Beven Keith J. (Orcid ID: 0000-0001-7465-3934)

On (in)validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for purpose.

Keith Beven[1] and Stuart Lane[2]

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

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

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.14704

On (in)validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for purpose?

Abstract

Model invalidation is a good thing. It means that we are forced to reconsider either model structures or the available data more closely, that is to challenge our fundamental understanding of the problem at hand. 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, epistemic errors in model inputs may well exert a very strong control over how accurate we might expect model predictions to be when compared against evaluation data that might also be subject to epistemic uncertainties. We suggest that both modellers and referees should treat model validation as a form of Turing-like Test, whilst being more explicit about how the uncertainties in observed data and their impacts are assessed. Eight principles in formulating such tests are presented. Being explicit about the decisions made in framing an analysis is one important way to facilitate communication with users of model outputs, especially when it is intended to use a model simulator as a 'model of everywhere' or 'digital twin' of a catchment system. An example application of the concepts is provided in Part 2.

You know the famous line that [philosopher] Isaiah Berlin borrowed from a Greek poet, "The fox knows many things, but the hedgehog knows one big thing"? The better forecasters were like Berlin's foxes: self-critical, eclectic thinkers who were willing to update their beliefs when faced with contrary evidence, were doubtful of grand schemes and were rather modest about their predictive ability. The less successful forecasters were like hedgehogs: They tended to have one big, beautiful idea that they loved to stretch, sometimes to the breaking point. They tended to

be articulate and very persuasive as to why their idea explained everything. The media often love hedgehogs.

Philip E. Tetlock, 2006

1. Background: On model validation

In modelling environment systems such as river catchments we know only too well that we can reproduce the complexity of catchment response with only limited accuracy (see, for example, Beven, 2001, 2002; 2012, 2019a,b). There are very good reasons for this, particularly a lack of full knowledge about the inputs and outputs for the part of the system being represented, even the spatial limits of the system itself (e.g. Khan *et al.*, 2014; Kauffeldt et al., 2015; Beven and Smith, 2015), and a lack of full knowledge about the representation of the (usually nonlinear and interacting) processes that control the responses (e.g. Wagener et al., 2021; Beven and Chappell, 2021). These knowledge, epistemic or deep uncertainties may well be different in the future that we are trying to predict than they were in the past where we might have some observed data that can be used for model evaluation, a problem of inference that also limits the application of purely data-based methods (Beven, 2020; Wagener et al., 2022). They should be distinguished from the aleatory uncertainties that can be treated as random variability and to which the full power of statistical theory can be applied.

It is the epistemic uncertainties that make model validation, in the opinion of some, impossible (see, for example, Stephenson and Freeze, 1974; Oreskes et al., 1994; Oreskes, 1997; Beven, 2013, 2015; Rougier and Beven, 2013). But, following George Box (1979), we might still like to know if some models might be useful or fit-for-purpose in some sense, even when we expect them to be wrong in some way perhaps not yet known. This is particularly important when a model simulator is to be used as a 'model of everywhere' or 'digital twin' of a catchment system to make predictions of how that catchment might behave under possible future conditions. In such cases, we wish to try to get the 'right results for the right reasons' (Kirchner, 2006) and to avoid using models that should not be considered fit-for-

purpose (e.g. Hrachowitz et al., 2014). Thus, model invalidation is an important part of the modelling as a learning process that underlies the 'models of everywhere' concept (e.g. Beven, 2007; Blair et al., 2019).

Our aim is to show how model validation should be really a process of model invalidation, through an extended and pro-active form of a necessarily subjective Turing-like Test. We develop our argument through our primary expertise in hydrological and hydraulic modelling, but we suggest that the discussion has wider relevance to a range of environmental models. In what follows we provide an overview of concepts of model validation; discuss the importance of observed data in model hypothesis testing and invalidation; and discuss the concept and conditionality of fitness-for-purpose. The idea of a Turing-like Test for fitness-for-purpose is then introduced as a way of framing our expectations about model performance, with suggestions for some principles underlying such a concept in application to environmental models. In Part 2 of this study we show how these concepts might be implemented in an illustrative case study and discuss how we might learn from model invalidation to make advances in knowledge and understanding of hydrological processes.

2. Model validation: an overview

The concept of model validation has been a concern in hydrological modelling (and other areas of environmental modelling) for a long time (e.g. Stephenson and Freeze, 1974; Konikow and Bredehoeft, 1992; Oreskes et al., 1994; and the papers in Anderson and Bates, 2001, and Beisbart and Saam, 2019). Vit Klemeš (1986) proposed a hierarchical approach to model validation that would test both the applicability of a model at a site and the transferability of a model to other sites or other climatic conditions. The former involved a split-sample test, the latter differential split-sample, proxy-basin tests.

There have been few reported studies that have gone beyond the simple split-sample test. Perhaps the best known in hydrology is that of Jens-Christian Refsgaard (1997; Refsgaard and Knudsen, 1996) who showed that model calibration could not entirely compensate for differences between sites and sample characteristics (see also Seibert et al., 1996). Ewen and Parkin (1996) also proposed a "blind" validation test for hydrological models when treating

catchments as if ungauged so that no calibration is possible. In their study, a variety of tests were set prior to making model runs using a version of the SHE model. Even allowing for uncertainty in model parameters, not all tests were passed (Parkin et al., 1996; Bathurst et al. 2004). This has not, however, prevented the SHE model from continuing to be widely used (e.g. Refsgaard et al., 2010), and raises the question as to what is being falsified: is it only the particular conditions under which the model was being applied, since the model and its framework as a whole have continued to be used in later applications?

Another widely used model, the Soil Water Assessment Tool (SWAT, Arnold et al., 1998) has, in various forms been applied to hundreds of catchments world wide (Gassman et al., 2007). SWAT is provided with a database of parameter values such that it can be applied in ungauged basins, but it can also be calibrated against historic data. Arnold et al. (2012) discuss a framework for calibration and validation of SWAT. Validation of a model in their sense is when the outputs are "sufficiently accurate" in a split-sample sense; "sufficiently accurate" being purpose specific (see also Van Griensven et al., 2008, as an example of using split-sample tests for fitness-for-purpose of the SWAT model). In a recent application of SWAT, however, Hollaway et al. (2018a) showed that SWAT could not provide sufficiently accurate simulations of both hydrograph and phosphorus outputs from a catchment even after conditioning on a calibration period and allowing for uncertainties in the evaluation data. The model did not appear to be fit-for-purpose in this case.

Similar model failures have been reported for the INCA-P quality model by Dean et al. (2009), for the WEPP erosion model by Brazier et al. (2001) and for TOPMODEL in predicting saturated areas (Beven and Kirkby, 1979; Güntner et al., 1999), flood frequency (Blazkova and Beven, 2009), streamflow in all flow conditions in a small catchment (Choi and Beven, 2007), and the storm to storm variability in stream chloride concentrations (Page et al., 2007). These failures raise again the question as to whether they are due to an incorrect model structure, an incorrect parameterisation, or the suitability of the boundary conditions and auxilliary relations involved in the application. This might be particularly the case when conditions are changing in either the forcing or catchment characteristics. Fowler et al (2016) and Wagener et al. (2022) show how split sample testing represents a challenge for climate change impact studies because the validity of predictions depends on the validity of both the model and its

parameterisation, something that may not hold if tested under changed conditions. It is then necessary to be careful not to confound rejection of a parameterization with rejection of the model structure.

It has also been suggested that some failures might be the result of inadequate sampling of the model space (e.g. Vrugt and Beven, 2018). This might be more likely when the models considered have many parameter values to be estimated by calibration and where we rely, to a greater or lesser extent, on specific assumptions or the theory of a model to interpret data in ways that can be used in calibration or testing. This is referred to as the theory-ladenness of data and reflects the point that when we compare data with a model we are not comparing a theoretically-based model of the world directly with reality, but rather with a data-based model of the world (Oreskes, 1997; Odoni and Lane, 2010): both data and numerical models are representations of the world (see also the statistical theory of reification in Goldstein and Rougier, 2009, and the critique of reification in Briggs, 2014).

Young et al. (1996) proposed and illustrated an alternative modelling approach that is based upon using data-based mechanistic models to identify the dominant modes of behavior in a system based upon analysis of observations; it is these dominant modes that a simulation model needs to be able to reproduce (see also Young, 2013; and the hydrological signatures approach in Hrachowitz et al., 2014). This is nicely illustrated by Nearing et al. (2016b) who have proposed a methodology for testing models relative to a purely data-based approach by considering measures of information in explaining the test data. This allows a model to be assessed in terms of the entropy of the observations of interest relative to the information that can be extracted using only data-based models. Where the data-based methods can be shown to explain more information than a theory-based model then that might be a reason to reject the theoretical model and explore the reasons why the data-based model performs better. Their work has shown that many models fail such a test (though this could be because data-based models can better compensate for physical inconsistencies in the observations, such as those shown in Beven and Westerberg, 2011, Beven and Smith, 2015, and Beven, 2019a).

It is also the case that the supposed validation of a model in one test does not mean that it is applicable generally, or that it will be valid for all possible model applications (see the discussion of SWAT applications above). This is a form of Hume's Problem of Induction (e.g. Beven and Lane, 2019) in the sense that a model that performs well at a site for one set of conditions (in time and space) cannot be expected to perform well for all possible future boundary conditions. The past is not necessarily a guide to future performance, especially when there are epistemic uncertainties about future initial and boundary conditions, a point first made in hydrological modelling by Stephenson and Freeze (1974) (see also Konikow and Bredehoeft, 1992, in respect of groundwater models and Lane et al., 2005, for hydraulic models of channel and floodplain flows). In the case of models with large numbers of parameters, such as SWAT and WEPP, it is likely that even if successful models could be found they might be over-fitted to the calibration data, with a danger of poor performance in prediction when the data uncertainties may be quite different.

There has been significant discussion of validation concepts in other domains of environmental science. A philosophical discussion is provided by Oreskes et al. (1994), following on the papers by Konikow and Bredehoeft (1992) and Anderson and Woessner (1992) in groundwater modelling. Oreskes et al. suggest that validation (implying strength of belief from its Latin root) is preferable to verification (also from the Latin, implying truth) since no model can ever be considered as a true representation of reality; it can only be considered an approximation (although the French title of the Klemeš 1986 paper uses the noun *vérification* to translate operational testing). Verification should only be used in the sense of some proof (preferably using formal mathematical methods that are not generally used in environmental modelling) that a computer code is correct in its implementation. In ecological modelling, Rykiel (1996) uses verification and validation in a similar way and also points to the conditionality of model validation, a qualification of the conditions under which a model might be considered validated based on past performance and method of evaluation.

Thus, verification is a necessary (and perhaps often over-looked) precursor of validation but a verified model in this formal mathematical sense does not necessarily mean that it is valid as fit-for-purpose. Verification requires us to show that our model predictions are internally consistent and reproducible (Hutton et al., 2016, see also Imbert, 2019) even before they are

confronted by and shown to be in some sense consistent with the observations (or not). This type of verification may occur at a number of different levels, the most basic being that decisions regarding the computational solution taken by a modeler (spatial discretization; time steps, convergence criteria, relaxation coefficients etc.) are not inadvertently impacting model predictions (see, for example, Kavetski and Clark, 2010; 2011; Metcalfe et al., 2015; Smith et al., 2021). This may extend to reproducibility between modelers using the same model; or even between models of the same system produced by the same modeler or different modelers. None of these evaluations need make redress to observations, such that a model may be verified but not yet validated and Hutton et al. (2016) suggest that this type of verification needs a community level shift in how we make our code transparent and usable by others.

In environmental hydraulics, there has been much emphasis upon the importance of verification of computer codes as a necessary precursor to validation. Lane and Richards (2001) drew attention to the existence of very different interpretations of the status of a numerical model: according to the American Society of Mechanical Engineers (ASME), the predictions of a model should be taken as verified or correct if its application has followed a series of controls on the numerical accuracy of the associated model solution, that is, it is verified in the sense of Oreskes *et al.* (1994). For the ASME, testing against observational data should not be a substitute for verification nor should such testing be a necessary requirement for labelling a model as acceptable and usable. Lane *et al.* (2005) argue against this position in relation to channel and floodplain flows, noting that these criteria fail to capture the conditionality of model applications that follows from the dependence on boundary conditions, geometry and the need for auxiliary relations to make models solvable (e.g. turbulence closure; wall treatments) that themselves may have a restricted range of applicability. The determination of model acceptability cannot be reduced to simply verification that a computer code is a proper solution of the underlying nonlinear mathematical equations (as already recognised by Stephenson and Freeze in 1974). Similar arguments will apply to models based on the approximate solution of dynamic nonlinear equations in other domains, such as atmospheric and ocean circulation models and subsurface flow models with their own requirements of boundary and auxiliary conditions.

It follows that a model, which is apparently acceptable in one situation, is not necessarily acceptable in another, even after allowing for model calibration or modification of the other auxiliary conditions required to make a model run (Morton, 1993). Key here is recognition that some realisations of a model may be more or less acceptable than others, depending on the ways in which uncertainties in input data/parameters propagate through to model predictions. The aim then might be to set some plausibility criteria or limits of acceptability that help to identify those simulations that are 'acceptable' or 'behavioural'. This is the basis of inferential approaches to parameter estimation, such as with the set-theoretic approaches of Keesman and van Straten (1991), the GLUE methodology (e.g. Beven, 2006, 2009, 2012, 2016; Beven and Binley, 2013; Vrugt and Beven, 2018) or the analogous concepts in Approximate Bayesian Computation (Nott et al., 2012; Vrugt and Sadegh, 2013; Sadegh and Vrugt, 2014). Here, limits of acceptability are used to define plausible model simulations and then the uncertainty that remains is quantified and presented as part of the primary model outputs.

To make progress we suggest that it is necessary to replace the notion of model validation, and all the debates around it, with the complementary notion of model invalidation (see also Beven, 2018; Beven and Lane, 2019). It is an important consideration as to when a model should be considered as ***not*** fit-for-purpose, but this will depend on both the requirements of the purpose, and the quality of observed data available for model evaluation. There are a growing number of studies of the uncertainties associated with hydrological data that could form the starting point for model invalidation (e.g. Harmel et al., 2009, 2014; Khan et al., 2014; Krueger et al., 2010; McMillan et al., 2012; Beven et al., 2011; Westerberg et al., 2011; Beven and Smith, 2015; Coxon et al., 2015; McMillan and Westerberg, 2015; Westerberg and McMillan, 2015; Hollaway et al., 2018b; Kiang et al., 2018; Ehlers et al., 2019; McMillan et al., 2022).

A critical measure that might then be used is whether there is any overlap at all between the distribution of uncertain observations, and the distribution of model predictions. If there is no overlap, then this might be considered as a reason for invalidating that model and finding something better. Even then, however, any particular observation might be considered as an outlier or not very important to the purpose of the application (Harmel et al., 2014). It has

been suggested that the use of a limited set of more extreme events might be of greatest value in model evaluation and testing (e.g. Singh and Bardossy, 2012) since they are more likely to reveal model deficiencies. This then raises the question, however, as to which events are truly informative, which might be disinformative (Beven and Smith, 2015) and how many such outliers should be allowed before a model is invalidated. This latter will be necessarily a subjective decision since it might be difficult to construct robust significance tests for epistemic errors rather than the random errors of statistical theory (see e.g. Frigg et al., 2014). In particular, making an analogy with statistical theory, we could perhaps allow failure on no more than 5% of the observations, but this might not be appropriate when the observational data that are of most interest for an application might make up that 5%, such as when a hydrological model consistently under predicts the largest peaks when used in assessing flood risk but does well in predicting the other 95% of the observational series (see Part 2 of this paper and also Colquhoun, 2014, Briggs, 2014, for critiques of this approach in statistical hypothesis testing).

This then suggests that there should be an expectation that it will not be possible to be entirely objective about model validation when faced with epistemic sources of uncertainty and error. However, good practice should entail being transparent about how uncertainties in model and observations are assessed, what quality measures are used, and how a model invalidation or rejection is to be defined (Beven et al., 2018). It has been proposed before that the subjective assumptions that underlie an analysis should be recorded in a condition tree or audit trail to facilitate communication with users of any model predictions (e.g. Beven and Alcock, 2012; Beven et al., 2014a,b). Here, we suggest extending that concept to include the conditions for model (in)validation. Invalidation implies that it is necessary to do better in some way: either to find a better model structure that is fit-for-purpose, or to better represent the environmental (i.e. boundary) conditions to which the model is being applied, or to improve the evaluation data (Beven and Lane, 2019).

This is then a way of progressing understanding rather than continuing to rely on model predictions that have not been evaluated as fit-for-purpose. If we follow philosopher of science Isabelle Stengers (2005) finding that our model predictions are not fit-for-purpose may be just one way of arriving at statements that do not "*say what is, or what ought to be, but*

[to] provoke thought, a proposal that requires no other verification than the way in which it is able to "slow down" reasoning and create an opportunity to arouse a slightly different awareness of the problems and situations mobilising us" (Stengers, 2005, 994). That is, showing that something is not fit-for-purpose has the potential to advance science through forcing us to search for other approaches, model structures, data etc., rather than to simply accept the model and observed data that we have if it is still associated with large and nonstationary prediction errors even after some form of calibration or conditioning on the observations (see also Thompson and Smith, 2019).

We stress that how fitness-for-purpose is evaluated will depend on the purpose. This will be different for models that might be used for a limited purpose (e.g. flood forecasting, where empirical adequacy might be sufficient), and models that aim to demonstrate a scientific understanding of how a catchment responds to rainfall (where conflicts with qualitative perceptual process understanding might be significant, see, for example, Beven and Chappell, 2021; Wagener et al., 2021, 2022). Defining invalidation criteria might be quite different for different cases. In particular, we should not confuse empirical adequacy with fitness-for-purpose. We should aim to get the 'right results for the right reasons' (Kirchner, 2006; Lane et al., 2011; Lane, 2012; Beven and Chappell, 2021), but this will depend on the purpose. In flood forecasting, for example, a model that predicts water levels rather than discharges and which therefore does not maintain mass balance but which uses adaptive updating to compensate for lack of knowledge of catchment inputs and flood discharges during flood events will generally be more fit-for-purpose than a model constrained by mass balance in predicting peak levels and timing. A model aimed at understanding, however, will require quite different criteria for evaluation (such as getting the patterns and amounts of overland flow correct, or the 'young water fraction' correct where tracer data are available).

3. Model (in)validation and hypothesis testing

Validation has long been considered as a form of hypothesis testing (Overton, 1977; Rykiel, 1996; Sornette et al., 2007; Clark et al., 2011; Baker, 2017; Pfister and Kirchner, 2017; Beven, 2018). Holling (1978) takes a strong Popperian falsification position on this, suggesting that

models can never be validated, they can only be invalidated (see also Beven and Lane, 2019). That is in line with the discussion of this paper, but we note that if multiple models satisfy some basic limits of acceptability (i.e. survive invalidation) they might still be associated with differing strengths of validation, analogous with Popper's varying degrees of *verisimilitude* in theory testing. Methods for hypothesis testing are well developed in statistical theory, based on treating errors as if they were fundamentally aleatory (after allowing for possible structure in the error series such as bias, heteroscedasticity, and covariation). In fact, statistical theory does not ever reject a model as a hypothesis, it will only give it a diminishingly small likelihood. It does, however, provide tools for deciding whether one model has a significantly higher likelihood than another (such as the use of Bayes Ratios and various information criteria). Thus, although a model is not necessarily rejected it might be superseded by another that could be considered as more valid in the sense of higher likelihood or belief.

The question then is how to assess the likelihood when the model uncertainties are primarily epistemic rather than aleatory in nature. Sornette et al. (2007) provide an iterative methodology for updating the degree of belief in a model as more data become available. The approach uses a statistical likelihood but also a subjective weighting parameter to weight the contribution of any new error information in a way that might depend on the framing of a particular application and expectations about the uncertainties associated with particular types of observations. This will be most applicable when the number of observations is small. As the number increases (as with discharge time series in hydrology) the assumption of a statistical error model tends to stretch the likelihood surface unreasonably (see discussion is Beven and Smith, 2015, and Beven, 2016). In such cases, a different approach will be necessary.

This will especially be the case when even the model with the highest likelihood or belief might be associated with significant error. Assessing the structure and parameters of an error model is, in fact, an important part of statistical hypothesis testing in that it will inform the type of likelihood function to be used. Such a framework does not itself, however, allow the user to decide whether any model is good enough or fit-for-purpose. It is often (even if not always) the case in hydrological modelling (and undoubtedly in other environmental domains) that models that appear to do well in calibration, do not do so well when applied to another period

of data or another data set (e.g. Choi and Beven, 2007; Blazkova and Beven, 2009; Coron et al., 2012; Hrachowitz et al., 2014), even when adding a statistical error model. This is a strong indication that either the model structure or the data are subject to epistemic errors, such that the structure of the errors is not stationary and therefore not well represented by a stationary statistical error model.

This raises another issue in model testing and validation. As in statistical hypothesis testing it will be possible to make both Type I and Type II errors, either accepting a model that will provide poor predictions or rejecting a model that would have provided good predictions just because of errors in the calibration data. Any assessment of fitness-for-purpose is therefore necessarily conditional on the decisions and assumptions made in the evaluation. This reflects the conditional nature of any validation exercise, but when carried out in the context of invalidation allows for the interesting case of all models tried being invalidated. The question then, of course, is why?

4. Fitness-for-purpose as a Turing-like Test

The Turing Test is a well-known concept from Artificial Intelligence and here we propose it as a means of addressing the challenge of deciding when a model should be deemed as fit-for-purpose. Turing (1950) proposed that a suitable test for machine intelligence was whether a human interrogator could tell the difference between the responses of another human or a computer program. As Turing posed the question: "*Are there imaginable digital computers which would do well in the imitation game*?". The concept has generated significant discussion in the field of artificial intelligence (e.g. French, 2000; Oppy and Dowe, 2016). A similar challenge has been proposed as a Turing-like test for simulation models used in the environmental sciences in the form: "Can a group of experts tell the difference between a sequence of observations in space and/or time and a model simulation?" [1]. If not, then it might be concluded that a model should be considered as fit-for-purpose. Of course, if we consider all journal referees to be experts of this type, then we would conclude that all

[1] For example by Jonty Rougier, University of Bristol, David Harel (2005), and Tim Palmer (2016) in different areas of computer simulations.

published model outputs should be considered as fit-for-purpose (although in some cases the comparison of model and observations can be obfuscated by, for example, already calibrating or bias-correcting model predictions using a set of past observations).

The concept of a Turing-like Test for environmental models raises some interesting issues. The first recognises that any expert is partially or fully bound by their prior experience and the disciplines within which that experience has developed, a problem frequently identified in relation to environmental policy-making (e.g. Brock and Carpenter, 2007; Klaey et al., 2015). In such an instance, viewing a model as fit-for-purpose is clouded by a particular idea of what makes a model fit-for-purpose that is not simply informed by the model that is under consideration. This is the sense in which reviewers are often chosen because of their knowledge of a particular frame of reference, that is they are the right kind of hedgehog. Past experience suggests that they can become rather prickly when their modelling concepts are questioned. Such a hedgehog will have a strong tendency to assess the model within its own frame of reference, that is the extent to which the model conforms to the paradigm within which it has been built. Many declared model successes are of this type. But, if we return to the definitions of verification and validation, this is not based on any invalidation test but is rather analogous with code to code verification; that is a check of conformity with the basic established principles that guide the modelling strategy, not the extent to which the paradigm itself is right, and the real world is being adequately imitated. Hedgehogs can often be rather short-sighted.

The second is whether a group of experts would be able to consider whether a model is sound or not, ***without*** access to at least some observations from the system under consideration (a catchment and its ***pertinent*** characteristics in the hydrological case). Exercises in simulating the response of catchments treated as ungauged, with access to only soil, geology, topography and land use maps have not proven very successful in the past because of the difficulty of translating such information into values of model parameters (e.g. Refsgaard et al., 1997). Declaring success, of course, will also depend on just what measures we define to evaluate the model as being fit-for-purpose for a particular context: as, for example, in the Ewen and Parkin framework discussed earlier (see Parkin et al., 1996; Bathurst et al., 2004, for applications).

Classically the Turing Test is a qualitative subjective decision (e.g. Palmer, 2016, in the context of climate models), but should ideally have the status of being auditable (Beven et al., 2014b). The subjectivity of such decisions will nearly always be the case for the refereeing of scientific papers that present modelling results in the academic literature, but also in the normal processes of internal and external refereeing in consultancy projects. Referees will sometimes point out when a model or period of data should be rejected, perhaps even after publication (see, for example, Beven, 2009). The basis for such decisions is not, however, always auditable given the information provided.

There are also examples of model inter-comparison projects, where the performance of multiple competing models is assessed e.g. 1D (Environment Agency, 2005) and 2D (Environment Agency, 2010) hydraulic models; the PILPS intercomparison of land surface parameterisations (e.g. Henderson-Sellars et al., 1996); the MOPEX comparison of hydrological models of Duan et al. (2006); the Distributed Model Intercomparison Project, (DMIP, Smith et al., 2013); or the benchmarking of land surface parameterisations of Nearing et al. (2016b). However, many of these models are often set up to test cases that they ought to be capable of reproducing and not cases that are representative of all possible applications of the model. For example, benchmarking of 1D flood inundation models in the United Kingdom (Environment Agency 2005) used 12 tests (e.g. subcritical flows, supercritical flows, triangular channel, Ippen wave, a looped flow divergence and convergence, weirs, side spills etc.) on the basis of either their suitability for analytical (i.e. direct) solution or because they represented important in-channel stream structures. The later study of 2D flood inundation models for the UK Environment Agency did not include a single evaluation against field scale data (there was a dam break case tested against laboratory scale data, and some tests involved field scale topographies but not level or velocity observations, see Environment Agency, 2013). These did not prevent conclusions being drawn about the acceptability of different model codes for different purposes by the experts involved, even though the tests were based essentially only on model-to-model comparisons (i.e. more a verification than validation exercise for real applications).

Other model intercomparison exercises (such as DMIP and PILPS), however, have allowed for validation on an observed data set that was not available to the different modelling teams with and without prior model calibration. The results of both these exercises were instructive. In the DMIP project (Smith et al., 2013), a collection of distributed hydrological models was compared with a lumped conceptual model as a benchmark. Performance of the distributed models based on prior estimates of the parameters was variable relative to the benchmark but was improved in all cases by calibration against observed discharge data (without assessment of data uncertainty). Some of the models performed poorly both in terms of long-term bias, reproduction of snow water equivalents and simulation of flood hydrographs. The conclusion, however, was that the models satisfied the National Weather Service criterion of success (less than 5% bias in predicted discharges on average), low cumulative runoff errors and high values of modelling efficiency. None of the models were explicitly rejected. This represents a Turing-like Test based on the expertise of 32 authors as hydrological modelling experts and, presumably, some additional referees.

To take one example from the series of PILPS inter-comparison experiments with land surface parameterisations, Nijssen et al. (2003) compare 21 different model formulations in an application to a large-scale catchment in Northern Scandinavia. The models gave highly variable results, albeit capturing the "broad patterns of snowmelt and runoff". Some models showed improved performance after calibration on smaller catchment data. One was rejected after failing a "consistency test" (a form of verification based on an internal water balance error of more than 3 mm per year). The greatest differences occurred during the snowmelt period, but the authors noted the difficulty of interpreting the differences because of the complexity of the schemes and dependence on the chosen parameter sets. In this case 26 authors chose not to reject any of the remaining 20 models. This, perhaps, demonstrates a greater allegiance to an epistemic community than to getting the right results for the right reasons. A more recent study in the same region, however, showed that multiple land surface models failed to capture the information content of the observations (as captured in an entropy measure) to the same extent as a purely data-based method (Nearing et al., 2016b). The conclusion of that study was that the physical basis of these models added no information towards explaining the data.

The above discussion implies that application of a Turing-like Test will need to evolve in a much broader and pro-active sense, beyond traditional model-data comparison. Identification of fitness-for-purpose implies a wide spectrum of influences on what is both fitness and purpose. That is, the expert judgement needs to happen 'upstream', and itself be subject to a Turing-like test (e.g. as part of defining the tender documents in commissioning research), before any kind of application of such a test to a model study. It may also be worth considering whether the idea of the Turing-like Test could be made more quantitative, even for cases involving significant epistemic uncertainties based on a formal expert elicitation of what might be expected in terms of model capabilities for a given application. It may also require some reflection upon more than just the end point of the modelling process (when a modeller thinks that they have got the model as good as they can get). Turing (1950) did not deny that computers had to be *made* to imitate, at least until they were able to learn how to imitate themselves. A focus on imitation as an end point, then, overlooks the performative nature of modelling in hydrology (see Lane, 2012) where performance is not only the end point but also the all-too-rarely-documented steps that a modeller goes through to develop trust that their model is providing a correct imitation.

5. Some principles for a Turing-like Test for model plausibility/(in)validation and fitness-for-purpose

From this discussion of a set of issues surrounding the potential for model evaluation by (in)validation, it is possible to extract some principles for discussing and defining an appropriate Turing-like Test as a means of going beyond the types of assessment of model acceptability common in the current literature. These principles are consistent with the Guidelines for Good Practice in dealing with epistemic uncertainties discussed in Beven et al. (2018).

1. Definitions of 'fitness' for the purpose of a project should be agreed amongst the relevant stakeholders, taking expert advice as necessary, before a Turing-like Test is applied. Criteria of fitness may be both quantitative, where observed data are available for model evaluation, and/or qualitative.

2. Models should not be expected to perform better than the observed data on which runs are based and evaluated. A critical evaluation of the data for consistency and uncertainties, independent of the model being studied, should therefore be a pre-requisite for model evaluation.

3. Models should not contradict secure evidence on the nature of system response and still be considered fit-for-purpose.

4. Evaluation should have the aim of getting the right results for the right reasons and not focus only on the need to make a decision.

5. Evaluation should allow for the possibility that all models might be rejected (invalidated) using criteria that allow for input and other observational uncertainties.

6. Past performance provides the only information about future performance, but the results of a Turing-like Test will always be conditional and the problem of induction and possibility of future surprises remain.

7. Achieving objective evaluation of models in the face of epistemic uncertainties can be a challenge: evaluations and evaluators should themselves be evaluated in terms of their fitness-for-purpose.

8. The basis for the definition of "fitness" should be recorded in an audit trail that will allow later review of the process, including the expected sources of uncertainty. This audit trail should include an account of the activities the modeler has used to gain trust in the model they think is fit for purpose.

6. So what should a Turing-like Test for models look like?

These principles do not, however, provide a sufficient basis for an evaluation methodology. In particular, a model might be considered useful even if it explicitly omits some evidence about the nature of system response, to simplify model structure and implementation, while still providing an acceptable match to key observations. Clearly, some types of evidence are

more important than others in informally applying a Turing-like Test for a particular application so that what constitutes acceptable performance will be context dependent. In particular, when the epistemic uncertainty of input data is likely to be significant, it will be very difficult to construct realistic realisations of the input uncertainties, and consequently any expectations of performance in reproducing observational data after the inputs are processed through a nonlinear model structure. This is an important issue in many domains of environmental modelling, but one that is often ignored. Thus, we need to think carefully about setting limits of acceptability in such cases. This requirement for thoughtfulness is the most important aspect of this pro-active Turing-like Test methodology being proposed.

To pass a Turing-like Test, a model must provide outputs that convince some set of relevant "experts" that it is an adequate, acceptable or a behavioural representation of the response of interest for a particular purpose. This judgment should allow for uncertainties in the available data, and for the potential biases of the experts themselves and, in particular, their relative cognitive behaviour (that is their tendency to be fox-leaning or hedgehog-leaning after Tetlock, 2006). It is relative because it depends on the stance of the expert to the wider modelling framework and approach, as well as its detail, that is being assessed. This suggests a way for defining an appropriate Turing-like Test based on methods that have been developed for expert elicitation (see for example O'Hagan et al., 2006; Krueger et al., 2012; Cooke, 2014; Aspinall and Cooke, 2013; Aspinall and Blong, 2015). The Classical Model Structured Expert Judgment (SEJ) method (Cooke, 1991), for example, is based on weighting the judgment of experts based on a preliminary set of questions in the relevant domain of expertise before they give advice on a particular application. This approach has been generally found to give better results than equal weighting. In the Bayesian approach of O'Hagan et al. (2006) distributions associated with the required information can be updated as more information is obtained from experts and model evaluations. Given the potential for epistemic uncertainties in observed data, models, and expert knowledge, however, fuzzy approaches to expert elicitation might also have value (e.g. Krueger et al., 2012). Such approaches can provide a structured framework for setting limits of acceptability in model evaluation.

In recent applications of the Generalised Likelihood Uncertainty Estimation (GLUE) methodology, model evaluations have been based on setting sensible limits of acceptability before viewing the model outputs. The application of constraints in this way has much in common with the blind validation approach of Ewen and Parkin (1996) but can take more explicit consideration of the potential for epistemic uncertainties in the input and evaluation data. Tests might include both data specific to the application catchment, and hydrological signatures for the expected behaviour in different climates, geologies and land uses, an approach that has been used within the Prediction of Ungauged Basins (PUB) framework (e.g. Gupta et al., 2008; Yilmaz et al., 2008; Wagener and Montanari, 2011; Kelleher et al., 2017) and in assessing climate impact models (Wagener et al., 2022). Such an exercise will focus attention on both the potential sources of uncertainty and what we might realistically expect of a model given the data limitations in any modelling project (see Part 2 of this paper). It also consistent with principle 5 above in that it does not preclude model invalidation (e.g. Parkin et al., 1996; Page et al., 2007; Dean et al., 2009; Graeff et al., 2009; Hollaway et al., 2018a).

7. Implementation of the Turing-like Test concept

There is an interesting logical conflict here. A scientific model is only ever conditionally valid, subject to further testing, but needs to provide reliable evidence if the model outputs are to be used in inferences or decision- and policy-making (e.g. Frigg et al, 2014; Roussos et al., 2021). Reliable evidence implies that a simulation model should be right for the right reasons or fit-for-purpose (rather than just demonstrating some success in reproducing how the system has worked in the past – a purely data-based model can usually do that, sometimes better, Young, 2013; Nearing et al., 2016a,b, 2021). In Part II of this study the implementation of these concepts using limits of acceptability will be discussed and an illustrative example application will be developed in the context of hydrological and hydraulic modelling.

ACKNOWLEDGEMENTS

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. Thanks are due to the original

referees on these papers, Thorsten Wagener and Erwin Zehe, for comments that led to improvements.

DATA STATEMENT

This paper does not depend on any observational data.

REFERENCES

Anderson, M.G. and Bates, P.B. (Eds.), 2001, *Model Validation: Perspectives in Hydrological Science*, Wiley: Chichester

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

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment – Part 1: model development. J. Am. Water Resour. Assoc. 34, 73–89

Arnold, J.G., Moriasi, D.N., Gassman, P.W., Abbaspour, K.C., White, M.J., Srinivasan, R., Santhi, C., Harmel, R.D., Van Griensven, A., Van Liew, M.W. and Kannan, N., 2012. SWAT: Model use, calibration, and validation. *Transactions of the ASABE*, *55*, 1491-1508.

Aspinall, W. & Blong, R., 2015. Volcanic Risk Management. Chapter 70 in: *The Encyclopedia of Volcanoes*, Second Edition (eds H. Sigurdsson & four others), Academic Press: 1215-1234.

Aspinall, W.P., & Cooke, R.M., 2013. Expert Elicitation and Judgement. In *Risk and Uncertainty assessment in Natural Hazards*. Rougier, J.C., Sparks R.S.J., Hill, L. (eds). Cambridge University Press, Chapter 4, 64-99.

Baker, V.R., 2017. Debates—Hypothesis testing in hydrology: Pursuing certainty versus pursuing uberty. *Water Resources Research*, *53*(3): 1770-1778.

Bathurst, J. C., Ewen, J., Parkin, G., O'Connell, P. E., & 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, 74-94.

Beisbart, C and Saam N J (Eds.), *Computer Simulation Validation - Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*, Cham: Springer.

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, 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, 2009. Comment on ‘‘Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?’’ by Jasper A. Vrugt, Cajo J. F. ter Braak, Hoshin V. Gupta and Bruce A. Robinson, *Stoch. Environ. Res. Risk Assess.*, 23, 1059–1060

Beven, K J, 2012. Causal models as multiple working hypotheses about environmental processes, *Comptes Rendus Geoscience, Académie de Sciences*, Paris, 344, 77–88

Beven, K J, 2013*.* So how much of your error is epistemic? Lessons from Japan and Italy. *Hydrological Processes,* 27, 1677–1680

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, 2015. What we see now: event-persistence in predicting the responses of hydro-eco-geomorphological systems? *Ecological Modelling,* 298, 4-15

Beven, K J., 2016. EGU Leonardo Lecture: Facets of Hydrology - epistemic error, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrol. Sci. J.* 61, 1652-1665

Beven, K J, 2018. On hypothesis testing in hydrology: why falsification of models is still a really good idea, WIRES Water, 5: e1278.

Beven, K. J., 2019a. 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., 2019b, How to make advances in hydrological modelling, *Hydrology Research*, 50(6): 1481-1494

doi: 10.2166/nh.2019.134

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

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

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

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

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, Lamb, R, Leedal, D T, and Hunter, N, 2014a, Communicating uncertainty in flood risk mapping: a case study, *Int. J. River Basin Management.*, 13, 285-296

Beven, K. J., Leedal, D. T., McCarthy, S., 2014b, Framework for assessing uncertainty in fluvial flood risk mapping, CIRIA report C721, available 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 J. Hydrol. Eng., 20, A4014010

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*, 18, 2769-2783,

Beven, K. J. and Lane, S.N., 2019. Invalidation of models and fitness-for-purpose: a rejectionist approach, Chapter 6 in: Beisbart, C. & Saam, N. J. (eds.), *Computer Simulation Validation - Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*, Cham: Springer, 145-172.

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*, https://doi.org/10.1016/j.envsoft.2019.104521

Blazkova, S., and Beven, K., 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.

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.

Boyle, D. P., Gupta, H. V., and Sorooshian, S., 2000. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, 36, 3663– 3674

Box, G. E. P., 1979, Robustness in the strategy of scientific model building, in Launer, R. L.; Wilkinson, G. N. (Eds.), *Robustness in Statistics*, Academic Press, pp. 201–236.

Brazier, R. E., Beven, K. J., Freer, J. and Rowan, J. S., 2000, Equifinality and uncertainty in physically-based soil erosion models: application of the GLUE methodology to WEPP, the Water Erosion Prediction Project – for sites in the UK and USA, *Earth Surf. Process. Landf.*, 25, 825-845.

Brock, A.A. and Carpenter, S.R., 2007. Panaceas and diversification of environmental policy. *Proceedings of the National Academy of Science*s, 104, 15206-11

Buytaert, W and Beven, K J, 2009, Regionalisation as a learning process, *Water Resour. Res.*, 45, W11419, doi:10.1029/2008WR007359.

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

Clark, M.P., Kavetski, D. and Fenicia, F., 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9), W09301, doi:10.1029/2010WR009827

Colquhoun D. 2014 An investigation of the false discovery rate and the misinterpretation of p-values. *R. Soc. Open Sci.* 1: 140216

Cooke, R M, 1991, *Experts in uncertainty: Opinion and Subjective Probability in Science*, Oxford University Press: Oxford.

Cooke, R M, 2014, Messaging climate change uncertainty, *Nature Climate Change* 5, 8-10

Coron, L., Andréassian, V., Perrin, C., Lerat, J., J.Vaze, J., Bourqui, M. and Hendrickx, F, 2012. Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, *Water Resour. Res.,* 48, W05552

Coxon, G., Freer, J., Westerberg, I., Wagener, T., Woods, R. and Smith, P. 2015. A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research*, 51, A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations

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

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, 3-17.

Ehlers, L.B., Sonnenborg, T.O. and Refsgaard, J.C., 2019. Observational and predictive uncertainties for multiple variables in a spatially distributed hydrological model. *Hydrological Processes*, *33*(5): 833-848.

Environment Agency 2005. Benchmarking of hydraulic river modelling software packages: project Overview. R&D Technical Report: W5-105/TR0 Defra/Environment Agency Flood and Coastal Defence R&D Programme. Bristol: Environment Agency.

Environment Agency, 2013, Benchmarking the latest generation of 2D hydraulic modelling packages, Final Technical Report Project SC120002, Bristol: Environment Agency, ISBN: 978-1-84911-306-9

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

Fowler, K. J.A., Peel, M.C., Watson, A. W., Zhang, L., Peterson, T.J., 2016. Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models. *Water Res. Res.,* 52, 1820-1846.

French, R. M., 2000, The Turing Test: The first 50 years, Trends in Cognitive Sciences, 4(3), 115-122

Frigg, R., Bradley, S., Du, H., Smith, L.A., 2014, Laplace's demon and the adventures of his apprentices, Philosophy of Science, 81(1), 31-59.

Gassman, P.W., Reyes, M.R., Green, C.H., Arnold, J.G., 2007. The soil and water assessment tool: Historical development, applications, and future research directions. *Trans. ASABE*, 5, 1211–1250.

Goldstein, M. and Rougier, J., 2009. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, 139, 1221-1239.

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 Processess*, *23*(5), pp.702-718.

Güntner, A., Uhlenbrook, S., Seibert, J. and Leibundgut, C., 1999. Multi-criterial validation of TOPMODEL in a mountainous catchment. *Hydrological Processes*, *13*(11), pp.1603-1620.

Gupta, H.V., Wagener, T. and Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes*, 22, 3802-3813.

Harel, D., 2005, A Turing-like test for biological modelling, Nature Biotechnology, 23(4): 495-496

Harmel, R.D., Smith, D.R., King, K.W., Slade, R.M., 2009. Estimating storm discharge and water quality data uncertainty: a software tool for monitoring and modeling applications. *Environmental Modelling & Software*. 24, 832e842.

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

Henderson-Sellers, A., McGuffie, K. and Pitman, A.J., 1996. The project for intercomparison of land-surface parametrization schemes (PILPS): 1992 to 1995. *Climate Dynamics*, 12, 849-859.

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.* 558, 607-624.

Hollaway MJ, Beven KJ, 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.

Holling, C.S., 1978. *Adaptive Environmental Assessment and Management.* John Wiley & Sons, New York, NY, 377 pp.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H.H.G. and Gascuel-Odoux, C., 2014. Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. *Water resources research*, *50*(9), pp.7445-7469.

Impert, C., 2019, The multidimensional epistemology of computer simulations: novel issues and the need to avoid the drunkard's search fallacy, Chapter 43 in: Beisbart, C. & Saam, N. J. (eds.), *Computer Simulation Validation - Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*, Cham: Springer, 1029-1055.

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.

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

Kavetski, D. and Clark, M.P., 2011. Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing. *Hydrological Processes*, *25*(4), pp.661-670.

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.

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*, 54, 7149-7176.

Kirchner, J. W., 2006, Getting the right results for the right reasons. Linking measurements, analyses and models to advance the science of hydrology, *Water Resources Research,* 42, W03S04

Klaey, A., Zimmermann, A.B. and Schneider, F., 2015. Rethinking science for sustainable development: Reflexive interaction for a paradigm transformation. *Futures*, 65, 72-85

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

Konikow, L. F. and Bredehoeft, J. D., 1992. Groundwater models cannot be validated?, *Advances in Water Resources*, 15, 75–83.

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

Krueger, T., Page, T., Hubacek, K., Smith, L. and Hiscock, K., 2012. The role of expert opinion in environmental modelling. *Environmental Modelling & Software*, 36, 4-18.

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. and Richards, K.S., 2001. The 'validation' of hydrodynamic models: some critical perspectives. In Bates, P.D. and Anderson, M.G. (editors) *Model validation for hydrological and hydraulic research*, Wiley: Chichester, 413-38.

Lane, S.N., Hardy, R.J., Ferguson, R.I. and Parsons, D.R., 2005. A framework for model verification and validation of CFD schemes in natural open channel flows. In: Bates, P.D., Lane, S.N. and Ferguson, R.I. (eds.), *Computational Fluid Dynamics: applications in environmental hydraulics*, Wiley, Chichester, 169-91

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

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

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

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*, e14481.

Metcalfe, P., K.J. Beven, and J. Freer, 2015, Dynamic Topmodel: a new implementation in R and its sensitivity to time and space steps, *Environmental Modelling and Software*, 72, 155-172.

Morton, A., 1993. Mathematical models: questions of trustworthiness. *British Journal for the Philosophy of Science*, 44, 659-674.

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

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

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C. and Gupta, H.V., 2021. What role does hydrological science play in the age of machine learning?. *Water Resources Research*, *57*(3), p.e2020WR028091.

Nijssen, B., Bowling, L.C., Lettenmaier, D.P., Clark, D.B., El Maayar, M., Essery, R., Goers, S., Gusev, Y.M., Habets, F., Van Den Hurk, B. and Jin, J. et al., 2003. Simulation of high latitude hydrological processes in the Torne–Kalix basin: PILPS Phase 2 (e): 2: Comparison of model results with observations. *Global and Planetary Change*, 38, 31-53.

Nott, D. J., Marshall, L., & Brown, J., 2012. Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection?. *Water Resources Research*, 48,, W12602

Odoni, N. and Lane, S.N., 2010. Knowledge-theoretic models in hydrology. *Progress in Physical Geography*, 34, 151-71.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E. and Rakow, T., 2006. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

Oppy, G., Dowe, D., 2016, The Turing Test, Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/archives/spr2019/entries/turing-test/ (last accessed 19.08.2019)

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*, Springer, Dordrecht, 207-217.

Oreskes, N., Shrader-Frechette, K. and Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263, 641-646.

Overton, S., 1977. A strategy of model construction. In: C. Hall and J. Day (Editors), *Ecosystem Modeling in Theory and Practice: An Introduction with Case Histories.* John Wiley & Sons, New York. Reprinted 1990, University Press of Colorado.: 49-73.

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.

Palmer, T.N., 2016: A personal perspective on modelling the climate system. Proceedings of the Royal Society A472, https://doi.org/10.1098/rspa.2015.0772.

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, 595-613.

Pfister, L. and Kirchner, J.W., 2017. Debates—Hypothesis testing in hydrology: Theory and practice. *Water Resources Research*, *53*(3), 1792-1798.

Refsgaard, J.C., 1997. Parameterisation, calibration and validation of distributed hydrological models. *J. Hydrology,* 198, 69–97.

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

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

Roussos, J., Bradley, R. and Frigg, R., 2021. Making confident decisions with model ensembles. *Philosophy of Science*, *88*(3), pp.439-460.

Rykiel Jr, E.J., 1996. Testing ecological models: the meaning of validation. *Ecological Modelling*, 90, 229-244.

Sadegh, M., and Vrugt, J.A., 2014. Approximate bayesian computation using Markov chain Monte Carlo simulation: DREAM (ABC). *Water Resources Research*, 50, 6767-6787.

Seibert, J., Uhlenbrook, S., Leibundgut, C. and Halldin, S., 2000. Multiscale calibration and validation of a conceptual rainfall-runoff model. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, *25*(1), pp.59-64.

Seibert, J., and J. J. McDonnell, 2002. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Res. Res*., 38, 1241.

Singh, S.K. and Bárdossy, A., 2012. Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, *38*, pp.81-91.

Smith, A., Tetzlaff, D., Kleine, L., Maneta, M. and Soulsby, C., 2021. Quantifying the effects of land use and model scale on water partitioning and water ages using tracer-aided ecohydrological models. *Hydrology and Earth System Sciences*, *25*(4), pp.2239-2259.

Smith, M., Koren, V., Zhang, Z., Moreda, F., Cui, Z., Cosgrove, B., ... and Anderson, E., 2013. The distributed model intercomparison project–Phase 2: Experiment design and summary results of the western basin experiments. *Journal of Hydrology*, 507, 300-329.

Sornette, D., Davis, A. B., Ide, K., Vixie, K. R., Pisarenko, V., and Kamm, J. R. , 2007. Algorithm for model validation: Theory and applications. *Proceedings of the National Academy of Sciences*, 104, 6562-6567.

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

Stephenson, R. and Freeze, R.A., 1974. Mathematical Simulation of Subsurface Flow Contributions to Snowmelt Runoff, Reynolds Creek, Idaho, *Water Resour. Res*., 10, 284-298.

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

Thompson, E.L. and Smith, L.A., 2019. Escape from model-land. *Economics*, *13*(1).

Turing, A. M., 1950. Computing machinery and intelligence. *Mind*, 59, 433-460.

Van Griensven, A , Meixner , T., Srinivasan, R. and Grunwald, S., 2008, Fit-for-purpose analysis of uncertainty using split-sampling evaluations, *Hydrological Sciences Journal*, 53, 1090-1103.

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

Vrugt, J. A., & Sadegh, M. 2013, Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, *49*, 4335-4345.

Vrugt, J A and Beven, K J, 2018, Embracing Equifinality with Efficiency: Limits of Acceptability Sampling Using the DREAM(LOA) algorithm, *J. Hydrology,* 559, 954-971

Wagener, T. and McIntyre, N. 2005. Identification of rainfall-runoff models for operational applications. *Hydrological Sciences Journal*, 50, 751-67

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

Wagener, T., Reinecke, R. and Pianosi, F., 2022. On the evaluation of climate change impact models. *Wiley Interdisciplinary Reviews: Climate Change*, *13*(3), p.e772.

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., 2021. Knowledge gaps in our perceptual model of Great Britain's hydrology. *Hydrological Processes*, *35*(7), p.e14288.

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C.-Y., 2011, Calibration of hydrological models using flow-duration curves, *Hydrol. Earth Syst. Sci*., 15, 2205-2227,

Westerberg, I.K. and McMillan, H.K., 2015, Uncertainty in hydrological signatures, *Hydrology and Earth System Sciences*, 19, 3951-3968.

Young, P.C., 2013. Hypothetico-inductive data-based mechanistic modeling of hydrological systems. *Water Resources Research*, 49, 915-935.

Young, P. C., Parkinson, S. and Lees, M., 1996, Simplicity out of complexity in environmental modelling: Occam's razor revisited. *Journal of Applied Statistics,* 23, 165-210