How to make advances in hydrological modelling

President's Invited Address, BHS Symposium, University of Westminster, September 2018

Keith Beven
Lancaster Environment Centre
Lancaster University
Lancaster, UK
k.beven@lancaster.ac.uk

Abstract

After some background about what I have learned from a career in hydrological modelling, I present some opinions about how we might make progress in improving hydrological models in future including how to decide whether a model is fit for purpose; how to improve process representations in hydrological models; how to take advantage of Models of Everywhere. Underlying all those issues, however, is the fundamental problem of improving the hydrological data available for both forcing and evaluating hydrological models. It would be a major advance if the hydrological community could come together to prioritise and commission the new observational methods that are required to make real progress.

1. Some Background

My first attempt at a hydrological model was produced as an undergraduate student at the University of Bristol in about 1970. It was an attempt to model the famous Lynmouth Flood in 1952. It was programmed in Algol and physically existed as a pack of punched cards that needed to be fed into a card reader every time a run was made (compilation errors, run-time errors and, eventually, production runs included). The primary data available were rainfall records, so the only "calibration" data were indirect post-flood estimates of a peak discharge. This was a highly sediment laden flow that transported some huge boulders, so any such estimate would have been highly uncertain. Even so, that simple study taught me a lot about the importance of antecedent conditions in trying to predict flood discharges; the wetting of the catchment prior to the flood was extremely important (as has also been the case in many more recent cases of flash flooding in the UK).

In starting my PhD at the University of East Anglia in Norwich in 1971, I made a survey of hydrological models in the literature. Even at that time there was a plethora of different models. With the more widespread availability of digital computers in the late 1960s, many PhD projects and consultants were producing their own models. Most of these were conceptual models of the Stanford Watershed Model type, which itself was the PhD project of Norman Crawford at Standford University under the direction of Ray K. Linsley (Crawford and Linsley, 1966). This model was the foundation of the Hydrocomp consultancy and I met both of them when I was able to participate, while still a PhD student myself, at the first UK Hydrocomp workshop. The Hydrocomp Simulation Programme in Fortran (HSPF) was later adopted by the US EPA and remains in use as a freely available tool. When I gave up my count of models at over 100 in 1971, I was already asking the question of how can we do better?

My response was to try and be objective. To base a model on best physical principles, and to measure rather than calibrate the model parameters. Al Freeze was already advocating this in the Freeze and Harlan paper in Journal of Hydrology in 1969, and implementing it using finite difference methods at the Thomas J. Watson research centre of IBM at Yorktown Heights. I took a slightly different strategy, using finite element methods to solve the Richards equation so that my hillslopes and soil horizons on those hillslopes could look more natural in the discretisation grid (I had a physical geography rather than engineering degree after all!). I did not have quite the same resources as Al Freeze. The model was implemented as two full boxes of computer cards (with all the same issues of compilation and run-time errors, but now with many more cards to get through the card reader successfully) and ran on an ICL1904 mainframe computer. I also carried out the laboratory work necessary to determine all the soil moisture characteristics on soil cores and the field work necessary for channel cross-sections and roughness. The model was applied to the East Twin catchment in the Mendips that had been studied by Darrell Weyman (1970, 1973) for his PhD and the results were really rather bad, a fact noted by my PhD examiner, Terrence O'Donnell. They were finally published as part of my Dalton Lecture paper (Beven, 2001) which tells the story of how that experience shaped my research career.

I was fortunate to then work with Mike Kirkby as a post-doc at the University of Leeds on the development of Topmodel (see Beven and Kirkby, 1979) based on Mike's concept of the topographic index. Given my experience of physically-based modelling I was more than happy to take another approach, but still one that allowed the results of the modelling to be mapped back into space (I was again in a Geography department). The only problem of that was that both the topographic analysis that went into Topmodel, and the analysis of the spatial nature of the results had to be done manually. There were no Digital Terrain Models, and computer outputs were still on lineprinter paper. We were also running a nested catchment experiment, with both rainfall and stream level data recorded on paper charts so a lot of time was spent just getting the data into computer compatible form (e.g. Beven and Callen, 1979). One of the nice outcomes of that project was that we demonstrated a model structure that could be applied successfully based on field measured parameters (Beven et al., 1984). I also learned that parameter optimisation would not necessarily use the model concepts in the correct way (see, for example, Fig. 14 of Beven and Kirkby, 1979).

My experience with these different types of models proved valuable in being appointed (actually as a "mathematic modeller" despite my geography degree) at the Institute of Hydrology (IH) in Wallingford in 1977. Part of my time was devoted to the SHE (Système Hydrologique Européen) project, a joint initiative with the Danish Hydraulics Institute (DHI) and SOGREAH in France, funded by a European Community loan. This was another attempt at producing a complete "physically-based" hydrological model and was led by Mike Abbott who had successfully dealt with the numerical issues of solving the shallow water equations in hydraulics which were the basis of the DHI MIKE series of simulation packages. Before I joined IH I had participated in the first SHE meeting at Wallingford, and the minutes record that, as a result of my PhD modelling experience, I raised many of the problems that would be met in the SHE project particularly in the decoupling of the saturated and unsaturated zone solutions. This was a pragmatic decision to reduce dimensionality, based on the available computer resource, but was the main reason why it was 1986 before the first SHE

applications appeared (Abbott et al., 1986a,b; Bathurst et al. 1986). It was later relaxed as computer power increased and much later both the MIKE-SHE and SHETRAN versions of SHE were implemented with fully 3D partially saturated Darcy-Richards subsurface solutions (see Ewen et al., 2000; Graham and Butts, 2005). Speed could still be an issue, however, and MIKE-SHE has also been used with conceptual groundwater storage components in some applications (see the history of SHE in Refsgaard et al., 2010).

After three years in Wallingford, in 1979 I moved to the University of Virginia and was able to return to working with Topmodel. I took advantage of its computational speed and the availability of a CDC6600 mainframe computer to start making Monte Carlo runs of the model in around 1980. This soon showed that there were many runs of the model with different parameter sets that gave more or less equivalent results, something that was later developed into the equifinality concept (Beven, 1993, 2006) though equifinality had already been mentioned in my PhD thesis (Beven, 1975). It was also the origin of the Generalised Likelihood Uncertainty Estimation (GLUE) methodology, although I did not have the confidence to publish this until much later (Beven and Binley, 1992). Returning to Wallingford in 1982, I told the Director, Jim McCulloch, that I would not work on SHE but there was also funding for another physically-based model the Institute of Hydrology Distributed Model or IHDM that had been started by Liz Morris. We rewrote the IHDM, producing Version 4, that was based on finite element rather than finite difference methods. It was therefore rather similar to my PhD model but with better numerics and finer discretisations because of more computer resource. The numerics of the IHDM were later improved still further by Ann Calver and Winifred Wood (Calver and Wood, 1989) but remained subject to the problems of using the Richards equation as a representation of flows in real soils (see, for example, Beven, 1989, a paper that started out as a commentary on the first 1986 SHE applications).

In 1985 I moved to Lancaster and continued work on 3D finite element modelling with Andy Binley (e.g. Binley et al., 1989a,b; Binley and Beven, 1992); Topmodel and the development of Dynamic Topmodel with Jim Freer (Beven and Freer, 2001); modelling flow and transport for water quality (e.g. Page et al., 2007; Dean et al. 2009; Hollaway et al., 2018a); pollutant dispersion and flood forecasting using the Data-Based Mechanistic (DBM) methods of Peter Young (e.g. Wallis et al., 1989; Young and Beven, 1994); and a wide range of applications of the GLUE methodology (e.g. Beven, 2009, 2016, and the references therein). Some of that work proved controversial, in particular about whether informal likelihood measures and rejection criteria could replace formal statistical methods in model evaluation (see, for example, Beven, 2006b, 2008; Andréassian et al., 2007; Todini and Mantovan, 2007; Hall et al., 2007). However, controversy encourages harder thinking about what is important and what is required to go beyond the norms of the current paradigm and make real advances.

This background frames the comments about how to make advances in hydrological modelling that are set out in the following sections. This essentially updates the final chapter of Beven (2012a). I concentrate on what I see as the three most important issues. These are: how to decide whether a model is fit for purpose; how to improve process representations in hydrological models; how to take advantage of Models of Everywhere. Underlying all those issues, however, is the fundamental problem of improving the hydrological data available for both forcing and evaluating hydrological models.

2. The need to improve hydrological data for model applications

Hydrological data are highly uncertain (see, most recently, Beven, 2019). This is true for the most basic of quantities, such as rainfall at a point, discharge at a point (particularly at the highest and lowest flows), and actual evapotranspiration fluxes (sort of at a point). It is even more problematic if we are interested in the water balance over a catchment area because there are uncertainties in catchment area rainfall, snowfall, evapotranspiration fluxes and storages. The issue is greater because in general the uncertainties involved are the result of a lack of knowledge (i.e. *epistemic* uncertainties) rather than random variability (the *aleatory* uncertainties) (see, for example, Kauffeldt et al., 2009; Beven, 2016a; Westerberg et al., 2016; Wilby et al., 2017). In some cases, we choose to treat data uncertainties *as if* they are aleatory because the convenience of the statistical techniques available (e.g. kriging for the interpolation of areal rainfalls; repeat measurements for ADCP estimates of flows; the choice and fitting of flood frequency distributions). A good example is the use of statistical regression for fitting rating curves for the conversion of observed water levels to discharges. It is often assumed that some simple power law will hold over the range of the data (with or without an offset, with or without multiple segments). This might be satisfactory within the range of the actual gaugings, at least if they are not too variable and if effects such as weed growth and a mobile bed are negligible, but in some cases such extrapolations can be potentially misleading (see for example Beven et al., 2012; McMillan and Westerberg, 2015; Hollaway et al., 2018; and the comparison of Kiang et al., 2018).

There is some movement towards the use of extrapolations based on hydraulic modelling of a gauging site, particularly for overbank flows. The review of the Sheepmount rating curve at Carlisle after the 2005 flood is a good example. Consulting engineers were commissioned by the Environment Agency to revisit the rating curve at this site using hydraulic modelling, since the recorded water level was over a metre higher than the highest measured discharge. This led to a significant increase in the estimated discharge relative to that produced by extrapolation of the rating curve fitted to the discharge measurements. The revised rating was then used to estimate the even higher flood peak from Storm Desmond in 2015. However, such estimates are very dependent on the estimation of effective roughness coefficients for the out-of-bank conditions, which is necessarily uncertain. Extrapolated discharge estimates are still often cited without any associated uncertainty range even though there is evidence that effective roughness might be model structure dependent and vary with peak magnitude (e.g. Romanowicz and Beven, 2003; Pappenberger et al., 2006).

These experiences led to Beven et al. (2011) and Beven and Smith (2015) suggesting that some catchment data might be *disinformative* in deciding whether a model is acceptable or not. They identified events that gave exceedingly high or exceedingly low runoff coefficients in a rapid response catchment in the north of England. Clearly if a model is constrained by mass balance, but the data for an event suggest a runoff coefficient greater than 1, then the model is going to produce residuals that reflect the deficiencies in the original data, not only from any failure of the model (see also the examples in global data sets included in Kauffeldt et al., 2013). In this case the problem is quite evident, and if such data are included in model evaluation will lead to bias in inference about parameter values and in predicted outcomes, especially if simple evaluation measures based on the sum of squared errors are used. There

will, however, be many other periods of data when the effects on model evaluation will be subtle and difficult to allow for.

The conclusion of this is that we need to be much more careful about considering the value of the available data in model evaluation, and that we need better observational techniques, not only for the inputs and outputs in the water balance equation but also for internal state variables. In the latter case, there is still a great deal of epistemic uncertainty about subsurface flow pathways on hillslopes (and in valley bottoms). Where internal state data are used there can also be incommensurability between observed variables and simulated variables (e.g. soil moisture at a point relative to the soil moisture output at the discrete element scale of a distributed model). There have been some advances, such as the COSMOS measurement of soil moisture over an area, but that has both variable effective depths and areal extent depending on the levels of near-surface moisture (Zreda et al., 2012; Evans et al., 2016; Baroni et al., 2018).

We also know enough from tracing experiments and the nature of the physics to conclude that the Richards equation should not be used in modelling flow through soils (in fact, we argued this in Beven, 1989, and Binley et al., 1989a,b, nearly 30 years ago). It is based on the wrong experiment that excluded the possibility of preferential flows in focussing on capillary equilibrium conditions. This might be more applicable under relatively dry conditions but even then, the physics itself suggests that the usual form should not be used if there is any heterogeneity of soil properties within the scale of the application, which is, of course, *always* the case (see also Beven, 2012a, 2018b; Beven and Germann, 1992; 2013). However, we have no good (non-destructive) measurement techniques at scales of interest with which to study vertical and downslope preferential flows and recharge. Those detailed observations that have been done have suggested that the flows can be highly localised, highly variable, and subject to complex connectivity issues in space and time (e.g. Freer et al., 1997; Jensco et al., 2009; McGuire et al., 2010; Klaus and Jackson, 2018).

In fact, we are not interested in such detail (except in terms of scientific understanding) and it might be better to develop new measurement techniques at larger scales that would integrate over the detail. If, for example, we had an effective and affordable gravity anomaly technique for total water storage over an area; coupled with a method for measuring stream discharges that was sufficiently accurate to determine incremental discharges downstream in a river network, then we might be able to infer much more useful process relationships than those we have currently. However, as a community we have not been at all pro-active about deciding on priorities for measurement requirements and commissioning new techniques.

The satellite community have done so much more effectively (including the SWOT launch planned for 2021 which will be of some hydrological interest), but from a hydrological point of view satellite imaging has always had potential but not actually been that useful, apart from generating digital terrain data, particularly LIDAR that has led to significant improvements in, for example, flood inundation mapping. Even then, however, there are both aleatory and epistemic uncertainties associated with the treatment of the digital numbers (how to deal with vegetation and buildings; small scale features such as walls and hedges on flood plains; later infilling of sinks or burn-in of channels in the terrain to get

consistent flow lines; determination of catchment boundaries etc) that will have an effect on any model outputs when compared with observations. Most other remote sensing is also associated with epistemic uncertainties, including rainfall and soil moisture estimation, with the result that it provides only some qualitative and uncertain indication of patterns in the landscape relevant to hydrology.

Improving the quantity and quality of hydrological data is essential to what follows, in particular in deciding on whether particular models might be fit-for-purpose. Note that this paper is about how to make improvements in hydrological simulation models. It is not about models used for forecasting, i.e. modelling using data assimilation for getting the best real time n step ahead predictions with minimal uncertainty (see Beven and Young, 2013, for a discussion of different types of model prediction). Forecasting does not necessarily require process representations, nor physical constraints such as mass balance that may not be a feature of the available data. This is particularly true in forecasting flood events when there may be poor sampling of the most intense rainfalls and the discharge rating curve may be subject to epistemic uncertainties. Data assimilation is then a valuable tool in improving forecasts.[1] Far better to forecast levels and use data assimilation to compensate for the limitations in the input data (see for example Romanowicz et al., 2006; Leedal et al., 2010).

Here I shall be interested in the representation and simulation of hydrological processes in the context of not only reproducing historical behaviours but also future behaviours under change. Even a cursory survey of the literature will reveal that this is a challenge and difficult to achieve. Hydrological systems are complex and nonlinear, and we have little in the way of techniques for studying patterns of processes at the catchment scale. We rely on the way in which catchments act as integrators over small scale complexity and heterogeneity in resorting to calibration of simple model representations against the very discharge data that we want to predict. That clearly helps in getting better reproduction of discharges without change but not necessarily for the right reasons. Getting good results for the wrong reasons could then be misleading when we want to simulate the impacts of change (rarely is any consideration given to change during a calibration period, but see Merz et al., 2011; Peel and Blöschl, 2011; Harrigan et al., 2014). In the past I have had some success in making predictions using only measured parameters (e.g. Beven et al. 1984), but also some notable failures (e.g. Beven, 2001).

3. Evaluating hydrological models as fit-for-purpose

We know very well that the process representations used in hydrological models are only approximations to the real-world complexity of surface and subsurface flows. It is also obvious that the epistemic issues with hydrological data mean that we would not expect even a perfect model to provide perfect predictions. We see this in the comparisons of observed

[1] Note that while I consider data assimilation to be essential in forecasting, I do not consider it to be good practice to use data assimilation to compensate for model deficiencies in simulation modelling, especially if there is no attempt to learn from the data assimilation about how a model might be in error. There have been a number of such studies in the literature. Clearly it is not possible to use data assimilation to compensate for model deficiencies in simulating the impacts of future changes. It is better then to attempt to produce a realistic estimate of the associated uncertainties, both aleatory and epistemic.

and predicted variables in a multitude of academic papers and reports to clients. Sometimes, indeed, it seems that the predictions are rather poor, especially if models are applied without calibration as if to an ungauged catchment. Calibration is generally helpful in finding parameter sets that give predictions that are closer to the observations, at least in the calibration period. When a split record evaluation is also done, it is common to find that the model performance is not so good in the validation period or under different seasonal or climate conditions (Refsgaard and Knudsen, 1996; Freer et al., 2003; Choi and Beven, 2007; Coron et al., 2014; Dakhlaoui et al., 2017; Pool et al., 2017; Fowler et al, 2016, 2018). This might be the result of over-fitting an overparameterised model; it might be because the model is producing good results in calibration for the wrong reasons; it might be only because the forcing data errors are quite different in the validation period. For more severe testing (see Klemes, 1986; Refsgaard and Knudsen, 1996; Ewen and Parkin, 1997; Seibert, 2003) it is often difficult to declare any form of success.

We can think about models as hypotheses about how a hydrological system functions (e.g. Beven, 2012b, 2018a). Thus, testing whether a model should be considered as fit-for-purpose can be considered a form of hypothesis testing, with the possibility of rejecting models that do not fit the evaluation data to some defined level of acceptability. Model rejection in this sense is a good thing; it means that we need to make some improvements, either to the model structure or to the data that we are using with the model (Beven, 2018a). Clearly, methods for hypothesis testing are well developed in statistics, under assumptions that variables can be considered to have aleatory variability. However, when we *know* that we are dealing with epistemic uncertainties it might be incoherent to use simple statistical assumptions (e.g. Beven et al. 2008). This is evident, for example, in the use of formal likelihood functions in model evaluation that, particularly for long time series, can give quite a misleading impression of the relative merits of different models and parameter sets (e.g. Beven and Smith, 2015; Beven, 2016a).

In assessing fitness-for-purpose, of course, we do need to consider what is the purpose. We can differentiate between two major types of purpose (though each could have a variety of subdivisions). The first is in the use of models to test the science, i.e. the understanding of how a hydrological system might function. This might involve the more detailed consideration of the internal states and other detail in experimental plots and catchments, and how they differ from responses reported from elsewhere. The second is in the use of models for decision making. The important factor then is that the model should make predictions of the future behaviour of a hydrological system that will not deviate too far from what would happen under the assumed boundary conditions. This might allow a greater degree of approximation to be considered to be acceptable, especially if decisions are being taken at larger scales (such as in the methods used for the UK National Flood Risk Assessment that is currently under revision). A particular feature of this second purpose is that the results cannot really be tested, even if a model has survived a validation test, since the future boundary conditions are necessarily unknown or epistemically uncertain (see for example the post-audit analyses of groundwater models in Konikow and Bredehoeft, 1992, where some models failed only because of poor assumptions about the future boundary conditions). We might hope, of course, that as the science evolves, the purpose of improving understanding will feed into the purpose of decision making, with a better theoretical basis for moving from

local scales to national scales and for assessing changes in parameter values but we are not there yet (see below).

The question remains of how should we test models as hypotheses in the face of epistemic uncertainties? Beven and Lane (2019) suggest that one way of looking at this problem is in the form of testing for model invalidation (see also Beven, 2018a). There is, of course, a long history of applying such tests, at least implicitly in the form of not invalidating a model based on its simulated outputs. Every time a referee accepts a paper with model results for publication, s/he is essentially applying such a test. Every time a report is presented to a client, then the authors of that report have applied such a test. Every time a report is accepted by the client (perhaps after an independent assessment by another consultant) then such a test has been applied. Most of these judgments are qualitative and subjective, albeit that they may be supported by some quantitative measures (such as quoting the Nash-Sutcliffe efficiency despite all of its faults as a measure of calibration or validation performance).

It is therefore interesting to speculate about what information such a group of experts would require in order to make such an invalidation more rigorous, both in the use of models for predicting an ungauged catchment and in the case where some output observations are available to evaluate model runs. One interesting feature of this strategy is that there is a possibility for the users of the model outputs, such as decision or policy makers or stakeholders affected by a decision, to be involved in such a process in considering not only the acceptability of the model outputs but also the assumptions that contribute to the outputs (see Beven, 2018a, and the condition tree approach of Beven et al., 2014).

There is actually a precedent for this type of approach in the "blind validation" approach of Ewen and Parkin (1996). This requires the modeller (in their case) to define some criteria for acceptability prior to making any model runs. Model parameters were estimated from past experience and no prior model calibration was allowed. The range of simulated outcomes was then compared with available observations of flows and internal state data (assumed at that time to be known accurately). Blind validation was applied to the SHE model by Parkin et al. (1996) and Bathurst et al. (2004). In both cases, the model simulations failed to meet all the defined validation criteria. In the application of Parkin et al. (1997) the model failed 1 out of 4 tests; in the case of Bathurst et al. (2004) 2 out of 10 tests were failed. This was despite the criteria for success being rather relaxed and some model simulations being excluded on the basis of expert evaluations. These failures do not seem to have had much effect on the use of the SHE model elsewhere. In fact, the failures are not mentioned at all in the SHE review paper of Refsgaard et al. (2010), which includes just a brief passing mention of the development of model testing methods based on the Klemes (1986) concepts. There have been no other applications of this blind validation methodology, to my knowledge, though it has much in common with the setting of limits of acceptability within the Generalised Likelihood Uncertainty Estimation (GLUE) methodology (see Beven, 2006a, 2009, 2016a) that has led to some other model invalidations (e.g. Page et al., 2007; Dean et al., 2009; Liu et al., 2009; Hollaway et al. 2016a).

One of the issues in this type of evaluation is, again, the data being used to both drive and test a model as hypothesis. Since we do not expect a model to be better than the data it is

used with, any invalidation test should first make some assessment and allowance for the uncertainties, both epistemic and aleatory, associated with those data, although in some (wet) cases any model that gets the water balance separation approximately correct might provide quite good measures of performance (e.g. Seibert et al., 2018). How uncertain do we expect the inputs used to force the model to be? If we have observations of the system response, how uncertain are those observations relative to the variables predicted by the model? We do not expect this assessment of uncertainty to be a simple statistical variability (though lacking better knowledge we might choose to treat it as such). We are not used to framing model testing in this way (and indeed perhaps we have avoided it because these are very difficult questions to resolve when we expect the nature of errors in the inputs to vary from event to event, and parameter interactions to be complex). Data uncertainty also raises the issue of how to avoid Type I hypothesis testing errors (accepting a model hypothesis that is not fit-for-purpose because of the data uncertainties) and Type II errors (rejecting a model hypothesis that would be fit-for-purpose because of the data uncertainties). The former is more problematic but should hopefully be reduced as new data or different types of data are added to the assessment. Such difficulties should not, however, stop us from thinking more deeply about how to make an invalidation test more rigorous.

A further feature that might be considered is whether a model contradicts some secure evidence on the nature of the system response. If that is the case, it should not be considered as fit-for-purpose. We want to base decisions on predictions from a model that, as far as possible, is producing the right results for the right reasons. A nice example of this appears in the very first Topmodel paper (Beven and Kirkby, 1979) where it was shown that optimising the model parameters resulted in using the model structure in a way that contradicts the theory on which it was based by using the subsurface store with a very low time constant to control the timing of fast runoff. There are also examples from other domains, such as climate models (e.g. Liepert and Lo, 2013). Thus, how to show that a model is giving the right results for the right reasons should be a subject for some deeper thought (see for example, Kirchner, 2006).

An interesting possibility that arises from applying more rigorous testing to model applications in hydrology is that all the models tried might be rejected as fit-for-purpose. This invalidation might be for different parameter sets in a single model structure; it might extend to multiple model structures. There are published examples of where all the models tried have been rejected (see most recently the case of the SWAT model in Hollaway et al., 2018, in an application to a small UK catchment). As noted earlier such model rejection is really a good outcome, in that it requires either that we do better modelling or find better data, or that we find some other way of making decisions within an adaptive management framework. We should, note, however, that even where more rigorous invalidation testing is carried out, the results will always be conditional on the information that is to hand now. The future remains epistemically uncertain, and the possibility of future surprise remains. That should not, however, be a reason for relaxing the testing. It should still be considered as poor practice to relax rejection criteria just because a decision needs to be made. That may not result in a good decision if the model is not fit-for-purpose or if the decision is sensitive to the uncertainty in model predictions.

Improving process representations in hydrological models

The concept of being able to reject models as hypotheses has an important implication; that we might be able to learn from the nature of the rejection to refine the representation of hydrological processes and systems where this is shown to be necessary. In this context model rejection is a good outcome. It is the starting point for where creativity of analysis and thought is required for doing better in the future.

It is already possible however to make some suggestions as to what such innovations might look like, particularly if we want process representations that will satisfy the needs to predict both flow and transport within a consistent framework. This assumes a greater importance when we start to accept the limitations of gradient based continuum approaches such as the Buckingham-Richards equation (which I have argued need to be reconsidered since Beven, 1989). Such a framework is required to consider both velocities (in predicting conservative transport) and celerities (in predicting flows). Since celerities are generally different and faster than velocities, it follows that any process representation should be length scale dependent, i.e. different scales of spatial discretisation might require different parameter values. The difference between velocities and celerities will also be state dependent, suggesting that at any scale the hysteresis on the storage-flux response will change with system state. This has been shown numerically using the Multiple Interacting Pathways (MIPs) model by Davies and Beven (2015).

The MIPs model allows velocity distributions to be specified as part of a random particle representation of all the water in the flow domain. Celerities follow from the filling and emptying of storage in the system. It is a computationally expensive modelling strategy and therefore has to date been restricted to small scale applications. While there is still much to explore in the interaction between scale of discretisation, time step, velocity distributions and transition probabilities it does have the type of consistent framework that might be valuable in future. Zehe and Jackisch (2016, Jackisch and Zehe, 2018) have taken a somewhat similar approach including a more explicit consideration of the effects of capillarity. Such approaches might be one way of approaching a theory of scale dependent process representations for both flow and transport.

I have argued before (e.g. Beven, 2006b, 2012a) that there is already a useful framework within which new process representations might be embedded. This is the Representative Elementary Watershed (REW) framework (see, for example, Reggiani et al. 2000; Reggiani and Schellekens, 2003). This sets out a framework of mass, energy and momentum equations that is common for any spatial discretisation. However, those balance equations need closure, I.e. a way of defining the flux terms of mass, energy and momentum at the boundaries of each discrete element, together with how those fluxes depend on the internal states of the system. I believe that this will lead to closure schemes based on hysteretic relationships between element storages and boundary fluxes. A move in this direction would, of course, be greatly enhanced by the availability of the relevant storages or fluxes at the element scale and it may be (again) that real progress will await the availability of new measurement techniques. What we should not do, however, is to continue to ignore the implications of the difference between velocities and celerities and the scale dependent and hysteretic nature of hydrological responses at the element scale.

It is perhaps worth pointing out that the asymmetry of the unit hydrograph or linear transfer functions derived at catchment scales, is a representation of hysteresis in the storage-flow relationship. But as a linear model, it relies on a way of processing the inputs to represent the effects of nonlinearity and antecedent conditions in predicting the catchment response at a wider range of conditions. I could speculate that if input, storage and output data were available for discrete elements of the landscape (or arbitrary REWs) then a transfer function modelling framework such as the Data-Based Mechanistic approach developed by Peter Young (e.g. 1998; Young and Beven, 1994) would be a suitable way of deriving closure schemes at the required scale. The parameters of such a model would then be quite different to those we use today: the time constants for the linear transfer function and some coefficients for nonlinear processing the input sequence. Given additional tracer data, it might also be possible to derive a consistent set of concepts relating parameters for both flow and transport within such a framework (e.g. Harman, 2019). The emphasis is, again, on making the right type of data available, initially at research locations so that we can learn about how to produce closure schemes that might be applicable more widely.

But I could also speculate that rather than accepting a limitation to the linear transfer function or unit hydrograph, with its constant time distribution of contributions of effective rainfall to the hydrograph, perhaps there will be other ways of analysing such data that might more explicitly reflect antecedent states and input intensities at the required scale of discretisation. There are methods of developing hysteretic functions that have been applied to hydrological systems (e.g. O'Kane and Flynn, 2007; Appelbe et al., 2009) but these also have some rather strong assumptions. Given recent developments in data mining techniques, might this be a way of deriving the *forms* of functions that would be applicable more widely, that would suggest quite different process representations than those being used today?

Hydrological Models of Everywhere

The other advance that is certainly going to have a major impact on modelling practice is the much more widespread availability of spatial predictions of hydrological models on the internet. I first suggested a Models of Everywhere concept more than a decade ago (Beven, 2007; Beven and Alcock, 2012) but it is only relatively recently that this has become computationally easier to implement and computer scientists have become more interested in the problem of producing facilitating software (e.g. Blair et al. 2018).

What is critical to this Models of Everywhere concept is that the predictions are sufficiently fine resolution that local stakeholders can relate to them directly. The concept is therefore quite different to providing the global "hyperresolution" simulations presented, for example, by Wood et al. (2011). Hyperresolution in their sense is of the order of 1km (see Bierkens et al., 2015) and while there may be some variables that local stakeholders can relate to at that scale, there will also be a great deal of hyperresolution ignorance about what parameters and variables at that scale might mean (see, for example, the discussion in Beven et al. 2015). There is a movement to finer resolution, continental scale simulations, such as the HydroBlocks of Chaney et al. (2016) which is based on Dynamic Topmodel. At much finer scales, such as the 2m scale used in producing the UK pluvial flooding maps, the ability of people with local knowledge to provide feedback on the model outputs is much more

direct. In this case modelling becomes much more of a learning process, driven by the feedback about where the model predictions are demonstrably wrong. It is a learning process about *places* that starts to reflect the uniqueness of places in terms of both learning about appropriate effective parameter values and learning about appropriate process representations (see Beven, 2000). The possibility of local feedback on the acceptability of model simulations will change the nature of the modelling process in fundamental ways. While we might start with general model structures that are applied to places as in the past, what we need are methods of learning about places from the availability of local data, effective ways of obtaining new data for different purposes, and making use of local (perhaps qualitative) knowledge and expertise (see, for example the study of Landström et al., 2011).

There is an interesting issue in the question of data assimilation in applications of Models of Everywhere. Clearly, we would wish to use all the useful information available to test models locally and to ensure that we get the right results for the right reasons. This might include whatever quantitative data might be available but might also be a matter of learning how to use "soft" data in model evaluations (see, for example, Seibert and McDonnell, 2002; Fenicia et al., 2008; Winsemius et al., 2009). We would also like to re-evaluate models as more data are made available. But, as noted earlier, there have been cases where data assimilation is used simply to compensate for model deficiencies by updating model states so as to get a better predicted outcome. If the purpose of modelling is real time forecasting into the near future, then that might be acceptable or even advisable. Where the purpose is for simulation and assessing the impacts of future change then we should be very wary of compensating for important model deficiencies. For Models of Everywhere we might want to do both forecasting and simulation, in which case it will be important to learn from the process of data assimilation for forecasting in improving the model formulation for simulation. There have been few studies (to my knowledge) that have done so (but see the learning from nonstationarity of Westra et al, 2014, as an example of the type of analysis that might lead to model modifications). More generally, forecasters have been satisfied with using data assimilation to get better forecasts, simulation modellers have been satisfied with using calibration to either find an optimal model or constrain the associated uncertainty. Perhaps we can do better, or at least be a little more thoughtful in applying models. The feedback from users once Models of Everywhere visualisations are more widely available may force us to do so.

Conclusions

I have written about these issues in many past papers (including Beven, 2016b) but this has been a useful opportunity to bring the strands of thought about the future of hydrological modelling in one place. I do think that hydrology remains a field of inexact science that is still greatly constrained by observational limitations and it would be really good to see the community make a real effort to decide on what its priorities should be and then move to commission what is needed (as has happened for example with the SWOT satellite). The process might be long but the benefits to the science would be great, including for testing models as hypotheses, developing new process representations and constraining predictive uncertainties.

The role of Models of Everywhere in improving modelling capability will also make for an

interesting future. What new techniques for learning about places and for learning from clear errors in representing the response of places will need to be developed? And how can new types of knowledge be used to constrain uncertainties? What should the learning framework for both quantitative and qualitative information look like, including the issue of distinguishing information from disinformation. This are issues that are relevant to a wider range of research areas than hydrology which is just one of many inexact environmental sciences (Beven, 2002, 2019).

There is a particularly interesting aspect of uncertainty for the modeller in this context. A realistic assessment of uncertainty in predicting how places respond will mean that the modeller is much less likely to be obviously wrong in those predictions. This is clearly a good thing (at least from a modeller's point of view) but should not preclude an effort being made to carry out model testing and find ways of reducing that predictive uncertainty.

As I said in the talk on which this paper is based, I am ending my career with much more uncertainty than when I started as a young PhD student in 1971. But that is a good thing - it means that there is still so much good research to do in the closely linked areas of novel observational methods, closure schemes and model testing, theoretical development and learning about places. In particular, learning about the assessment of epistemic uncertainties will also lead to the development of methods for reducing those uncertainties. The near future could be an exciting time for hydrological research and practice.

Acknowledgements

Over a long career, I have been fortunate to work and collaborate with many excellent hydrological modellers and experimentalists; a number that is really too long to list here. I will just mention the contribution of Dr. Peter Metcalfe who had only recently started work on the Q-NFM project led by Dr Nick Chappell at Lancaster University before his sudden death in a climbing accident. Working with Peter on the problem of modelling distributed natural flood management measures in this project (NERC grant no. NE/R004722/1) was instrumental in my thinking again about how to improve hydrological models. I am also grateful to Jan Seibert and 2 anonymous referees who made some useful suggestions for relevant papers and improvements to the paper and presentation.

References

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P.E. and Rasmussen, J., 1986. An introduction to the European Hydrological System—Systeme Hydrologique Europeen, “SHE”, 1: History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, 87(1-2): 45-59.

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'connell, P. E., & Rasmussen, J., 1986. An introduction to the European Hydrological System—Systeme Hydrologique Europeen, “SHE”, 2: Structure of a physically-based, distributed modelling system. *Journal of Hydrology*, 87(1-2): 61-77.

Andréassian, V., Lerat, J., Loumagne, C., Mathevet, T., Michel, C., Oudin, L. and Perrin, C., 2007. What is really undermining hydrologic science today?. *Hydrological Processes*, *21*(20): 2819-2822.

Appelbe, B., Flynn, D., Mcnamara, H., Philip, O.K., Pimenov, A., Pokrovskii, A., Rachinskii, D. and Zhezherun, A., 2009. Rate-independent hysteresis in terrestrial hydrology. *IEEE Control Systems*, *29*(1): 44-69.

Baroni, G., Scheiffele, L.M., Schrön, M., Ingwersen, J. and Oswald, S.E., 2018. Uncertainty, sensitivity and improvements in soil moisture estimation with cosmic-ray neutron sensing. *Journal of Hydrology*, 564: 873-887.

Bathurst, J.C., 1986. Physically-based distributed modelling of an upland catchment using the Systeme Hydrologique Europeen. *Journal of Hydrology*, 87(1-2): 79-102.

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. *Journal of Hydrology*, *287*(1-4): 74-94.

Beven, K.J., 1975. A deterministic spatially distributed model of catchment hydrology, PhD thesis, University of East Anglia, Norwich.

Beven, K.J. 1989, Changing ideas in hydrology: the case of physically based models. *Journal of Hydrology*, 105: 157-172.

Beven, K.J. (1993), Prophecy, reality and uncertainty in distributed hydrological modelling, *Advances in Water Resources*, 16, 41-51.

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

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

Beven, K J, 2002, Towards a coherent philosophy for environmental modelling, *Proc. Roy. Soc. Lond. A,* 458, 2465-2484.

Beven, K J, 2006a, A manifesto for the equifinality thesis, *Journal of Hydrology*, 320: 18-36.

Beven, K J, 2006b, On undermining the science?, *Hydrological Processes*, 20: 3141-3146.

Beven, K J, 2006c, The Holy Grail of Scientific Hydrology: $Q_t = H(\underset{\leftarrow\leftarrow}{S}\underset{\leftarrow\leftarrow}{R})A$ as closure, *Hydrology and Earth Systems Science,* 10: 609-618.

Beven, K J, 2007, Working towards integrated environmental models of everywhere: uncertainty, data, and modelling as a learning process. *Hydrology and Earth System Science,* 11(1): 460-467.

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

Beven, K J, 2009, *Environmental Modelling – An Uncertain Future*, Routledge: London

Beven, K J, 2012a, *Rainfall-Runoff Modelling – The Primer,* 2nd edition, Wiley-Blackwell: Chichester

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

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

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

Beven, K J, 2018a, On hypothesis testing in hydrology: why falsification of models is still a really good idea, *WIRES Water*, DOI: 10.1002/wat2.1278.

Beven, K J, 2018b, A Century of Denial: Preferential and Non-Equilibrium Water Flow in Soils, 1864 – 1984, *Vadoze Zone Journal*, in press.

Beven, K. J., 2019, Towards a new paradigm for testing models as hypotheses in the inexact sciences, Proc. Royal Soc. London A, submitted

Beven, K. J. and Alcock, R., 2012, Modelling everything everywhere: a new approach to decision making for water management under uncertainty*, Freshwater Biology,* 56: 124-132, doi:10.1111/j.1365-2427.2011.02592.x

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

Beven, K.J., Callen, J.L., 1979, 'HYDRODAT: A system of FORTRAN computer programs for the preparation and analysis of hydrological data from charts. British Geomorphological Research Group, *Technical Bulletin 23*.

Beven, K J, Cloke, H., Pappenberger, F, Lamb, R and Hunter, N, 2015. Hyperresolution information and hyperresolution ignorance in modelling the hydrology of the land surface. SCIENCE CHINA Earth Sciences, 58 (1): 25-35.

Beven, K J and Freer, J, 2001, A Dynamic TOPMODEL, *Hydrol. Process.,*15(10), 1993-2011.

Beven, K.J., Germann, P. 1982, Macropores and water flow in soils, *Water Resources Research*, 18(5), 1311-1325.

Beven, K. J. and Germann, P. F., 2013, Macropores and water flow in soils revisited, *Water Resources Research,* 49(6): 3071-3092, DOI: 10.1002/wrcr.20156

Beven, K.J., Kirkby, M.J., 1979, A physically-based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24(1): 43-69.

Beven, K.J., Kirkby, M.J., Schofield, N., Tagg, A., 1984, 'Testing a physically-based flood forecasting model (TOPMODEL) for three UK catchments, *Journal of Hydrology*, 69: 119-143.

Beven, K. J. and Lane, S., 2019, Invalidation of models and fitness-for-purpose: a rejectionist approach, in: Beisbart, C. & Saam, N. J. (eds.), Computer Simulation Validation - Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives, Cham: Springer, to appear 2019

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

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

Beven, K J, Smith, P J, and Freer, J, 2008, So just why would a modeller choose to be incoherent?. *Journal of Hydrology*, 354,15-32.

Beven, K., Smith, P. J., and Wood, A., 2011, On the colour and spin of epistemic error (and what we might do about it), *Hydrol. Earth Syst. Sci.,*. 15, 3123-3133, doi: 10.5194/hess-15-3123-2011.

Beven, K., and P. Young, 2013, A guide to good practice in modeling semantics for authors and referees, *Water Resources Research*, 49(8): 5092-5098 DOI: 10.1002/wrcr.20393.

Bierkens, M.F., Bell, V.A., Burek, P., Chaney, N., Condon, L.E., David, C.H., de Roo, A., Döll, P., Drost, N., Famiglietti, J.S. and Flörke, M., 2015. Hyper-resolution global hydrological modelling: what is next? "Everywhere and locally relevant". *Hydrological Processes*, *29*(2), pp.310-320, https://doi.org/10.1002/hyp.10391

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

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

Binley, A.M. and K.J. Beven (1992), Three-dimensional modelling of hillslope hydrology, *Hydrological Processes*, 6, 347-359.

Blair, G. S., K. Beven, R. Lamb, R. Bassett, K. Cauwenberghs, G. Dean, N. Hunter, E. Edwards, V. Nundloll, F. Samreen, W. Simm, R. Towe, 2018, Models of Everywhere Revisited: A Technological Perspective, *Environmental Modelling and Software*, submitted.

Calver, A. and Wood, W.L., 1989. On the discretization and cost-effectiveness of a finite element solution for hillslope subsurface flow. *Journal of Hydrology*, *110*(1-2): 165-179.

Chaney, N.W., Metcalfe, P. and Wood, E.F., 2016. HydroBlocks: a field-scale resolving land surface model for application over continental extents. *Hydrological Processes*, *30*(20): 3543-3559.

Choi, H. T., and K. J. Beven, 2007, Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, *Journal of Hydrology*, *332*(3-4), 316–336, doi: 10.1016/j.jhydrol.2006.07.012.

Coron, L., V. Andréassian, C. Perrin, M. Bourqui, and F. Hendrickx, 2014, On the lack of robustness of hydrologic models regarding water balance simulation: A diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments, *Hydrology and Earth System Sciences,* 18(2): 727–746, doi: 10.5194/hess-18-727-2014.

Crawford, N. H. and Linsley, R.K., 1966, Digital Simulation in Hydrology: Stanfrod Watershed Model IV, Technical Report 39, Department of Civil Engineering, Stanford University, CA.

Dakhlaoui, H., Ruelland, D., Tramblay, Y., Bargaoui, Z., 2017. Evaluating robustness of conceptual rainfall-runoff models under climate variability in Northern Tunisia*. Journal of Hydrol*ogy, 550, 201–217

Davies, J and Beven, K J, 2015, Hysteresis and scale in catchment storage, flow, and transport, *Hydrological Processes.* 29(16): 3604-3615, DOI: 10.1002/hyp.10511.

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

Ewen, J. and Parkin, G., 1996. Validation of catchment models for predicting land-use and climate change impacts. 1. Method. *Journal of Hydrology*, *175*(1-4): 583-594.

Ewen, J., Parkin, G. and O'Connell, P.E., 2000. SHETRAN: distributed river basin flow and transport modeling system. *ASCE Journal of Hydrologic Engineering*, *5*(3): 250-258.

Evans, J. G., H.C. Ward, J. R. Blake, E. J. Hewitt, R. Morrison, M. Fry, L. A. Ball, L.C. Doughty, J. W. Libre, O.E. Hitt, D. Rylett, R.J. Ellis, A.C. Warwick, M. Brooks, M.A.Parkes, G.M.H. Wright, A.C. Singer, D.B. Boorman, A . Jenkins, 2016, Soil water content in southern England derived from a cosmic-ray soil moisture observing system – COSMOS-UK: Soil water content in southern England – COSMOS-UK, *Hydrological Processes*, DOI: 10.1002/hyp.10929

Fenicia F. , McDonnell J.J. , and Savenije, H.H.G., 2008, Learning from model improvement: On the contribution of complementary data to process understanding. *Water Resources Research,* 44: W06419.

Fowler, K.J., Peel, M.C., Western, A.W., Zhang, L. and Peterson, T.J., 2016. Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models. *Water Resources Research*, *52*(3), pp.1820-1846. https://doi.org/10.1002/2015WR018068

Fowler, K., G. Coxon, J. Freer, M. Peel, T. Wagener, A. Western, R. Woods and L. Zhang, 2018, Simulating runoff under changing climatic conditions: a framework for model improvement, *Water Resources Research*, doi: 10.1029/2018WR023989

Freer, J., K. Beven, and N. Peters, 2003, Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure, in *Calibration of Watershed Models*, edited by Q. Duan, H. Gupta, S. Sorooshian, A. Rousseau, and R. Turcotte, pp. 69–87, doi: 10.1029/WS006p0069.

Freer, J, McDonnell, J, Beven, K J, Brammer, D, Burns, D, Hooper, R P and Kendal, C, 1997, Topographic controls on subsurface stormflow at the hillslope scale for two hydrologically distinct small catchments, *Hydrol. Process.*, 11(9), 1347-1352.

Freeze, R.A. and Harlan, R. L., 1969, Blueprint for a physically-based, digitally-simulated, hydrologic response model, *J. Hydrology,* 9:237-258.

Graham, D. N. & Butts, M. B. 2005 Flexible integrated watershed modelling with MIKE SHE. In: Singh, V. P. & Frevert, D. K. (eds) *Watershed Models*. CRC Press, Boca Raton, Florida,pp. 245-272.

Hall, J., O'Connell, E. and Ewen, J., 2007. On not undermining the science: Coherence, validation and expertise. Discussion of Invited Commentary by Keith Beven. *Hydrological Processes: An International Journal*, *21*(7), pp.985-988.

Harman, C., 2019, Age-ranked storage-discharge relations - a unified description of spatially-lumped flow and water age in hydrologic systems, *Water Resources Research*, submitted

Harrigan, S., Murphy, C., Hall, J., Wilby, R.L. and Sweeney, J. 2014. Attribution of detected changes in streamflow using multiple working hypotheses. *Hydrology and Earth System Sciences*, 18, 1935-1952.

Hollaway, M.J., Beven, K.J., Benskin, C.McW.H., Collins, A.L., Evans, R., Falloon, P.D., Forber, K.J., Hiscock, K.M., Kahana, R., Macleod, C.J.A., Ockenden, M.C., Villamizar, M.L., Wearing, C., Withers, P.J.A., Zhou, J.G., Haygarth, P.M., 2018a, Evaluating a processed based water quality model on a UK headwater catchment: what can we learn from a 'limits of acceptability' uncertainty framework?, *J. Hydrology.* DOI: 10.1016/j.jhydrol.2018.01.063

Hollaway M.J., Beven K.J., Benskin, C.McW.H., Collins, A.L., Evans, R., Falloon, P.D., Forber, K.J., Hiscock, K.M., Kahana, R., Macleod, C.J.A., Ockenden, M.C., Villamizar, M.L., Wearing, C., Withers, P.J.A., Zhou, J.G., Barber, N. J. and Haygarth, P.M. 2018b, A method for uncertainty constraint of catchment discharge and phosphorus load estimates. *Hydrological Processes*. 32:2779- 2787. DOI: 10.1002/hyp.13217

Jackisch, C. and Zehe, E., 2018. Ecohydrological particle model based on representative domains. *Hydrology and Earth System Sciences*, *22*(7): 3639-3662.

Jencso, K. G., McGlynn, B. L., Gooseff, M. N., Wondzell, S. M., Bencala, K. E., & Marshall, L. A. (2009). Hydrologic connectivity between landscapes and streams: Transferring reach-and plot-scale understanding to the catchment scale. *Water Resources Research*, 45, W04428. doi.org/10.1029/2008WR007225

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., and Westerberg, I. K.: Disinformative data in large-scale hydrological modelling, *Hydrol. Earth Syst. Sci.*, 17, 2845-2857, https://doi.org/10.5194/hess-17-2845-2013, 2013.

Kiang, J.E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I.K., Belleville, A., Sevrez, D., Sikorska, A.E., Petersen-Øverleir, A. and Reitan, T., 2018. A Comparison of Methods for Streamflow Uncertainty Estimation. *Water Resources Research*: doi.org/10.1029/2018WR022708

Kirchner, J.W., 2006. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3). DOI: 10.1029/2005WR004362

Klaus, J. and Jackson, C.R., 2018. Interflow Is Not Binary: A Continuous Shallow Perched Layer Does Not Imply Continuous Connectivity. *Water Resources Research*. doi.org/10.1029/2018WR022920

Klemes, V ., 1986, Operational testing of hydrological simulation models, *Hydrol. Sci. J.*, 31 (1), 13-24, 1986.

Konikow, L.F. and Bredehoeft, J.D., 1992. Ground-water models cannot be validated. *Advances in Water Resources*, *15*(1): 75-83.

Landström, C., S. J. Whatmore, S. N. Lane, N. Odoni, N. Ward, and S. Bradley, 2011. Coproducing flood risk knowledge: redistributing expertise in critical 'participatory modelling'. *Environment and Planning A,* 43 (7): 1617–1633.

Leedal, D T, J. Neal, K. Beven, P. Young and P. Bates, 2010, Visualization approaches for communicating real-time flood forecasting level and inundation information*, J. Flood Risk Management,* 3: 140-150

Liepert, B.G. and Lo, F. 2013. CMIP5 update of 'Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models'. *Environmental Research Letters*, 8, 029401.

Liu, Y., Freer, J., Beven, K. and Matgen, P., 2009. Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology*, *367*(1-2): 93-103.

McGuire, K. J., & McDonnell, J. J. (2010). Hydrological connectivity of hillslopes and streams: Characteristic time scales and nonlinearities. *Water Resources Research*, 46, W10543. DOI: 10.1029/2010WR009341

McMillan, H.K. and Westerberg, I.K., 2015. Rating curve estimation under epistemic uncertainty. *Hydrological Processes*, *29*(7): 1873-1882.

Merz, R., Parajka, J. and Blöschl, G., 2011. Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research*, *47*(2). DOI:10.1029/2010WR009505

O'Kane, J.P. and Flynn, D., 2007. Thresholds, switches and hysteresis in hydrology from the pedon to the catchment scale: A non-linear systems theory. *Hydrology and Earth System Sciences*, *11*(1): 443-459, https://doi.org/10.5194/hess-11-443-2007

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

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. *Journal of Hydrology*, *175*(1-4): 595-613.

Pappenberger, F, Matgen, P, Beven, K J, Henry J-B, Pfister, L and de Fraipont, P, 2006, Influence of uncertain boundary conditions and model structure on flood inundation predictions*, Advances in Water Resources,* 29(10): 1430-1449, doi:10.1016/j.advwatres.2005.11.012

Peel, M.C. and Blöschl, G., 2011. Hydrological modelling in a changing world. *Progress in Physical Geography*, *35*(2): 249-261.

Refsgaard, J. C., and Knudsen, J., 1996, Operational validation and intercomparison of different types of hydrological models*, Water Resources Research*, 32 (7): 2189-2202.

Refsgaard, J.C. and Storm, B., 1990. Construction, calibration and validation of hydrological models. In *Distributed Hydrological Modelling* (pp. 41-54). Springer, Dordrecht.

Reggiani, P., Sivapalan, M. and Hassanizadeh, S.M., 2000. Conservation equations governing hillslope responses: Exploring the physical basis of water balance. *Water Resources Research*, *36*(7): 1845-1863.

Reggiani, P. and Schellekens, J., 2003. Modelling of hydrological responses: the representative elementary watershed approach as an alternative blueprint for watershed modelling. *Hydrological Processes*, *17*(18): 3785-3789.

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

Romanowicz, R, Young, P C and Beven, K J, 2006, Data assimilation and adaptive forecasting of water levels in the River Severn catchment, UK, *Water Resources Research,* 42, W06407, doi:10.1029/2005WR004373

Seibert, J., 2003. Reliability of model predictions outside calibration conditions. *Nordic Hydrology,* 34, 477-492, doi: 10.2166/nh.2003.028.

Seibert, J. and McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, 38(11), pp.23-1.

Todini, E. and Mantovan, P. 2007. Comment on: 'On undermining the science?' by Keith Beven. *Hydrological Processes* 21(12): 1633–1638.

Wallis, S.G., Young, P.C., Beven, K.J. (1989), Experimental investigation of the aggregated dead zone model for longitudinal solute transport in stream channels, *Proc. Inst. Civ. Eng.*, Part 2, 87, 1-22.

Westerberg, I.K., Wagener, T., Coxon, G., McMillan, H.K., Castellarin, A., Montanari, A. and Freer, J., 2016. Uncertainty in hydrological signatures for gauged and ungauged catchments. *Water Resources Research*, *52*(3): 1847-1865.

Westra, S., M. Thyer, M. Leonard, D. Kavetski, and M. Lambert, 2014, A strategy for diagnosing and interpreting hydrological model nonstationarity, *Water Resources Research*, 50: 1–24, DOI: 10.1002/2013WR014719.

Weyman, D.R., 1970, Throughflow on hillslopes and its relation to the stream hydrograph, *Hydrological Sciences Bulletin*, 15: 25-33.

Weyman, D.R., 1973. Measurements of the downslope flow of water in a soil. *Journal of Hydrology*, *20*(3), pp.267-288.

Wilby, R.L., Clifford, N.J., De Luca, P., Harrigan, S.O., Hillier, J.K., Hodgkins, R., Johnson, M.F., Matthews, T.K.R., Murphy, C., Noone, S.J., Parry, S., Prudhomme, C., Rice, S.P., Slater, L.J., Smith, K.A., Wood, P.J. 2017. The "dirty dozen" of freshwater science: Detecting then reconciling hydrological data biases and errors. *WIREs Water*, 4: n/a, e1209. doi:10.1002/wat2.1209.

Winsemius H.C. , Schaefli B. , Montanari A. , and Savenije H.H.G., 2009, On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information . *Water Resources Research,* 45: W12422.

Wood, E.F., Roundy, J.K., Troy, T.J., Van Beek, L.P.H., Bierkens, M.F., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J. and Gochis, D., 2011. Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*(5), W05301, DOI: 10.1029/2010WR010090.

Young, P., 1998. Data-based mechanistic modelling of environmental, ecological, economic and engineering systems. *Environmental Modelling & Software*, *13*(2): 105-122.

Young, P.C. and K.J. Beven (1994), Data-based mechanistic modelling and the rainfall-flow non-linearity, *Environmetrics*, 5: 335-363.

Zehe, E. and Jackisch, C.: A Lagrangian model for soil water dynamics during rainfall-driven conditions, *Hydrol. Earth Syst. Sci.*, 20: 3511–3526, https://doi.org/10.5194/hess-20-3511-2016, 2016.

Zreda, M., W.J. Shuttleworth, X. Zeng, C. Zweck, D. Desilets, T. Franz, R. Rosolem, 2012, COSMOS: the COsmic-ray Soil Moisture Observing System, *Hydrol. Earth Syst. Sci*., 16 : 4079-4099, 10.5194/hess-16-4079-2012