Beven Keith J. (Orcid ID: 0000-0001-7465-3934)
Page Trevor John Charles (Orcid ID: 0000-0002-1684-6049)
Hankin Barry (Orcid ID: 0000-0001-7315-3321)

On (in)validating environmental models. 2. Implementation of a Turing-like Test to modelling hydrological processes

Keith Beven[1], Stuart Lane[2], Trevor Page[1], Ann Kretzschmar[1], Barry Hankin[1,3], Paul Smith[1,4], and Nick Chappell[1]

[1] Lancaster Environment Centre, Lancaster University, Lancaster UK

[2] Institute of Earth Surface Dynamics, University of Lausanne, Switzerland

[3] JBA Consulting, Warrington, UK

[4] Waternumbers, Lancaster, UK

Corresponding author: Keith Beven, k.beven@lancaster.ac.uk

Keywords: hypothesis testing; epistemic uncertainties; limits of acceptability; hydrologic model; hydraulic models

Funding: NERC Grant No. NE/R004722/1; Fondation Herbette, Lausanne

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/hyp.14703

On (in)validating environmental models. 2. Implementation of a Turing-like Test to modelling hydrological processes

Abstract

Part 1 of this study discussed the concept of using a form of Turing-like Test for model evaluation, together with eight principles for implementing such an approach. In this part, the framing of fitness-for-purpose as a Turing-like Test is discussed, together with an example application of trying to assess whether a rainfall-runoff model might be an adequate representation of the discharge response in a catchment for predicting future natural flood management scenarios. It is shown that the variation between event runoff coefficients in the record can be used to create some limits of acceptability that implicitly take some account of the epistemic uncertainties arising from lack of knowledge about errors in rainfall and discharge observations. In the case study it is demonstrated that the model used cannot be validated in this way across all the range of observed discharges, but that behavioural models can be found for the peak flows that are the subject of interest in the application. Thinking in terms of the Turing-like Test focusses attention on the critical observations needed to test whether streamflow is being produced in the right way so that a model is considered as fit-for-purpose in predicting the impacts of future change scenarios. As is the case for uncertainty estimation in general, it is argued that the assumptions made in setting behavioural limits of acceptability should be stated explicitly to leave an audit trail in any application that can be reviewed by users of the model outputs.

"Foxes beat hedgehogs. And the foxes didn't just win by acting like chickens, playing it safe with 60% and 70% forecasts where hedgehogs boldly went with 90% and 100%. Foxes beat hedgehogs on both calibration and resolution. Foxes had real foresight. Hedgehogs didn't."

Phillip Tetlock and Dan Gardner, 2015

"What are the grounds for credibility of a given hydrological simulation model?"

Vit Klemeš, 1986

A Turing-like Test as a framing of fitness-for-purpose

In Part 1 of this study we have set out the reasons for dealing with model evaluation as a form of pro-active Turing-like Test and set out eight principles that might underlie such a test. We have argued that because of lack of knowledge and the consequent inevitable epistemic uncertainties in the modelling process, this will necessarily involve some qualitative judgments as to what constitutes acceptable model performance, particularly in judging whether a model is getting the right results for the right reasons and might therefore be more robust in providing evidence about future performance under changed conditions. We have suggested that modellers need to be more like Tetlock's agile foxes than short-sighted hedgehogs in their approach, but that there is a need to be explicit about the assumptions that are made in model evaluation, providing an audit trail that can be reviewed by potential users of the modelling results.

Of course, in thinking about what constitutes fit-for-purpose and being robust about predicting the future we must take account of both our fundamental hydrological knowledge, and any evidence available from past observations that can be made available. What new observations should be made will also depend on the importance of the purpose and the resources allocated (Kelleher et al., 2017; Tauro et al., 2018; Beven et al., 2020). A model that is only based on learning methods or calibrated against past observations of one output variable may provide a more likely representation of those data (e.g. Nearing et al., 2021) but that does not necessarily make it a more likely model of the system being described, especially if is required to be used to provide evidence about future performance under changed future conditions (e.g. Ewen and Parkin, 1996; Brigode et al., 2013; Beven and Lane, 2019; Seibert et al., 2019; Hankin et al., 2019). Thus, there may be a difference between a mathematically best fit to past streamflow data (which may be overfitted in compensating for errors and uncertainties in forcing and evaluation data, see Beven, 2020) and the most robust or reliable representation of the system for the purpose at hand. Determining

robustness will then necessarily also depend on a choice of appropriate evaluation or likelihood measures in the face of epistemic errors in model structures and inputs (e.g. Graeff et al., 2009; Andréassian et al., 2012; Coxon et al., 2014; Beven, 2019b for hydrological examples). Such a choice is inherently difficult since robustness will require that both process representations and epistemic errors are in some sense "similar" in both evaluation and prediction periods, even if only for the input data. This similarity assumption has been the basis of the GLUE methodology in the past but, for good epistemic reasons, cannot necessarily be guaranteed. We should still expect future surprises and be prepared to learn from them.

So in what way should hydrological and hydraulic models be defined as being fit-for-purpose? In the literature, discussion of what should constitute fitness-for-purpose in hydrological and other environmental models has been somewhat conspicuous by its absence. It is inevitable that any decision about whether a model be considered adequate will be subjective (see Part 1). In any application, with different model structures and parameter sets, there will be a range of performances from the best available to unreasonable or implausible. The primary concern, however, should not only be where in that range we might decide that performance is adequate, but rather whether the best model available is itself adequate, or can be rejected as not sufficiently robust or fit-for-purpose for the type of predictions required.

Such a decision will depend on the purpose. It was suggested in Part 1 that we might expect to be more demanding for applications where we are testing for scientific understanding (when a model might simply be rejected on the a basis of prior perceptual understanding, e.g. Wagener et al., 2021a,b; Beven and Chappell, 2021), or where the decisions that depend on model predictions might be consequential in terms of major investment costs or risk of significant dis-benefits. We might also expect to be more demanding for model output variables that are directly required for such a decision than for variables that are not explicitly relevant to that decision (though clearly we also want to demonstrate that any model is, as far as possible, getting the 'right results for the right reasons' using all the available information). We might also expect that different researchers, practitioners or end-users of model outcomes might have different attitudes towards what constitutes fit-for-purpose in assessing risk, making their decisions, or formulating policy for catchment management.

A search of the literature concerned with model fitness-for-purpose in decision and policy-making suggests that it often depends on purely qualitative assessments (Boaz and Ashby, 2003; Fisher et al., 2009; Rijke et al., 2012) and therefore the attitude of the person making the assessment. Even where assessments are based on summary statistics, decisions will necessarily have a subjective element (e.g. Wagener et al., 2001, 2021a; Ritter and Muñoz-Carpena, 2013; Harmel et al., 2014). Such attitudes might involve past experience about the pedigree or utility of a particular model or type of model (e.g. Beven, 2001; Refsgaard et al., 2006, 2010; Francesconi et al., 2016) or an implicit recognition that some critical variables are expected to be just much more difficult to predict than others.

This then, of course, suggests that validation or hypothesis testing of models as fit-for-purpose is not only a matter of testing the statistical uncertainties of past performance, but depends on more subtle interrelationships between modellers and the users of the model outputs that are intrinsically linked to issues of future research funding or specific policy needs rather than simply aiming to get the right result for the right reason (see Lane, 2012; Part 1). As Lane (2012) argues: if the modeller is as important as the model; if modellers or modelling communities hold particular visions of what the right modelling strategy, and right model, are; and notably if modelling is being undertaken in a framework that emphasises producing results rather than slowing down reasoning, invalidation may be a markedly awkward goal. Following Tetlock's (2006) notions of the fox and the hedgehog, how can we become more like foxes rather than persisting as short-sighted hedgehogs?

The data for model (in)validation

The degree of wrongness of a model will necessarily depend on the quality of the observations used to drive and test it (e.g. Oudin et al., 2006; Kuczera et al., 2010; Krueger et al., 2010; Beven and Smith, 2015; Engeland et al., 2016). Those data will be subject to both random aleatory uncertainties and epistemic uncertainties (e.g. Beven and Westerberg, 2011; McMillan et al., 2012a, 2017, 2022; Moges et al., 2020). However, quality assurance of all available data needs to be done carefully. Ideally, data should be evaluated for consistency and errors ***prior*** to running any model, since we would not wish simply to exclude all periods of data that a model does not fit (Beven and Smith, 2015; Beven, 2019b). The degree of

uncertainty in the inputs will depend on the size of the catchment, how the inputs are measured, and may vary with type of event (e.g. convective versus frontal precipitation) and whether snow or poorly measured orographic enhancement or rain-shadow depletion of precipitation are important. The uncertainties are thus, in principle, expected to be both epistemically unknown and non-stationary in their characteristics, meaning that it is more difficult to generate different realisations of the inputs as a way of assessing the sensitivity of model outputs to these input uncertainties, which might vary from event to event in complex ways. In the case of the upland River Kent catchment (UK) being modelled in the example case study below, the estimates of catchment integrated rainfalls have been based on the interpolation of rainfalls using a form of co-kriging that allows for the effects of storm direction and the patterns of elevation. Most of the observation points for this input variable are, however, at lower elevations. This estimate is expected to better allow for elevation effects than simple linear interpolation between sparsely distributed raingauges but cannot itself be easily validated given the data available. Snow is not important for the period being modelled in this catchment but will create additional epistemic issues where it is a significant input to a catchment.

There have been studies on the impacts of input errors on model calibration (e.g. Kavetski et al., 2006; Oudin et al., 2006; Renard et al., 2010; Balin et al., 2010) and on the impact of rating curve errors (e.g. Liu et al., 2009; Blazkova and Beven, 2009; McMillan et al., 2010; Beven and Westerberg, 2011; Domeneghetti et al, 2013; Sikorska and Renard, 2017; Coxon et al., 2014; Hollaway et al., 2018a). Uncertainities in other data used for model evaluations have also been considered (e.g. snow depths in Blazkova and Beven, 2009; snow cover fraction in Schaefli, 2016 and Teweldebrhan et al., 2018; and geophysical information in Graeff et al., 2009). We also have evidence for this catchment that high wind speeds and local humidity deficits might result in significant local interception losses for some events, even in winter. Rarely have all such sources of epistemic uncertainty been considered. Bayesian inference can do this implicitly (e.g. Huard and Mailhot, 2006; Kavetski et al., 2006; Ajami et al., 2007; Reichert and Mieleitner, 2009; Renard et al., 2010; Balin et al., 2010) but in ways that interact with the model structure and parameter distributions and that cannot lead to model rejection, only larger input multipliers or residual variances that compensate for any model deficiencies.

In proposing a Turing-like test for model evaluation we are also suggesting that observations might indicate model deficiencies be allowed to "speak back" (Stengers, 2013), to change how we think about our underlying perceptions of the problem. In that way it could be argued that real progress is being made. Reacting as a hedgehog makes this difficult (see the discussion of the resistance to change in the history of preferential flow in Beven, 2018b). Discrepancies may be outliers that have to be dismissed, or 'parameterised out' of the model through modifying an auxiliary relation. But, they may also be the observations that force us to dismantle the basis upon which our predictions have been built (see Morton, 1993; Stengers, 2005; Baker, 2017), to perturb the dominant perceptual models and paradigms upon which we are relying. The question is when do discrepancies pass from being outliers to observations that are allowed to speak back and perturb what we think and what we do? That is a part of defining a thoughtful Turing-like Test.

An interesting example comes from the implementation of the mass balance equation in many environmental models. Similar issues can arise in the implementation of energy and momentum balances (see Reggiani et al., 2000, for a hydrological example). Because the boundary conditions for a model domain are subject to epistemic errors (Beven et al., 2011; Khan *et al.*, 2014; Kauffeldt et al., 2013; Fan, 2019; Safeeq et al., 2021) then it is quite possible that the available observational data will not satisfy such balance conditions. Use of such data to calibrate and to test a model for which mass balance is assumed to hold as a basic principle might then feed disinformation into the modelling process and result in bias in the predictions (e.g. Beven and Smith, 2015; Beven, 2019b). This issue has been recognised for a long time. The original hydrological Stanford Watershed Model of Crawford and Linsley (1966) had parameters that allowed the rainfall inputs and evapotranspiration estimates to be modified by a constant factor to help meet the mass balance requirement. The widely-used Sacramento model also included a parameter that allowed inputs to be adjusted (see e.g. Duan et al., 2006). The use of such parameters will also result in a strong parameter interaction in calibration and arose because of the hydrological understanding that either the rainfall estimates or the evapotranspiration estimates might not be accurate.

The use of rainfall multipliers to correct for water balance errors is an early example of the use of 'effective parameters' that are too convenient and too costly to dismiss because they can easily be adjusted to produce an apparently acceptable model prediction. Manning's *n* provides a second example of this in hydraulic modelling. Aronica et al., (1998), Werner et al. (2005), Pappenberger et al. (2007) and others show how allowing both floodplain and channel roughness n values to vary during model calibration provides many combinations of possibly acceptable solutions. Lane (2014) describes how attempts to constrain Manning's *n* in hydraulic models of river flow (such as the Conveyance Estimation System of Bramley, 2004) which invites modellers to enter a range of factors from channel sinuosity to vegetation) have failed because modellers need *n* as the critical parameter that can be used to make a model perform. When it is constrained, it is no longer so effective and it loses its versatility as a means of matching observations in calibration and validation. With its dimensionality ($m^{-1/3}s$), *n* should vary with depth and velocity, although this is rarely the case in practical model applications. Indeed, Manning (1877) in his original paper rejected what is now known as the Manning equation (in part because of its dimensionality issue) so this is another case where a more thoughtful fox-like approach to process representations is still needed. Pappenberger et al. (2005a) also provide an example of how post-flood level surveys with which a hydraulic model might be compared may themselves be subject to epistemic uncertainties. What is needed therefore is a way of assessing what might be expected of a model under a given set of conditions, independent of a particular model structure, i.e. based on allowing for data uncertainty and consistency in terms of any applicable physical principles such as mass and energy balance. Given the epistemic nature of sources of error, this will necessarily require expert input into defining a Turing-like Test.

Who then can be considered as an expert?

There is one theme that underpins the application of the Turing-like Test which is the notion of an 'expert'. The seventh principle in Part 1 of this study noted the need to at least record, if not test, the attitudes of the expert(s) involved in model evaluation. There are two points that need to be made here to emphasise this issue. The first is that the limited (as yet) research undertaken on social attitudes and responses in the field of hydrology and hydraulics has revealed the ways in which experts are conditioned by their personal and institutional

trajectories with respect to the models with which they work (Landström et al., 2011a, 2013; Lane, 2014). That is, careful attention needs to be given to the *a priori* conditioned knowledge that an expert uses to frame their evaluation, whether intended or not. The second is that work has suggested that 'experts' need not simply be those who have had their expertise certified. In relation to flood inundation modelling, expertise has been shown to be much more diffuse than might be imagined and includes those who have experienced flooding in other ways e.g. flood victims (Lane et al., 2011). The evidence that flood victims might provide has traditionally been seen as biased, as it may be bound to their lived experience of flooding which is commonly more than just "where water has gone", as well as a normative desire for a particular solution to be adopted. However, there is no evidence that such victims are any less biased than those modellers who advocate a particular modelling approach or model choice. Indeed, setting different kinds of expertise in juxtaposition may be a means of putting our knowledge to the test, exposing biases which we may not be aware of, and so helping us to reject those kinds of models (or data) in a Turing-like test.

Landström et al (2011b) followed two scientists as they worked with flood victims to develop a new approach to flood modelling for the town of Pickering (UK). They describe how such working caused the scientists to turn away from their normal academic community networks, and associated models and modelling practices, to develop a new modelling approach to reduce flood risk. This co-working questioned a series of hypotheses that the modellers (and the associated government agency) were making (e.g. the standard of defence to which the community wish to be protected, the feasibility of using floodplain storage to reduce flood magnitude in certain events) which in turn resulted in the adoption of a new flood risk management strategy for the town. The point is that application of a Turing-like Test of fitness-for-purpose can allow for the engagement of those traditionally excluded from model validation who may well, in bringing a different kind of knowledge and experience, reduce the substantial power that other vested interests may have in not rejecting a model.

More often, however, the experts have been the modellers themselves. There are some prior examples of this type of Turing-like test in hydrological modelling, mostly from cases of trying to evaluate model simulation runs for ungauged catchments for which no past discharge time series are available. In the blind validation of Parkin et al., (1996) some of the

ensemble of model runs based on the prior estimates of parameter ranges were rejected by the modellers acting as experts as unreasonable for the catchment under study, though the basis for that rejection was not made clear. Wagener and Montanari (2011) also discuss the use of information from hydrological streamflow signatures characteristic of 'similar' gauged catchments that might be used to constrain an initial set of model runs; while Kelleher et al. (2017) consider how this information might be combined with local information. Other suggestions have been made, including a small number of direct discharge measurements (e.g. Seibert and Beven, 2009; Jackisch et al., 2014), particularly if collected during extreme events (Singh and Bardossy, 2012); and sources of "soft" information (Seibert and McDonnell, 2002, 2013; Winsemius et al., 2009). Soft data are more qualitative information or spatially limited data that might be obtained from short field campaigns or post-event survey work to capture local knowledge with limited effort (e.g. the Section 19 post-event flood extent surveys in Environment Agency, 2016). A small number of direct measurements can also be useful in constraining model parameter values directly if these can be made directly scale commensurate, as in Beven et al. (1984).

A Case Study illustrating the concept of the Turing-like Test for the case of a rainfall-runoff model in the prediction of flood mitigation.

In defining a Turing-like Test in the case of rainfall-runoff models, the first thing to consider is the purpose. In the Case Study for this paper we are interested in predicting the impacts of various Natural Flood Management (NFM) measures (such as in-channel and off-line storage, reconnection of flood plain storage; and soil improvement and tree planting) on flood peaks. We therefore require a model that, as far as possible, can reproduce flood peaks within the limitations of the data available in a way consistent with knowledge of catchment processes during floods, and which can later allow for implementing the types of NFM measures to be considered. For the 209 km^2 River Kent catchment at Sedgwick in Cumbria, UK, we have chosen a form of Dynamic Topmodel (Smith and Metcalfe, 2022; Electronic Supplement Section B). At the catchment scale, we have observations of inputs and historical stream discharges but know that there are issues with the rainfall interpolation in this upland catchment, and rating curve extrapolations for the highest flood discharges. We also have

some information about patterns of flood inundation in past events, but no direct observations (e.g. saturated area measurements) of the internal hydrological processes. This is a common situation in hydrological modelling.

To define a Turing-like test for evaluating Dynamic Topmodel in predicting the situation prior to any introduced or natural changes we will make use of the historical rainfall and discharge data available. The test is posed in the form of setting limits of acceptability prior to making model runs. This approach has been used in different types of applications in the past within the Generalised Likelihood Uncertainty Estimation (GLUE) methodology in the form of the support for fuzzy measures (e.g. Franks et al., 1998; Freer et al, 2003, 2004; Blazkova et al., 2002; Blazkova and Beven, 2009; Page et al., 2007; Romanowicz and Beven, 2003; Pappenberger et al., 2007). The approach allows that the limits of acceptability might be chosen differently depending on the purpose, even for the same variable to be simulated.

One way of setting limits of acceptability in a way that reflects epistemic uncertainties in observed rainfall and discharge data, was suggested by Beven (2019b). It is applicable to fast-responding catchments where event runoff coefficients can be calculated based on recession curve extrapolation using a master recession curve. Beven and Smith (2015) and Beven (2019b) show how these can vary markedly for events of similar observed inputs, including events for which the apparent runoff coefficient is greater than 1. This might arise because of errors in the estimation of either the inputs or the outputs. For any event of interest, a distribution of runoff coefficients for similar events in terms of input volume and antecedent flows can then be determined and used to determine limits of acceptability for that event. Note that runoff coefficient as used here refers to estimates of total runoff volume that might have been expected if the next event had not occurred as explained in the Electronic Supplement to this paper. It avoids any arbitrary separation of a "baseflow" component as used in many calculations of runoff coefficients (e.g. Blume et al, 2007).

The method used here allows a database of event runoff coefficients to be built up that can be used to define distributions of potential runoff coefficients for "similar" events. These distributions are then used to define limits of acceptability that reflect the rainfall and discharge epistemic uncertainties implicit in the variation in event runoff coefficients [see

Electronic Supplement Section A]. Figure 1 shows the limits defined in this way for the River Kent catchment for part of the period used in model evaluation. The period includes the highest peak on record, produced by Storm Desmond on December 6th 2015. Note that the limits of acceptability take account of the calculated runoff coefficient for each event, so that the limits can be highly asymmetric around the observed value (and may not even include the observed for events with runoff coefficients much greater than 1). Note also that using the runoff coefficients as multipliers in this way allows for volumetric uncertainties but not for timing errors in the inputs or hydrograph shape. We have therefore relaxed the limits based on runoff coefficients alone to allow a +/- 2 hour timing error with a 15 minute time step (see also Pappenberger and Beven, 2004). The limits of acceptability shown in Figure 1 include this allowance. Figure 2 shows the range of simulations using a version of Dynamic Topmodel for which a wide range of prior parameter distributions have been sampled randomly [see Electronic Supplement Section B]. In this case there has been a sufficient 'run-in period' such that initialisation uncertainties will not be important. Figure 2 shows that the range of model behaviours based on prior estimates of model parameters is potentially rather wide, perhaps wider than a modelling expert would allow to be reasonable.

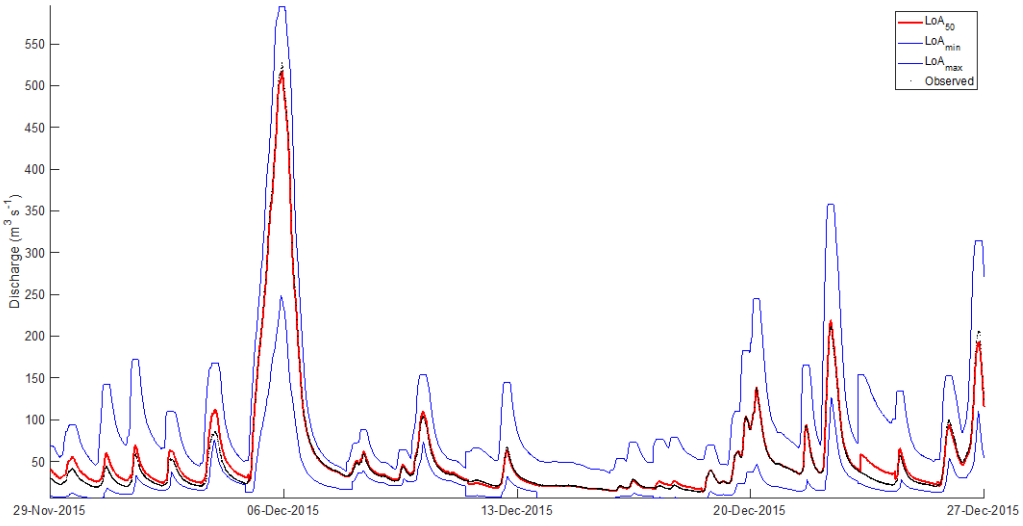

Figure 1. Observed discharge at 15-min resolution (black dots), Upper and Lower Limits of Acceptability based on distribution of runoff coefficient multipliers and +/- 2 hour timing error allowance (blue lines), and LoA 50th percentile (red) for a section of the evaluation period (including Storm Desmond on 6th December).

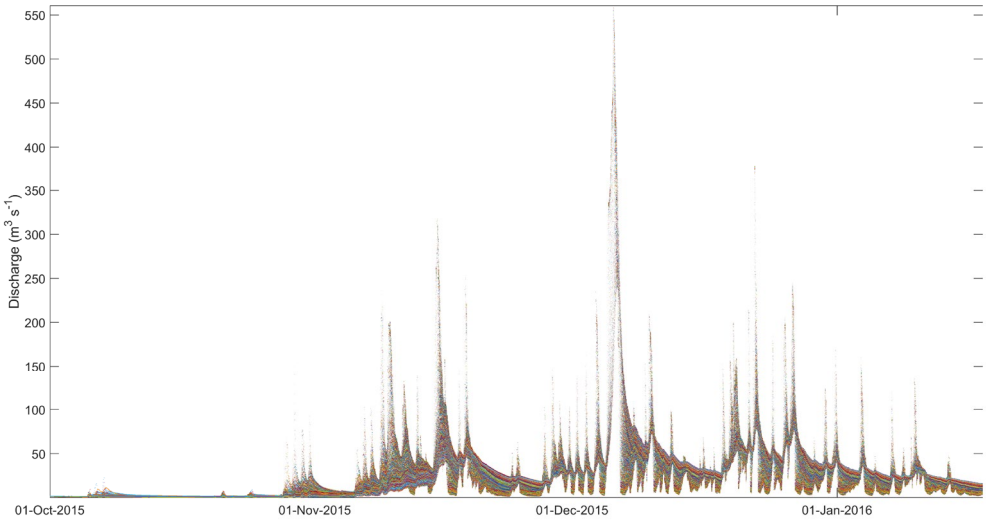

Figure 2. Observed (red) and simulated discharges at 15-min resolution (grey band) for the whole of the evaluation period over all 100,000 model runs.

Analysis of the full set of 100,000 simulations ,some of which are shown in Figure 2, shows that there are **no** model runs that always satisfy the limits of acceptability shown in Figure 1. Figure 3 shows a histogram of the normalised absolute deviations, *NAD*, here defined as:

$$NAD(i,t) = \left|\frac{Q_{sim}(i,t) - LoA_{50}(t)}{LoA(t)_u - LoA_{50}(t)}\right| \text{for } Q_{sim}(i,t) > LoA_{50}(t) \quad [1a]$$

or

$$NAD(i,t) = \left|\frac{Q_{sim}(i,t) - LoA_{50}(t)}{LoA_{50}(t) - LoA(t)_l}\right| \quad \text{for } Q_{sim}(i,t) < LoA_{50}(t) \quad [1b]$$

where *NAD(I,t)* is the normalised absolute deviation for simulation *I* at time *t*; $Q_{obs}(t)$ is the observed discharge at time *t*; $Q_{sim}(i,t)$ is the simulated discharge for the *i*th model run; and $LoA(t)_u$, $LoA_{50}(t)$ and $LoA(t)_l$ are the upper, median and lower limits of acceptability at time *t*. The normalisation is calculated with respect to the median of the distribution of runoff coefficient multipliers to allow for the case where the observed flow is very close to the upper

or lower limits, resulting in very large *NAD* values if the observed flows were used. It must be remembered, however, that those limits are an expression of the expectations about runoff coefficients from past events, so that the $LoA_{50}(t)$ value represents a median estimate of what the model might predict given the epistemic uncertainty in the observations. For the simulated flow to be within the limits of acceptability at all time steps therefore, the maximum *NAD* must be less than 1. Figure 3 shows that this is not the case for any of the 100,000 runs of Dynamic Topmodel.

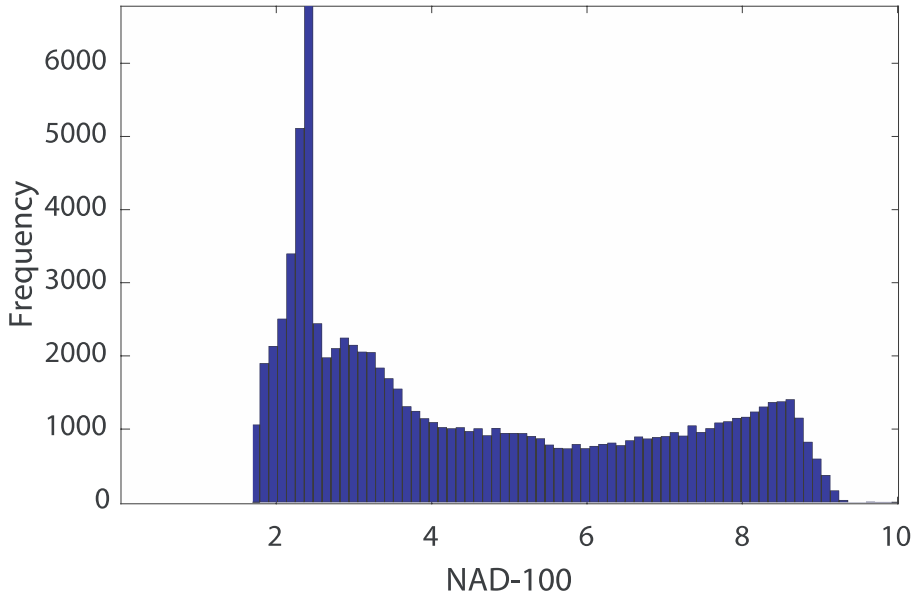

Figure 3. Maximum normalised absolute deviations as defined by equations [1] for all model runs for (100% compliance for all time steps requires values <1)

For this application there are no model runs that have a *NAD* < 1 for all time steps. On that basis, therefore, Dynamic Topmodel could be considered to be invalidated in this application to the Kent catchment. This is despite the implicit treatment of the epistemic uncertainties in the inputs and flow observations that are reflected in the runoff coefficient distributions (with the allowance for timing errors) used to construct the limits of acceptability. It is also despite the fact that the simulation runs show a range of Nash-Sutcliffe Efficiency and Kung-Gupta Efficiency values that extend to over 0.9 (see Figures 4 and 5), that many hydrological modellers would find quite acceptable.

We might consider that some allowance should be made for outliers, by analogy with a one tailed test in statistical practice. If a threshold is imposed at 99% compliance for *NAD*<1, then there are still no simulations that are behavioural. At 95% compliance for *NAD*<1, there is only 1 simulation that might be considered behavioural. At 90% compliance, there are 498 simulations that could be considered behavioural, but this also results in some simulations giving significant overprediction of the highest peaks (even where those have event runoff coefficients estimated from the observed data of greater than 1). This is one of the issues that arise with epistemic rather than aleatory uncertainties: allowing for non-compliance might mean that those non-compliant time steps are those of greatest interest, in this case highest discharge peaks during major flood events. This is why a thoughtful, Turing-like test is necessary.

For the purpose at hand, which is concerned primarily with predicting the peak flows during flood events before and after flood mitigation interventions have been introduced, we therefore concentrate on the simulation of the hydrograph peaks. Checking the simulations of the peaks in rank order (see Table 1) shows that there are model runs that might be considered behavioural in predicting the peaks within the limits of acceptability and within the timing error limits. There are 3249 of the 100,000 runs that can be considered as behavioural for all of the highest 26 peaks in the evaluation period, but none that survive the evaluation on all 40 peaks. Figure 4 shows a histogram of the NSE for these simulations. Interestingly, these do not include the highest NSE simulations, as shown in the 'dotty plot' projections of NSE against individual parameter values in Figure 5. Despite the constraints provided by the limits of acceptability. these dotty plots show that there is broad equifinality across the ranges of most of the parameters, with some constraint on the feasible range of the transmissivity decline parameter *m*.

Table 1. Percentage of the full set of 100,000 simulations that are behavioural in reproducing peak discharges ranked by magnitude (with rank 1 being Storm Desmond, the largest on record). As well as magnitude limits of acceptability, timing limits of +/- 2 hours are imposed.

Peak no.	1	2	3	4	5	6	7	8	9	10
% accepted	84.1	32.9	32.8	20.0	20.0	15.6	10.8	10.5	10.5	8.7
Peak no.	11	12	13	14	15	16	17	18	19	20
% accepted	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5
Peak no.	21	22	23	24	25	26	27	28	29	30
% accepted	8.5	8.5	8.5	8.5	8.5	3.6	0.9	0.9	0.9	0.9
Peak no.	31	32	33	34	35	36	37	38	39	40
% accepted	0.8	0.8	0.8	0.6	0.6	0.6	0.0	0.0	0.0	0.0

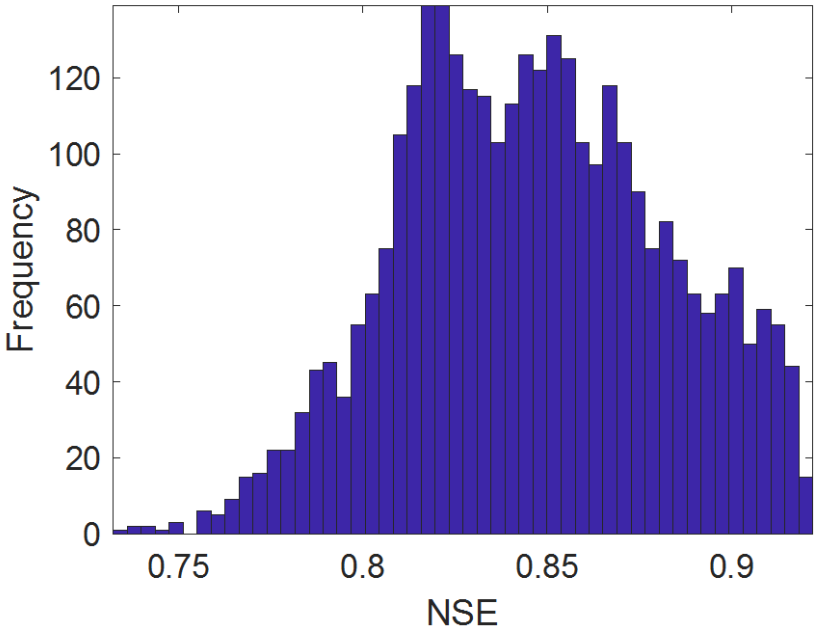

Figure 4. Histogram of Nash-Sutcliffe Efficiency values for the 3249 model simulations that are behavioural on the ranked peak evaluation.

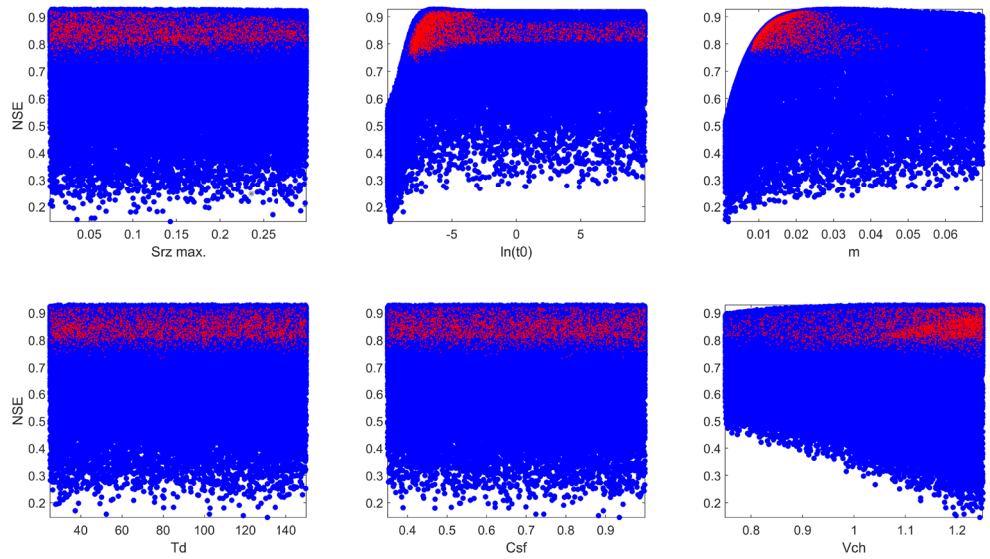

Figure 5. Plots of Nash-Sutcliffe Efficiency against model parameter values as projections of the full NSE response surface onto single parameter dimensions. In blue are all the 100000 simulation runs, in red the 3249 that are behavioural on the ranked peak evaluation. Parameters are defined in Table ES1 of the Electronic Supplement. The root zone initialisation parameter is not shown here given that a run-in period is used before any model evaluations.

The results of using these models to predict part of the full evaluation period, including the Storm Desmond event of December 2015 are shown in Figure 6. This has also been used, in a form of split record test to predict the peak discharges for two other periods that include major flood events causing property damage in the same catchment in 2005 and 2009 (Figures 7 and 8). It can be seen that in both cases the highest peaks are successfully predicted but, as progressively smaller peaks are included in the evaluation, the number of successful simulations decreases (to zero after 11 events in 2005; to zero after 15 events in 2009; see Table 2).

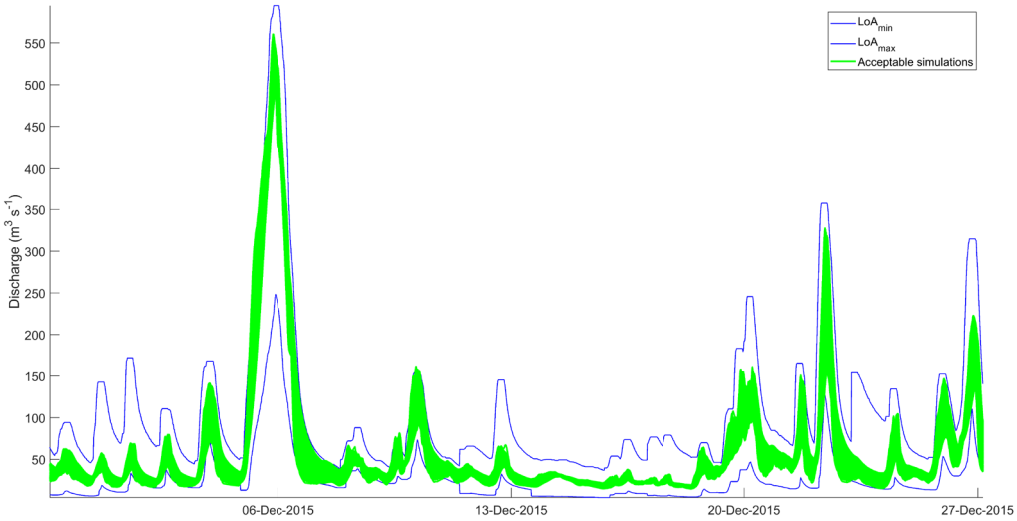

Figure 6. Upper and Lower limits of acceptability (blue lines) and range of the 3249 behavioural simulations after evaluation on the 2015 storm peak discharges

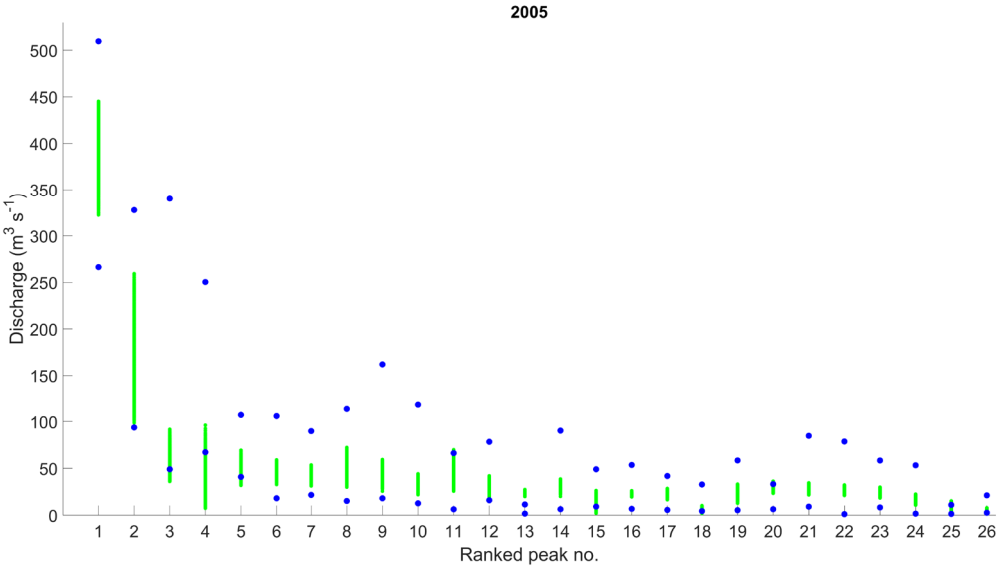

Figure 7: River Kent at Sedgwick (UK). Sequential rejection of 2015 behavioural simulations for ranked peaks in the 2005 evaluation period. Green dots are surviving simulations (with number labelled at each successive storm peak evaluation); blue dots are upper and lower limits of acceptability for peak magnitudes determined in the same way as 2015; crosses are observed peaks; timing limits of ± 2 hours also imposed as for 2015

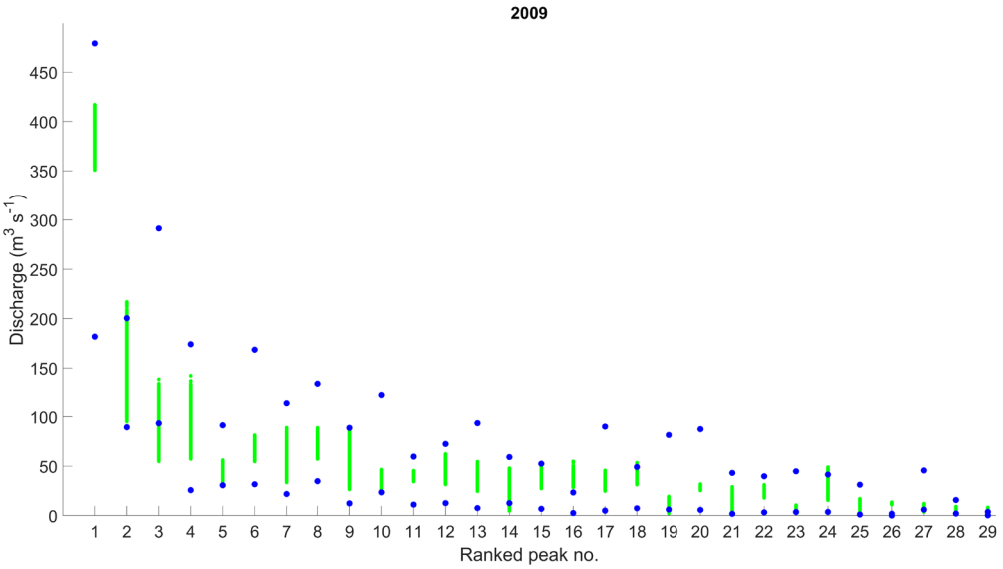

Figure 8: River Kent at Sedgwick (UK). Sequential rejection of 2015 behavioural simulations for ranked peaks in the 2009 evaluation period, River Kent at Sedgwick. Explanation as for Figure 6.

Table 2. Numbers of simulations from the 3249 that are behavioural in 2015, that remain behavioural in reproducing peak discharges ranked by magnitude for the 2005 and 2009 periods. In both periods the largest events were damaging flood events.

Peak no.	1	2	3	4	5	6	7	8	9	10
2005 peaks	3249	3249	1523	118	118	118	118	118	118	118
2009 peaks	3249	3197	1343	1343	1343	1343	1343	1343	1343	1343
Peak no.	11	12	13	14	15	16	17	18	19	20
2005 peaks	118	0	0	0	8	8	8	8	8	8
2009 peaks	1343	1343	1343	1074	1074	0	0	0	0	0

It is clearly then a Turing-like decision as to how to define which set of models might be considered as fit-for-purpose.. For the purposes of the project we have model simulations that survive 26 hydrograph peak predictions for the selected storm period in 2015, and for more than 10 of the major peaks in the test periods within 2005 and 2009. We have no models that survive evaluation by the limits of acceptability at all time steps, and for all hydrograph peaks over all three periods. It is therefore a subjective decision as to whether the success in predicting the larger events is adequate for the prediction of changes in peak flows for the current application focused on the effects of nature-based interventions on flood hydrographs. This will be explored in a future paper.

The relationship between fitness-for-purpose and hydrological understanding

We have shown how for this study, our model fails a limits of acceptability test at all time steps; but for the purpose of predicting peak flow before and after the introduction of flood mitigation measures it might still be useful in a Turing-like test sense. Such model evaluations are, however, just a start towards understanding whether that model might be giving the 'right results for the right reasons' (as discussed also by Kirchner, 2006). It can be accepted as fit-for-purpose only in the sense of reproducing storm peak discharges over the observed period within the span of model realisations as constrained by a minimal set of observations. The evaluation is necessarily limited and conditional, as with any model calibration against discharge data, because discharge is an integrative observation, so that it provides little information on how the discharge is being generated, or on how realistic the input data might be. We thus may not learn that much from a model that reproduces such an integral observable, only that it is conditionally acceptable. We should learn more from cases, such as those cited earlier, where the model is *not* shown to be acceptable.

One possibility in the current context of assessing flood peaks is to consider the patterns of inundation in the Kent catchment produced by the ensemble of discharge predictions at the Sedgwick gauging station of the previous section. Such patterns of inundation predicted for areas at risk of flooding, including the town of Kendal, could be evaluated using within event (e.g. aerial survey or satellite flood extent) or semi-quantitative post-event survey data (maximum water levels, peak timing estimates). To illustrate this, selected models from the ensemble of realisations of Dynamic Topmodel that were acceptable on the peak flow limits have been used in a model cascade (following the approach outlined in Hankin et al. 2019) to drive a 2D HEC-RAS model of the network in providing patterns of lateral inflows to the channel reaches (See Electronic Supplement Section C].

Running the model cascade for different model realisations reveals different modes of spatial behaviour in the predicted inundation depths and extent. Figure 8 highlights locations where the simulated patterns of inundation from two of the model runs diverge, with a difference that would be detectable using remotely sensed images of inundation or cheap level sensors

placed at critical locations in the floodplain. In the absence of satellite imagery for this event this is demonstrated using a surrogate; a spatial image produced by Flood Foresight modelling (https://www.jbaconsulting.com/floodforesight/) which generates real-time event-footprints with 30 m resolution based on interpolation of a scenario-library of flood depth grids from real-time gauge discharge data. Based on the obvious pathway that is present in simulation mc541 (see Figure 8), **all** similar ensemble members could now be rejected, while simulation mc1805 is retained, along with similar simulations. This type of approach to cascading modelling uncertainties using similar modes of behaviour and temporal and spatial constraints was previously used by Pappenberger et al. (2005b).

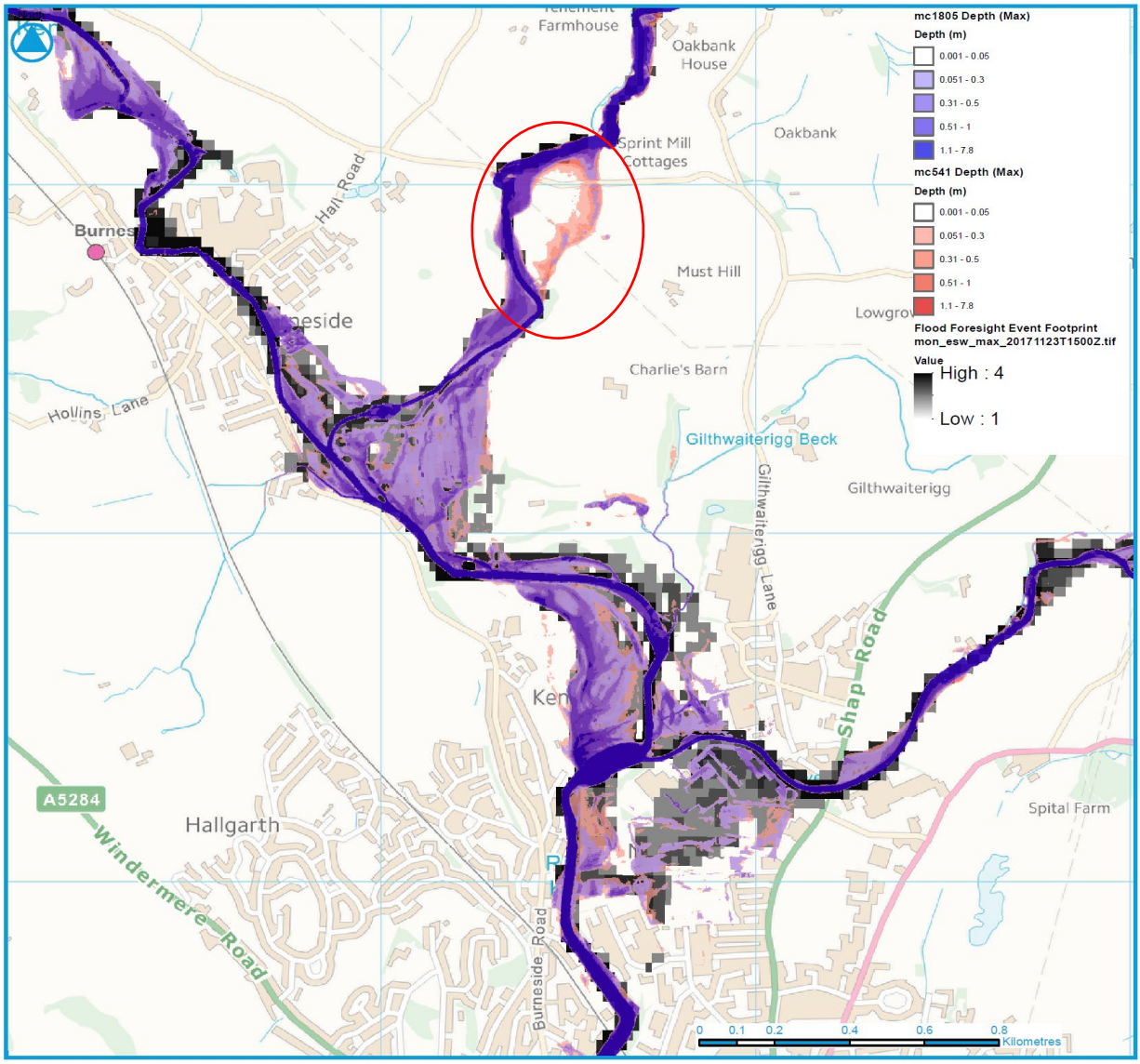

Figure 8: Eliminating key modes of behaviour based on surrogate remotely sensed data (greys to blacks) for HEC-RAS2D inundation predictions in the area north of Kendal, Cumbria, UK. Simulation mc541 (reds) and all

similar ensemble members are now rejected as the highlighted pathway was not observed. Simulation mc1805 and similar ensemble members (blues) are retained.

Surrogate data for inundation (based on another model) are clearly not the same as air-borne or satellite observed patterns of inundation. However, the latter can also be associated with significant epistemic uncertainties (see for example, Bates et al., 1997; Romanowicz and Beven, 2003; Pappenberger et al., 2005a, 2007; Di Baldassarre et al., 2009; Bates, 2012). In both cases, there is then an issue as to how much belief to place in a Turing-like Test that depends on comparing uncertain spatial model outputs to an alternative model (albeit conditioned on observed water depths at some a point of time) or uncertain remote sensing. Similar issues will arise in using local knowledge of such patterns.

There are other purposes for which hydrological models are used that might require different limits of acceptability. We have noted above some past applications that have defined limits based on different summary statistics (flow duration curves, flood frequency curves, snow depth and extent data, and environmental tracer data) rather than simply evaluating some goodness-of-fit index and accepting the best fitting models. This difference is critical, but requires that for a given purpose the assumptions that underlie the definition of the limits in any Turing-like Test are made explicit. This is where Principle 8 in Part 1 of this paper becomes important to provide an audit trail for later critical evaluation.

It is also interesting to pose the question about what critical observations are required to make the Turing-like Test more robust in evaluating the process representations. Because of the limitations of current observational techniques in hydrology (e.g. Beven et al., 2019), this requires further research. Spatial information (such as soil moisture, water table, or saturated areas) may be useful in assessing spatially distributed models but experience using such spatial information has generally had only moderate success, in part because of scale, commensurability and heterogeneity issues (e.g. Beven and Kirkby, 1979; Blazkova et al., 2002; Franks et al., 1998; Lamb et al., 1998; Freer et al., 2004; Clark et al., 2009; Dimitrova-Petrova et al., 2020), but it has also been used to argue for model invalidation (e.g. Barling et al., 1994). Aronica et al. (1998), used a set of fuzzy measures to constrain ensemble

predictions based on local knowledge of "time of arrival" of flood water passing through a narrow part of the Imera basin in Sicily. These data were used to form useful fuzzy constraints (i.e., flat-topped membership functions were used given to represent the uncertainty in the time of arrival provided by local observers). As noted earlier, the support of such fuzzy measures can also be interpreted as applying limits of acceptability for model validity as in the application in this paper.

Remote sensing information is also associated with epistemic uncertainties and is more often used as model input and in data assimilation than in model evaluations or hypothesis testing. There have been studies that have demonstrated the use of actual evapotranspiration and surface soil moisture estimates in model calibration (e.g. Herman et al., 2018; Wambura et al., 2018) but it is important to note that both are constructed variables and again these studies have not considered the epistemic uncertainties and commensurability issues involved. The uncertainties in closing the energy balance to construct actual evapotranspiration rates were demonstrated long ago (e.g. Franks and Beven, 1997), while estimates of surface soil moisture may become more uncertain under wet and dry extremes and provide limited information on the soil water profile (and deeper). Indeed, such surface soil moisture estimates have often rather been used for data assimilation to correct for model deficiencies (see, for example, the recent review of Babaiean et al., 2018). Environmental tracer data might be useful in testing different model structures (e.g. Vaché and McDonnell, 2006; Rinaldo et al., 2011; McMillan et al., 2012b; Harman, 2015; Benettin et al., 2015; Kirchner, 2016) but might require more parameters in a model to make proper use of it and a proper consideration of uncertainties in any flow separations (e.g. Joerin et al. 2003; Kirchner, 2019; Genereux, 2022). This is even more the case for non-conservative water quality data (e.g. Strömqvist et al., 2012; Hollaway et al., 2018b). Practice can clearly be improved but there is little guidance in the literature about what might be the most informative observations in the context of the Turing-like Test for testing models as hypotheses we propose here.

It is interesting to speculate about whether Turing-like tests of fitness-for-purpose could be "institutionalised" for different types of hydrological analyses and purposes in the same way as, say, certain Agencies have institutionalised methods for flood frequency analysis or

methods for the design of flood defences. This might become more important as we move into an era of "open science" with open source models (with multiple variants), databases of model parameters (with various uncertainty and commensurability issues), and open source data and crowd-sourced data (with more or less quality assurance). While those Agencies might be reluctant to define fitness-for-purpose themselves, perhaps it might be sufficient to insist that there is a proper and transparent audit trail that documents explicitly the activities that the modeler has used to gain trust that the model be considered as fit-for-purpose for that purpose (see Principle 8 in Part 1), including tests that might allow for model invalidation.

Conclusions

Having devised the theoretical universal computing machine in 1937, Turing went on to make its physical equivalent, later called a computer, and predicted there would be armies of people programming these devices in the future (Hodges, 2014). This has been the case in hydrology and hydraulics for the last 50 years or so, with the development of many different models and implementations of models, but without much rigorous testing of those models as hypotheses. Now those well-established armies of programmers are joined by the ranks of data-scientists, developing methods of semantic storage and deep learning (including for flood data, see e.g., Towe et al., 2020).

We suspect that the model application presented here is a common example of how model simulations might represent some features of the data available well, but not necessarily all features (see Beven, 2020 for a discussion of deep learning in this respect). The closer we look, including the use of internal state observations with all the commensurability issues of matching observed and model variables, the more likely this is to be the case. We have suggested that model invalidation is a good thing, even if there is a natural resistance amongst modellers (and referees) to reject models as fit for a particular type of purpose. Model invalidation means that we are required to do better, either in developing better model structures or providing better input and evaluation data. It is not easy, however, to decide when a model should be invalidated when we expect that the sources of uncertainty in environmental modelling will often be epistemic rather than simply aleatory in nature. In particular, we wish to avoid rejecting a model that might be useful in prediction because of

uncertainties in the input and evaluation data, but equally we do not want to accept models that contradict secure evidence on the nature of system response.

We have suggested that both modellers and referees should treat model validation as a form of Turing-like Test, but, in doing so, be more explicit about how the uncertainties in different types of observations and their impacts are assessed. Ideally, definitions of 'fitness' for a 'purpose' should be set up before a Turing-like Test is applied and themselves subjected to critical expert evaluation, that might include stakeholders from outside the modelling community with relevant local knowledge or experience (see, for example, Lane et al., 2011).

This might be considered as an example of the 'era of pragmatism' in model evaluation (see Ewen et al., 2012) but within which evaluation should allow for the possibility that all models might be invalidated. There are other examples of studies where all the models tried have been rejected (see discussion in Part 1). It should be remembered that while evaluations of past performance provide the only information about future performance - model validation is always conditional. Thus, a real possibility of future surprises remains, particularly when there are significant sources of epistemic uncertainty in predicting future change.

One of the most important aspect of such a test is being explicit about recording the decisions made in framing an analysis and for model invalidation. This allows a proper assessment by others and also facilitates communication with end-users of the model outputs in that the assumptions and audit trail can be discussed. Such explicit recording of assumptions of an analysis has been incorporated into the workflows of the CURE uncertainty estimation toolbox (https://www.lancaster.ac.uk/lec/sites/qnfm/credible). More research is required, however, on the types of observational data that might be used to differentiate models and process representations that are fit-for-purpose for particular purposes from those that are not.

Where this becomes most interesting is when none of the models tried satisfy the Turing-like Test defined for a particular purpose. This could be for a number of reasons. It could be that not enough model realisations have been run to find sets of parameters that would satisfy the plausibility condition, particularly in high dimensional model spaces. Both Zhang

et al. (2008) and Vrugt and Beven (2018) suggest ways of constructing an efficient search for model parameter sets that satisfy certain behavioural constraints. It could be that not enough account has been taken of uncertainties in the data. Even the best models are not immune to the 'garbage in – garbage out' principle. But, if we can show that something needs improving, then that is how science advances (albeit that it might be disappointing for the policy and decision makers who are awaiting the model outcomes to inform their decisions). It is also where the creativity in modelling lies, so we should look at model invalidation not as a failure but as a real opportunity to do better (and perhaps at the same time improve the quality of the observed data needed). That depends, of course, on being able to avoid acting like a Tetlock hedgehog because of our prickly vested interests in existing model structures.

ACKNOWLEDGMENTS

The origins of the ideas in this paper were developed whilst KB was a Herbette Scholar at the University of Lausanne. Further work on the papers has been carried out under the NERC funded Q-NFM project NE/R004722/1, led by Nick Chappell. The Environment Agency (CMBLNC – Cumbria and Lancashire) is thanked for the licenced release of rainfall, streamflow and raw rating data for the Kent catchment for the NERC Q-NFM project (Licence reference CL77737MG). We are grateful to Jim Freer and Gemma Coxon of Bristol University for help with applying the version of Dynamic Topmodel in their DECIPHeR modelling package in initial test runs used in an earlier version of this paper. Thanks are due to the original referees on these papers Thorsten Wagener and Erwin Zehe for comments that led to improvements to the papers.

DATA STATEMENT

Hydrometric data for the Kent catchment may be obtained under license from the UK Environment Agency. Inputs and outputs from the model runs have been archived at Lancaster University and are available on request from the authors.

REFERENCES

Ajami, N.K., Duan, Q. and Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, *43*(1).

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.H., Oudin, L., Mathevet, T., Lerat, J. and Berthet, L., 2012. All that glitters is not gold: the case of calibrating hydrological models. *Hydrological Processes*, *26*(14), 2206-2210.

Aronica, G, Hankin, B.G., Beven, K.J., 1998, Uncertainty and equifinality in calibrating distributed roughness coefficients in a flood propagation model with limited data*, Advances in Water Resources*, 22(4), 349-365.

Babaeian, E., Sadeghi, M., Jones, S.B., Montzka, C., Vereecken, H. and Tuller, M., 2019. Ground, proximal, and satellite remote sensing of soil moisture. *Reviews of Geophysics*, *57*(2), pp.530-616.

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

Balin, D., H. Lee, and M. Rode, 2010, Is point uncertain rainfall likely to have a great impact on distributed complex hydrological modeling?, Water Resour. Res., 46, W11520, doi:10.1029/2009WR007848.

Barling, R.D., Moore, I.D. and Grayson, R.B., 1994. A quasi-dynamic wetness index for characterizing the spatial distribution of zones of surface saturation and soil water content. *Water Resources Research*, *30*(4), 1029-1044.

Bates, P.D., 2012. Integrating remote sensing data with flood inundation models: how far have we got?. *Hydrological processes*, *26*(16), 2515-2521.

Bates, P.D., Horritt, M.S., Smith, C.N. and Mason, D., 1997. Integrating remote sensing observations of flood hydrology and hydraulic modelling. *Hydrological processes*, *11*(14), 1777-1795.

Benettin, P., Kirchner, J.W., Rinaldo, A. and Botter, G., 2015. Modeling chloride transport using travel time distributions at Plynlimon, Wales. *Water Resources Research*, *51*(5), 3259-3276.

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

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

Beven, K J, 2014. BHS Penman Lecture: "Here we have a system in which liquid water is moving; let's just get at the physics of it" (Penman 1965). *Hydrology Research*, 45, 727-736

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*, 17(1): 180153

Beven, K. J., 2019a, Validation and Equifinality, Chapter 34 in: Beisbart, C. & Saam, N. J. (eds.), *Computer Simulation Validation - Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*, Cham: Springer.

Beven, K. J., 2019b, Towards a methodology for testing models as hypotheses in the inexact sciences, Proceedings Royal Society A, 475 (2224), doi: 10.1098/rspa.2018.0862

Beven, K. J., 2020, Deep Learning, Hydrological Processes and the Uniqueness of Place, *Hydrological Processes,* doi: 10.1002/hyp.13805

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, Aspinall, W P, Bates, P D, Borgomeo, E, Goda, K, Hall, J W, Page, T, Phillips, J C, Simpson, M, Smith, P J, Wagener, T and Watson, M, 2018, Epistemic uncertainties and natural hazard risk assessment – Part 2: What should constitute good practice?, *Natural Hazards and Earth System Science*, https://doi.org/10.5194/nhess-18-1-2018, in press

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, *J. Hydrology*, 69, 119-143.

Beven, K. J. and Lane, S., 2019, Invalidation of models and fitness-for-purpose: a rejectionist approach, Chapter 5 in: Beisbart, C. & Saam, N. J. (eds.), *Computer Simulation Validation - Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*, Cham: Springer.

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

Beven, K J and Westerberg, I, 2011, On red herrings and real herrings: disinformation and information in hydrological inference, *Hydrological Processes*, 25, 1676–1680

Beven, K. J., Asadullah, A., Bates, P. D., Blyth, E., Chappell, N.A., Child, S., Cloke, H., Dadson, S., Everard, N., Fowler, H. J., Freer, J., Hannah, D.M., Heppell, C., Holden, J., Lamb, R., Lewis, H., Morgan, G., Parry. L., Wagener, T., 2020, Developing observational methods to drive future hydrological science: can we make a start as a community?, *Hydrological Processes, 34*(3), 868-873, DOI: 10.1002/hyp.13622

Beven, K. J. and Chappell, N. A., 2021, Perceptual perplexity and parameter parsimony, WIRES Water, e1530. https://doi.org/10.1002/wat2.1530

Blair, G.S., Beven, K.J., Lamb, R., Bassett, R., Cauwenberghs, K., Hankin, B., Dean, G., Hunter, N., Edwards, E., Nundloll, V., Samreen, F., Simm, W., Towe, R., 2019, Models of Everywhere Revisited: A Technological Perspective, Environmental Modelling and Software**,** 122, p.104521.

Blazkova, S, Beven, K, Tacheci, P and Kulasova, A, 2002, Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): the death of TOPMODEL?, *Water Resources Research*, 38(11), W01257, 10.1029/2001WR000912

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

Blöschl, G., Sivapalan, M., Savenije, H., Wagener, T. and Viglione, A. eds., 2013. *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press.

Blume, T., Zehe, E. and Bronstert, A., 2007. Rainfall—runoff response, event-based runoff coefficients and hydrograph separation. *Hydrological Sciences Journal*, *52*(5), pp.843-862.

Boaz, A. and Ashby, D., 2003. *Fit for purpose?: assessing research quality for evidence based policy and practice*. London: ESRC UK Centre for Evidence Based Policy and Practice.

Bramley, M. 2004. New tools for flood level estimation – conveyance and afflux estimation systems, http://www.river-conveyance.net/aes/documents/papers/13_Defra_2004b.pdf

Brigode, P., Oudin, L. and Perrin, C., 2013. Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change? *Journal of Hydrology*, *476*, pp.410-425.

CIWEM, 2017. Code of practice for the hydraulic modelling of urban drainage systems. https://www.ciwem.org/assets/pdf/Special%20Interest%20Groups/Urban%20Drainage%20Group/Code%20of%20Practice%20for%20the%20Hydraulic%20Modelling%20of%20Ur.pdf

Clark, M.P., Rupp, D.E., Woods, R.A., Tromp-van Meerveld, H.J., Peters, N.E. and Freer, J.E., 2009. Consistency between hydrological models and field observations: linking processes at the hillslope scale to hydrological responses at the watershed scale. *Hydrological Processes: An International Journal*, *23*(2), 311-319.

Coxon, G., Freer, J., Wagener, T., Odoni, N.A. and Clark, M., 2014. Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes*, *28*(25), pp.6135-6150.

Crawford, N.H. and Linsley, R.K., 1966. Digital Simulation in Hydrology - Stanford Watershed Model 4, Department of Engineering Technical Report no.39, Stanford University.

Di Baldassarre G, Schumann G, Bates PD. 2009. A technique for the calibration of hydraulic models using uncertain satellite observations of flood extent. *Journal of Hydrology* 367, 276–282.

Dimitrova-Petrova, K., Geris, J., Wilkinson, M.E., Rosolem, R., Verrot, L., Lilly, A. and Soulsby, C., 2020. Opportunities and challenges in using catchment-scale storage estimates from cosmic ray neutron sensors for rainfall-runoff modelling. *Journal of Hydrology*, *586*, p.124878.

Domeneghetti, A., Vorogushyn, S., Castellarin, A., Merz, B. and Brath, A., 2013. Probabilistic flood hazard mapping: effects of uncertain boundary conditions. *Hydrology and Earth System Sciences*, *17*(8), pp.3127-3140.

Duan Q., Schaake, J., Andreassian, V., Franks, S., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T.S., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E.F. 2006. Model Parameter Estimation Experiment (MOPEX): Overview and Summary of the Second and Third Workshop Results. *Journal of Hydrology*, 320(1-2), 3-17.

Engeland, K., Steinsland, I., Johansen, S.S., Petersen-Øverleir, A. and Kolberg, S., 2016. Effects of uncertainties in hydrological modelling. A case study of a mountainous catchment in Southern Norway. *Journal of Hydrology*, *536*, pp.147-160.

Environment Agency and Cumbria County Council, 2016. Kendal Flood Investigation Report Kendal (5th-6th December 2015). https://www.cumbria.gov.uk/eLibrary/Content/Internet/536/6181/42557103755.pdf\

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), pp.583-594.

Ewen, J., O'Connell, E., Bathurst, J., Birkinshaw, S.J., Kilsby, C., Parkin, G. and O'Donnell, G., 2012. Physically-based modelling, uncertainty, and pragmatism–Comment on:'Système Hydrologique Europeén (SHE): review and perspectives after 30 years development in distributed physically-based hydrological modelling'by Jens Christian Refsgaard, Børge Storm and Thomas Clausen. *Hydrology Research*, *43*(6): 945-947.

Fan, Y. (2019). Are catchments leaky? WIREs Water, 6(6), e1386. https://doi.org/10.1002/wat2.1386

Fisher, B., Turner, R.K. and Morling, P., 2009. Defining and classifying ecosystem services for decision making. *Ecological Economics*, *68*(3), 643-653.

Francesconi, W., Srinivasan, R., Pérez-Miñana, E., Willcock, S.P. and Quintero, M., 2016. Using the Soil and Water Assessment Tool (SWAT) to model ecosystem services: A systematic review. *Journal of Hydrology*, 535, pp.625-636.

Franks, S and Beven, K J, 1997, Estimation of evapotranspiration at the landscape scale: a fuzzy disaggregation approach, *Water Resources Research*, 33(12), 2929-2938.

Franks, S W, Gineste, Ph, Beven, K J and Merot, Ph, 1998, On constraining the predictions of a distributed model: the incorporation of fuzzy estimates of saturated areas into the calibration process, *Water Resources Research*, 34, 787-797.

Freer, J. E., K. J. Beven, and N. E. 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. N. Rousseau, and R. Turcotte, AGU Books, Washington, 69-87.

Freer, J, McMillan, H, McDonnell, J J and Beven, K J, 2004, Constraining Dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures, *J. Hydrology.*, 291, 254-277

Genereux, D.P., 2022. Addendum to "Quantifying Uncertainty in Tracer-Based Hydrograph Separations" for Three-Component Mixing Problems. *Water Resources Research*, *58*(2), p.e2022WR031987.

Graeff, T., Zehe, E., Reusser, D., Lück, E., Schröder, B., Wenk, G., John, H. and Bronstert, A., 2009. Process identification through rejection of model structures in a mid-mountainous rural catchment: observations of rainfall–runoff response, geophysical conditions and model inter-comparison. *Hydrological Processes: An International Journal*, *23*(5), pp.702-718.

Hankin, B. Metcalfe, P., Beven, K. and Chappell, N.A. 2019. Integration of hillslope hydrology and 2d hydraulic modelling for natural flood management. *Hydrology Research,* https://doi.org/10.2166/nh.2019.150 .

Harman, C. J. (2015), Time-variable transit time distributions and transport: Theory and application to storage-dependent transport of chloride in a watershed, *Water Resources Research*, doi:10.1002/2014WR015707.

Harmel, R.D., Smith, P.K., Migliaccio, K.W., Chaubey, I., Douglas-Mankin, K.R., Benham, B., Shukla, S., Muñoz-Carpena, R. and Robson, B.J., 2014. Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations. *Environmental Modelling & Software*, 57, 40-51

Herman, M.R., Nejadhashemi, A.P., Abouali, M., Hernandez-Suarez, J.S., Daneshvar, F., Zhang, Z., Anderson, M.C., Sadeghi, A.M., Hain, C.R. and Sharifi, A., 2018. Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. *Journal of Hydrology*, *556*, pp.39-49.

Hodges, 2014, *Alan Turing : The Enigma*. Vintage Press. ISBN 9781784700089

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., 2018, 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.* 558: 607-624. 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., Haygarth, P.M., 2018, 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.* 558, 607-624, Doi: 10.1016/j.jhydrol.2018.01.063

Huard, D., and A. Mailhot (2006), A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc," Water Resour. Res., 42, W07416, doi:10.1029/2005WR0046

Jackisch, C., Zehe, E., Samaniego, L., and Singh, A. K. (2014) An experiment to gauge an ungauged catchment: Rapid data assessment and eco-hydrological modelling in a data-scarce rural catchment, Hydrological Sciences Journal, 59, 2103-2125, doi: 10.1080/02626667.2013.870662, 2014

Joerin, C., K. J. Beven, I. Iorgulescu, A. Musy, 2002, Uncertainty in hydrograph separations based on geochemical mixing models, *J. Hydrology*, 255, 90-106.

Kavetski, D., G. Kuczera, and S. W. Franks, 2006, Bayesian analysis of input uncertainty in hydrological models: 2. Application, Water Resour. Res., 42, W03408, doi:10.1029/2005WR004376.

Kauffeldt, A., S. Halldin, A. Rodhe, C.-Y. Xu, and I. K. Westerberg, 2013. Disinformative data in large-scale hydrological modelling. Hydrology and Earth System Sciences, 17, 2845-2857.

Kelleher et al. 2017. Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. Hydrol. Earth Syst. Sci., 21, 3325–3352, https://doi.org/10.5194/hess-21-3325-2017

Khan, A., Richards, K.S., Parker, G.T., McRobie, A. and Mukhopadhyay, B., 2014. How large is the Upper Indus basin? The pitfals of auto-delineation using DEMs. *Journal of Hydrology*, 509, 442-53.

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), 1–5. doi:10.1029/ 2005WR004362

Kirchner, J. W., 2016, Aggregation in environmental systems – Part 2: Catchment mean transit times and young water fractions under hydrologic nonstationarity, Hydrol. Earth Syst. Sci., 20, 299–328, https://doi.org/10.5194/hess-20-299-2016

Kirchner, J. W., 2019, Quantifying new water fractions and transit time distributions using ensemble hydrograph separation: theory and benchmark tests, Hydrol. Earth Syst. Sci., 23, 303–349, https://doi.org/10.5194/hess-23-303-2019

Klemeš, V., 1986. Operational testing of hydrological simulation models, *Hydrological Sciences Journal,* 31, 13-24.

Kretzschmar, A., Tych, W. and Chappell, N.A., 2014. Reversing hydrology: Estimation of sub-hourly rainfall time-series from streamflow. *Environmental Modelling & Software*, 60,290-301.

Kretzschmar, A., Tych, W., Chappell, N.A. and Beven, K.J., 2016. Reversing hydrology: quantifying the temporal aggregation effect of catchment rainfall estimation using sub-hourly data. *Hydrology Research*, *47*(3), 630-645.

Krueger, T., J. Freer, J. N. Quinton, C. J. A. Macleod, G. S. Bilotta, R. E. Brazier, P. Butler, and P. M. Haygarth 2010), Ensemble evaluation of hydrological model hypotheses, Water Resour. Res., 46, W07516, doi:10.1029/2009WR007845.

Kuczera, G, Renard, B, Thyer, M and Kavetski, D., 2010, There are no hydrological monsters, just models and observations with large uncertainties!, Hydrological Sciences Journal, 55(6), 980—991, DOI: 10.1080/02626667.2010.504677

Lamb, R., K.J. Beven and S. Myrabø, S., 1998, Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model., *Advances in Water Resources*, 22(4), 305-317.

Landström, C., Whatmore, S.J. and Lane, S.N., 2011a. Virtual Engineering: computer simulation modelling for UK flood risk management. Science Studies, 24, 3-22.

Landström, C., Whatmore, S.J., Lane, S.N., Odoni, N., Ward, N. and Bradley, S., 2011b. Coproducing flood risk knowledge: redistributing expertise in critical 'participatory modelling'. Environment and Planning A, 43, 1617-33.

Landström, C., Whatmore, S.J. and Lane, S.N., 2013. Learning through computer model improvisations. Science, Technology and Human Values, 38, 678-700.

Lane, S.N., 2012. Making mathematical models perform in geographical space(s). Chapter 17 in Agnew, J. and Livingstone, D. *Handbook of Geographical Knowledge*. Sage, London.

Lane, S.N., 2014. Acting, predicting and intervening in a socio-hydrological world. *Hydrology and Earth System Sciences*, 18, 927-52.

Lane, S. N., Odoni, N., Landström, C., Whatmore, S. J., Ward, N. and Bradley, S., 2011. Doing flood risk science differently: an experiment in radical scientific method. Transactions of the Institute of British Geographers, 36, 15–36

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), pp.93-103.

McDonnell, J J and Beven, K J, 2014, Debates—The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities, and residence time distributions of the headwater hydrograph, *Water Resour. Res*., 50, doi:10.1002/2013WR015141.

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. *Hydrological Processes*, 24(10), pp.1270-1284.

McMillan, H.; Krueger, T.; Freer, J. 2012a, Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrol. Process*., 26, 4078–4111.

McMillan, H., Tetzlaff, D., Clark, M. and Soulsby, C., 2012b. Do time-variable tracers aid the evaluation of hydrological model structure? A multimodel approach. *Water Resources Research*, *48*(5).

McMillan, H.K.; Seibert, J.; Petersen-Overleir, A.; Lang, M.; White, P.; Snelder, T.; Rutherford, K.; Krueger, T.; Mason, R.; Kiang, J. 2017; How uncertainty analysis of streamflow data can reduce costs and promote robust decisions in water management applications. *Water Resour. Res*., 53, 5220–5228.

McMillan, H., Coxon, G., Sikorska-Senoner, A.E. and Westerberg, I., 2022, Impacts of observational uncertainty on analysis and modelling of hydrological processes: Preface. *Hydrological Processes*, p.e14481.

Moges, E., Demissie, Y., Larsen, L. and Yassin, F. 2020. Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis. *Water*, 13, 28, https://dx.doi.org/10.3390/w13010028

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., and Gupta, H. V., 2021, What Role Does Hydrological Science Play in the Age of Machine Learning?, *Water Resources Research, 57*(3), p.e2020WR028091.

Oreskes, N, 1997, Testing Models of Natural Systems: Can It be Done? in M. L. D. Chiara, K. Doets, D. Mundici

and J. Van Benthem (Eds.), Structures and Norms in Science, pp 207-217, DOI: 10.1007/978-94-017-0538-7_13, Springer, Dordrecht

Oudin, L., Perrin, C., Mathevet, T., Andréassian, V. and Michel, C., 2006. Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. *Journal of Hydrology*, 320(1-2), pp.62-83.

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.

Pappenberger, F and Beven, K J, 2004, Functional Classification and Evaluation of Hydrographs based on Multicomponent Mapping (M^x), *J. River Basin Management*, 2, 89-100.

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

Pappenberger, F., Beven, K.J., Hunter N., Gouweleeuw, B., Bates, P., de Roo, A., Thielen, J., 2005b, Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology and Earth System Science,* 9(4), 381-393.

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

Parkin, G., O'Donnell, G., Ewen, J., Bathurst, J. C., O'Connell, P. E., & 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), 595-613.

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

Refsgaard, J.C., Van der Sluijs, J.P., Brown, J. and Van der Keur, P., 2006. A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources*, *29*(11), pp.1586-1597.

Refsgaard, J.C., Storm, B. and Clausen, T., 2010. Système Hydrologique Européen (SHE): review and perspectives after 30 years development in distributed physically-based hydrological modelling. *Hydrology Research*, *41*(5), pp.355-377.

Reggiani, P., Sivapalan, M., Hassanizadeh, S.M., 2000. Conservation equations gov- erning hillslope responses. *Water. Resour. Res*. 38 (7), 1845e1863.

Reichert, P. and Mieleitner, J., 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resources Research*, *45*(10). W10402, doi:10.1029/2009WR007814

Renard, B., Kavetski, D., Kuczera, G., Thyer, M. and Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5), W05521, doi:10.1029/2009WR008328.

Rijke, J., Brown, R., Zevenbergen, C., Ashley, R., Farrelly, M., Morison, P. and van Herk, S., 2012. Fit-for-purpose governance: a framework to make adaptive governance operational. *Environmental Science & Policy*, 22, 73-84.

Rinaldo, A., Benettin, P., Harman, C.J., Hrachowitz, M., McGuire, K.J., Van Der Velde, Y., Bertuzzo, E. and Botter, G., 2015. Storage selection functions: A coherent framework for quantifying how catchments store and release water and solutes. *Water Resources Research*, *51*(6), pp.4840-4847.

Ritter, A., Muñoz-Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments*. J. Hydrol.* 480, 33e45.

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, 10.1029/2001WR001056

Safeeq, M., Bart, R.R., Pelak, N.F., Singh, C.K., Dralle, D.N., Hartsough, P. and Wagenbrenner, J.W., 2021. How realistic are water-balance closure assumptions? A demonstration from the Southern Sierra Critical Zone Observatory and Kings River Experimental Watersheds. *Hydrological Processes*, *35*(5), p.e14199.

Schaefli, B., 2016. Snow hydrology signatures for model identification within a limits-of-acceptability approach. *Hydrological Processes*, *30*(22), pp.4019-4035.

Seibert, J. and Beven, K., 2009, Gauging the ungauged basin: how many discharge measurements are needed?, *Hydrol. Earth Syst. Sci*., 13, 883-892.

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), 23-1.

Seibert, J. and McDonnell, J.J., 2013. Gauging the ungauged basin: relative value of soft and hard data. *Journal of Hydrologic Engineering*, 20(1), p.A4014004.

Seibert, J., Staudinger, M. and Meerveld, H.J., 2019. Validation and over-parameterization—experiences from hydrological modeling. In *Computer simulation validation* (pp. 811-834). Springer, Cham.

Sikorska, A.E. and Renard, B., 2017. Calibrating a hydrological model in stage space to account for rating curve uncertainties: general framework and key challenges. *Advances in Water Resources*, *105*, pp.51-66.

Singh, S. K. and Bardossy A: Calibration of hydrological models on hydrologically unusual events. Advances in Water Resources 38:81–9, 2012.

Smith, P J and Metcalfe, P, 2022, Package dynatop, https://cran.r-project.org/web/packages/dynatop/index.html (last accessed 10.06.2022)

Stadnyk, T.A. and Holmes, T.L., 2020. On the value of isotope-enabled hydrological model calibration. *Hydrological Sciences Journal*, *65*(9), 1525-1538.

Stengers, I., 2005. The cosmopolitical proposal. In Latour B. and P. Weibel (eds), *Making Things Public*, Cambridge MA, MIT Press, 994-1003

Stengers, I., 2013. *Au temps des catastrophes: résister à la barbarie qui vient*. Editions La Découverte, Paris.

Strömqvist, J., Arheimer, B., Dahné, J., Donnelly, C. and Lindström, G., 2012. Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale. *Hydrological Sciences Journal*, 57(2):.229-247.

Tauro, F., Selker, J., Van De Giesen, N., Abrate, T., Uijlenhoet, R., Porfiri, M., Manfreda, S., Caylor, K., Moramarco, T., Benveniste, J. and Ciraolo, G., 2018. Measurements and Observations in the XXI century (MOXXI): innovation and multi-disciplinarity to sense the hydrological cycle. *Hydrological Sciences Journal*, *63*(2), pp.169-196.

Tetlock, P.E., 2006, *Expert political judgement: how good is it? How can we know?* Princeton University Press, Princeton, New Jersey

Tetlock P and Gardner D, 2015, *Super-forecasting, The Art and Science of Prediction*, Penguin Random House: New York, p.69

Teweldebrhan, A.T., Burkhart, J.F. and Schuler, T.V., 2018. Parameter uncertainty analysis for an operational hydrological model using residual-based and limits of acceptability approaches. *Hydrology and Earth System Sciences*, *22*(9), pp.5021-5039.

Towe, R., Dean, G., Edwards, E., Nundloll, V., Blair, G., Lamb, R., Hankin, B., Manson, S., 2020, Rethinking data-driven decision support in flood risk management for a big data age, *Journal of Flood Risk Management*, *13*(4), p.e12652.

Vaché, K.B. and McDonnell, J.J., 2006. A process-based rejectionist framework for evaluating catchment runoff model structure. *Water Resources Research*, *42*(2).

Vrugt, J.A. and Beven, K.J., 2018. Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM (LOA) algorithm. *Journal of Hydrology*, *559*, pp.954-971.

Wambura, F. J., Dietrich, O. and Lischeid, G.: Improving a distributed hydrological model using evapotranspiration-related boundary conditions as additional constraints in a data-scarce river basin, Hydrol. Process., 32(6), 759–775, doi:10.1002/hyp.11453, 2018.

Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrol. Earth Syst. Sci. 5 (1), 13e26. http://dx.doi.org/10.5194/hess-5-13-2001.

Wagener, T., and Montanari, A. 2011. Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. Water Resour. Res*.,* 47, W06301, doi:10.1029/2010WR009469.

Wagener, T., Gleeson, T., Coxon, G., Hartmann, A., Howden, N., Pianosi, F., Rahman, M., Rosolem, R., Stein, L. and Woods, R., 2021a. On doing hydrology with dragons: Realizing the value of perceptual models and knowledge accumulation. *Wiley Interdisciplinary Reviews: Water*, *8*(6), p.e1550.

Wagener, T., Dadson, S.J., Hannah, D.M., Coxon, G., Beven, K., Bloomfield, J.P., Buytaert, W., Cloke, H., Bates, P., Holden, J. and Parry, L., 2021b. Knowledge gaps in our perceptual model of Great Britain's hydrology. *Hydrological Processes*, *35*(7), p.e14288.

Werner, M. G. F., Hunter, N. M., & Bates, P. D. (2005). Identifiability of distributed floodplain roughness values in flood extent estimation. *Journal of Hydrology*, *314*(1), 139-157.

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*(12), W12422, doi:10.1029/2009WR007706.

Zhang, Z., T. Wagener, P. Reed, and R. Bhushan 2008. Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization. *Water Resour. Res*., 44, W00B04, doi:10.1029/2008WR006833.