vogel, R. M., Lee, S. U., and Batchelder, R., 2008. Appraisal of the generalized
864 likelihood uncertainty estimation (GLUE) method. *Water Resour. Res.*, 44, W00B06,
865 DOI:10.1029/2008WR006822.
866 Todini E. and Mantovan P., 2007. Comment on: 'On undermining the science?' by Keith Beven.
867 *Hydrol. Process.*, 21(12), 1633–1638
868 Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., and Robinson, B. A., 2008. Treatment of input
869 uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain
870 Monte Carlo simulation. *Water Resour. Res.*, 44(12). W00B09,
871 DOI:10.1029/2007WR006720.
872 Vrugt, J. A., and M. Sadegh, 2013. Toward diagnostic model calibration and evaluation:
873 Approximate Bayesian computation, *Water Resour. Res.*, 49, DOI:10.1002/wrcr.20354.
874 Westerberg, I., Guerrero, J.-L., Seibert, J., Beven, K. J., and Halldin, S., 2011a. Stage-discharge
875 uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras,
876 *Hydrol. Process.*, 25, 603-613, DOI: 10.1002/hyp.7848.
877 Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E. and
878 Xu, C. Y., 2011b. Calibration of hydrological models using flow-duration curves. *Hydrol.*
879 *Earth Syst. Sci.*, 15(7), 2205-2227.
880 Wicks, J. M., Adamson M., and Horritt M., 2008. Communicating uncertainty in flood maps - a
881 practical approach, *Defra Flood and Coastal Management Conference*, Manchester.
882
883

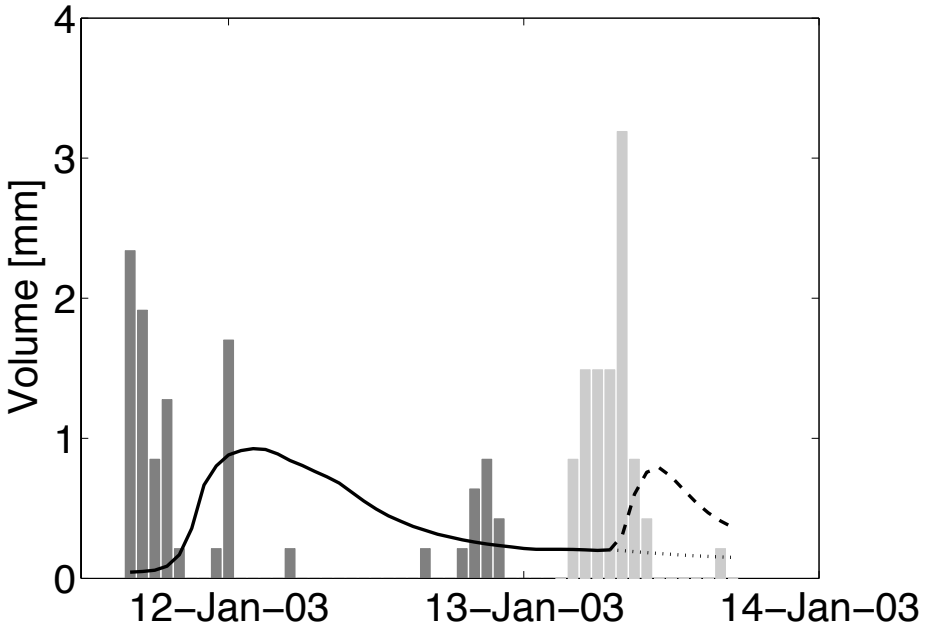
884 **Table 1. A classification of different types of uncertainty**
885

Type of Uncertainty	Description
Aleatory Uncertainty	Uncertainty with stationary statistical characteristics. May be structured (bias, autocorrelation, long term persistence) but can be reduced to a stationary random distribution
Epistemic Uncertainty (system dynamics)	Uncertainty arising from a lack of knowledge about how to represent the catchment system in terms of both model structure and parameters. Note that this may include things that are included in the perceptual model of the catchment processes but which are not included in the model. They may also include things that have not yet been perceived as being important but which might result in reduced model performance when surprise events occur.
Epistemic Uncertainty (forcing and response data)	Uncertainty arising from lack of knowledge about the forcing data or the response data with which model outputs can be evaluated. This may be because of commensurability or interpolation issues when not enough information is provided by the observational techniques to adequately describe variables required in the modelling process. May be a function of a limited gauging network, lack of knowledge about how to interpret radar data, or non-stationarity and extrapolation in rating curves.
Epistemic Uncertainty (disinformation)	Uncertainties in either system representation or forcing data that are *known* to be inconsistent or wrong. Real surprises. Will have the expectation of introducing disinformation into the modelling processes resulting in biased or incorrect inference (including false positives and false negatives in testing models as hypotheses)
Semantic / Linguistic Uncertainty	Uncertainty about what statements or quantities in the relevant domain actually mean (there are many examples in hydrology including storm runoff, baseflow, hydraulic conductivity, stationarity etc). This can partly result from commensurability issues that quantities with the same name have different meanings in different contexts or scales.
Ontological Uncertainty	Uncertainty associated with different belief systems. Relevant example here might be beliefs about whether formal probability is an appropriate framework for the representation of beliefs about the nature of model residuals. Different beliefs about the appropriate assumptions could lead to very different uncertainty estimates so that every uncertainty estimate will be conditional on the underlying beliefs and consequent assumptions.

886
887 **Table 2. Defining non-stationarity. Different classes of epistemic error**
888 **that lead to non-stationarity in model residual characteristics.**

Class	Source	Description
1	Non-stationarity of a stochastic process	Change over time that can be described by a deterministic function, including structure in model residuals that might compensate for consistent model or boundary condition error. All other variability will be stochastic in nature (see Koutsoyiannis and Montanari, 2015)
2	Non-stationarity in catchment characteristics	Expectation that model parameters and possibly structure representing catchment characteristics will change over time or space in a way that will induce model prediction error if parameters are considered stationary
3	Non-stationarity in boundary conditions	Expectation that model boundary conditions will change over time or space in a way that will induce model prediction error if boundary conditions are poorly estimated. In some cases may include disinformative data as defined in the text.
4	Non-stationarity in model residual characteristics	Expectation that the statistical characteristics of the model residuals will vary significantly in time and space because of epistemic uncertainties about the causes of the unpredictable model error. May result from arbitrary epistemic uncertainties in boundary conditions, long-term stochastic variability or inclusion of disinformative calibration data.

889
890
891



892
893 **Figure 1. Example of an event where the runoff coefficient based on the measured**
894 **rainfalls and stream discharges is about 1.4. This clearly violates mass balance and will**
895 **therefore be disinformative in calibrating a model that is constrained to maintain mass**
896 **balance to represent that catchment area.**
897

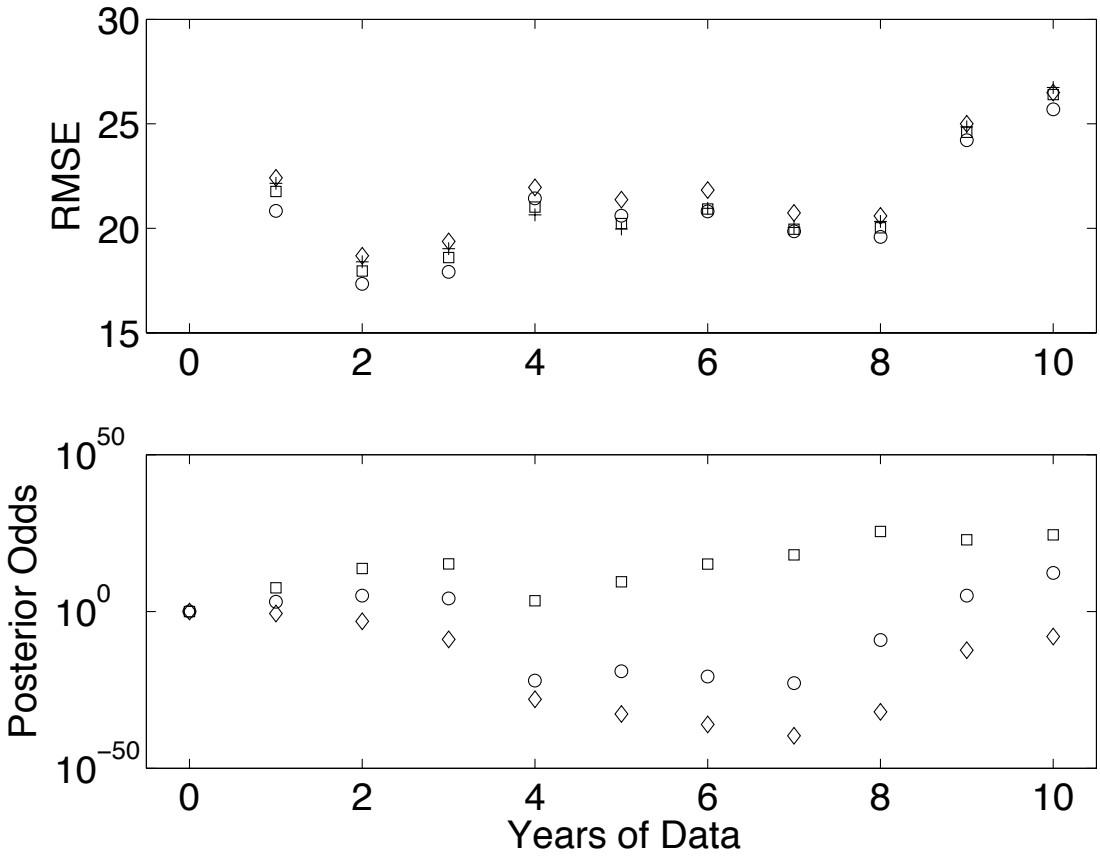
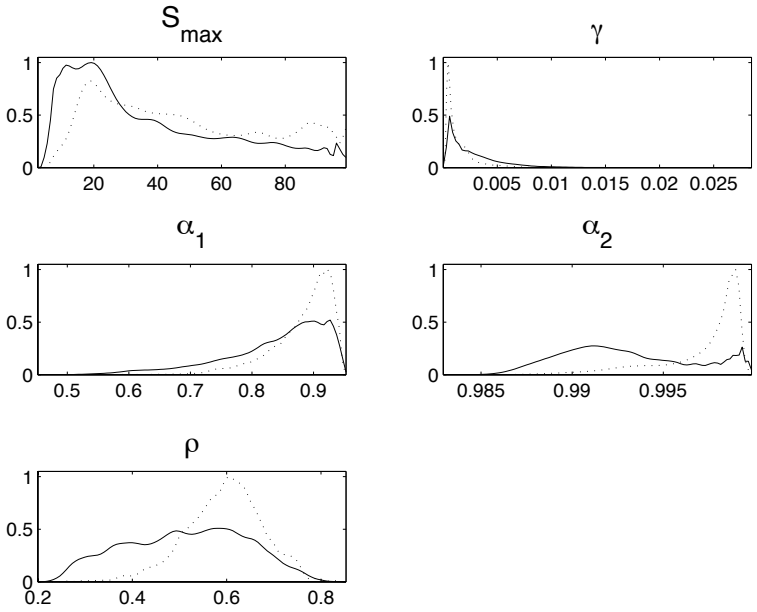
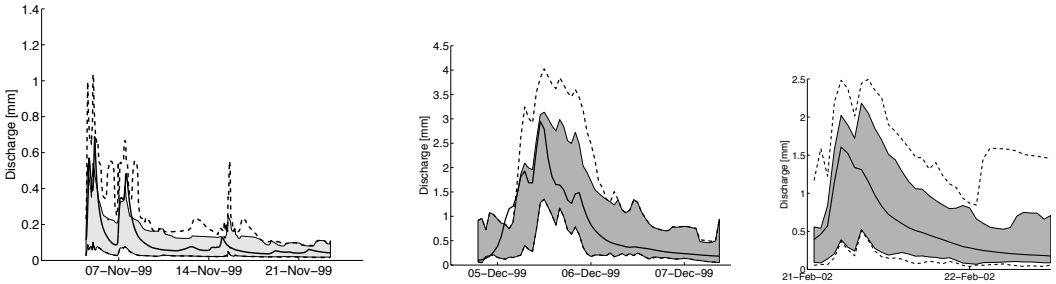


Figure 2. Top: Root mean square errors for four model parameter sets within the same model structure (a simple single tank conceptual rainfall-runoff model, see Beven and Smith, 2014). Bottom: Likelihood ratios or posterior odds for 3 of the models, relative to the first (+ symbol in upper plot), evaluated using a formal likelihood and updated after the addition of further years of model residuals. The formal likelihood used allows for a mean bias, constant variance and 1st order autocorrelation and assumes a Gaussian distribution of model residuals. While similar in root mean square error (and visual performance), the different models have likelihood ratios that evolve to be 10^{40} different as 6 years of data are added, followed by a rapid reduction in likelihood ratio over the next 3 years.



911
912 **Figure 3. Posterior probability density functions for model parameters evaluated both**
913 **with (solid line) and without (dotted line) calibration events classified as disinformative.**
914 **Further details of this study can be found in Beven and Smith (2014).**
915



916
917
918 **Figure 4. A sample of events taken from the model evaluation period. Each event is**
919 **treated as if it is either informative (shaded 95% prediction bounds) or disinformative**
920 **(dotted 95% prediction bounds). The first event is evaluated (a posteriori) as**
921 **disinformative, the last two as informative. Further details of this study can be found in**
922 **Beven and Smith (2014).**
923
924