# Benchmarking Hydrological Models for an Uncertain Future

Keith Beven[1]

[1]Lancaster University Lancaster Environment Centre

April 9, 2023

**Abstract**

This commentary discusses a framework for the benchmarking of hydrological models for different purposes when the datasets for different catchments might involve epistemic uncertainties. The approach might be expected to result in an ensemble of models that might be used in prediction (including models of different types) but also provides for model rejection to be the start of a learning process to improve understanding.

**Benchmarking Hydrological Models for an Uncertain Future**

Keith Beven

Lancaster Environment Centre, Lancaster University, UK

**Novel Aspects**

- A framework for model benchmarking is outlined
- Defining limits of acceptability for models while allowing for data uncertainties is emphasised
- If all models are rejected then it should instigate a learning process that will improve uncerstanding

**Abstract**

This commentary discusses a framework for the benchmarking of hydrological models for different purposes when the datasets for different catchments might involve epistemic uncertainties. The approach might be expected to result in an ensemble of models that might be used in prediction (including models of different types) but also provides for model rejection to be the start of a learning process to improve understanding.

**On benchmarking and intercomparisons of hydrological models**

One of the priority actions identified in the UK Flood Hydrology Roadmap (Environment Agency, 2022) concerns the issue of how to benchmark models for practical applications in flood hydrology. The aims would be two-fold: to ensure that the models used for operational applications can be considered as fit-for-purpose, and to provide a framework to make it easier for moving models from research into practice. Previous benchmarking exercises commissioned by the Environment Agency have been one-off projects for the comparison of 1D hydraulic models (Environment Agency, 2010) and later 2D hydraulic models (Environment Agency, 2013), but these were primarily model to model intercomparisons using hypothetical data sets rather testing for performance in real applications. At the time, there were good reasons for this: it established confidence in models giving consistent results without raising the additional concerns of data uncertainties in both model inputs and inundation datasets for evaluation. However, in the wider flood hydrology context, concerns about data and boundary condition uncertainties cannot be avoided. The question, therefore, is how data uncertainties might affect a benchmarking methodology.

There have been international intercomparisons of hydrological models in the past, including those organised by the World Meteorological Organisation (WMO) for real-time forecasting and snowmelt runoff models

(Sittner, 1976; Cavidias and Morin, 1986; Georgakakos and Smith, 1990;). Benchmarking has also been applied to land surface models, including for projects such as PILPS and PLUMBER (Henderson-Sellars et al., 1996; Abramowitz, 2012; Best et al., 2015; Haughton et al., 2016). More recently, model intercomparison and benchmarking projects have included DMIP and IHM-MIP projects for distributed models (e.g. Smith et al., 2004, 2012, 2013; Maxwell et al., 2014; Kollet et al., 2017); the Great Lakes Model Intercomparison project (e.g. Mai et al., 2022); benchmarking of NLDAS land surface models (e.g. Nearing et al., 2016, 2018); and the testing of model ensembles. These have taken the form either of testing which model provides the best simulations according to some metric (often using a split record test, e.g. Knoben et al., 2019); or testing against a benchmark model, either a chosen conceptual hydrological model (e.g. Newman et al., 2017; Seibert et al., 2018) or a purely data-based or machine learning model (e.g. Kratzert et al., 2019; Lees et al., 2021). Some benchmarking projects have also concentrated on seasonal and low flow forecasts (e.g. Nicolle et al., 2014; Girons-Lopez et al., 2021)

Experience from those intercomparisons involving hydrological models suggests that for most purposes there will be no model that can be considered as better than others: the relative performance will depend on which catchment is being simulated, which period or events are being simulated, and which performance measure or measures are chosen to do the evaluation. I have, of course, argued for a long, long time that the idea of an optimum hydrological model should be considered as untenable in favour of a concept of equifinality of models and parameter sets (e.g. Beven and Freer, 2001; Beven, 2006). Others have also suggested that the use of multiple metrics can reflect subjective judgments about the acceptability of different models (e.g. Gauch et al., 2022), though different experts might vary in their rankings (Crochemore et al., 2015).

Perhaps more interesting have been the benchmarking exercises involving comparisons with machine learning models (e.g. Nearing et al., 2021). In most of these studies it has been shown that the machine learning methods generally produce better predictions in both calibration and validation. This has included the training of the machine learning models on a large collection of catchments, when compared against models calibrated on single catchments (Figure 1). However, it is also the case that better does not always mean good. Distributions of the NSE efficiency across a large number of the US CAMELS dataset catchments show that there are some 10% of catchments where less than 50% of the variance in the discharge is captured by the models. Similar variation in performance has been reported in hydrological modelling studies of large numbers of catchments in France (Perrin et al., 2001) and the UK (Lane et al., 2019; Lees et al., 2021). So something else is also going on here, which clearly has an impact on benchmarking in the sense of whether models might be fit-for-purpose.

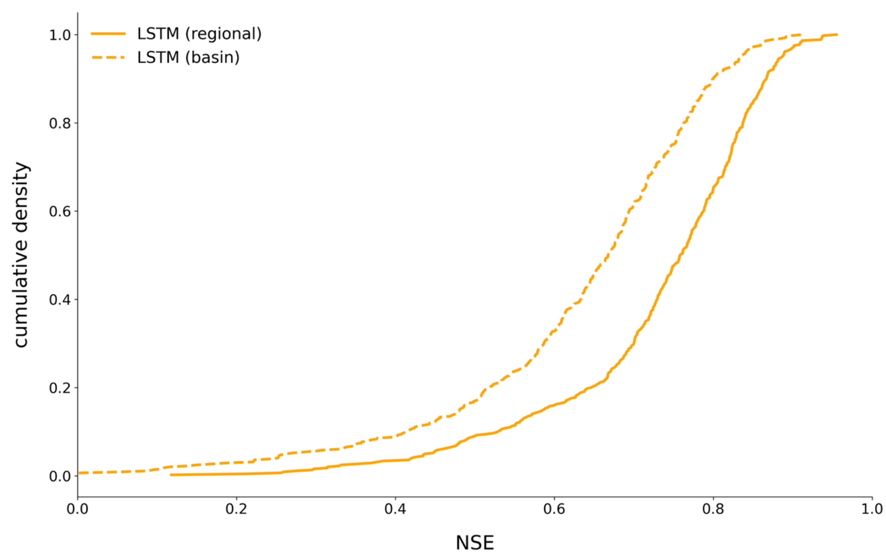

Figure 1. Cumulative distribution of NSE values over 531 catchments taken from the US CAMELS database using a single LSTM Deep Learning model trained over all the catchments (solid line) and separate LSTM models trained on the individual catchments (dotted line) (taken from Nearing et al., 2021)

**Benchmarking and data uncertainties**

There is also now increasing recognition of the way in which data and boundary condition uncertainties might influence how well models can be evaluated or tested as hypotheses about how catchment function (e.g. Beven and Binley, 1992; Beven and Freer, 2001; Liu et al., 2004; Coxon et al., 2014; Beven and Smith, 2015; McMillan et al., 2018; Beven, 2019; Beven and Lane, 2022; Westerberg et al., 2022). Clearly, we cannot expect any model to perform better than the quality of the data and boundary conditions it is supplied with, or of the data that are used in evaluation. This applies to both hydrological models and the machine learning methods that are intended to extract the maximum amount of information from the data. Indeed, Figure 1 suggests that averaging of potential observation errors across many catchments might be of value relative to training on only the data from a single catchment, even where those catchments include a wide range of physical characteristics. One interpretation of this is that epistemic errors in the observations might dominate model structural errors for some catchments (see also Beven, 2020).

That principle will also apply to both the hydrometric data and any tracer or geochemical data (e.g. Harmel et al., 2009; Krueger et al., 2009; Hollaway et al., 2018a). There have been a number of recent studies using "tracer-aided" model calibrations and evaluations (Birkel and Soulsby, 2015; Delavau et al., 2017; Smith et al., 2021; Stevenson et al., 2021) but these have not generally considered uncertainties in the data, and making use of such data will normally involve the introduction of additional parameters. For some water quality models, *many* more parameters might be involved (e.g. Hollaway et al., 2018b).

A particular aspect of epistemic uncertainty in the hydrometric data arises when the observations associated with individual events have runoff coefficients greater than 1 in catchments where the effects of snowmelt and longer term storage are not significant so that event-based coefficients can be calculated (Beven, 2019). Many hydrological models are constrained to satisfy mass balance and can therefore never reproduce an event that has a runoff coefficient greater than 1 (allowing for recession contributions for the previous event). Beven and Westerberg (2011) called such events disinformative events, in the sense of not providing useful information for model calibration (see also Beven et al., 2011; Beven and Smith, 2015; Beven et al., 2022b). Such events will also have an effect on the simulation of subsequent events since if the rainfall inputs for that event have been underestimated it will also impact the antecedent conditions for the next event. We can also envisage that there will also be events where the rainfall inputs are overestimated, with runoff coefficients artificially low, but these are much more difficult to identify securely. Such issues are a good argument for not imposing mass balance in flood forecasting models, but rather using data assimilation in real-time to compensate for errors in estimating the inputs.

For flood hydrology, there is also the issue of uncertainty in the estimation of flood peaks arising from rating curve uncertainties (e.g. Clarke, 1999; Costa and Jarrett, 2008; Westerberg et al., 2011; Domeneghetti et al., 2012; Coxon et al., 2015; McMillan and Westerberg, 2015). Uncertainties in rating curves can be estimated from statistical theory when the rating curves are fitted to observed discharges using regression methods. However, extrapolating well above the observed data points to estimating peak flows can also involve epistemic uncertainties as to the functional form of the curve (e.g. Hollaway et al., 2018a). In some cases, it might be possible to constrain the extrapolation using hydraulic modelling, but this can also introduce additional uncertainties in boundary conditions and roughness parameter estimates. Thus, in benchmarking models for flood flows it is important to consider such uncertainties.

**Benchmarking for a purpose**

This also raises a more general issue for benchmarking. What do we wish to benchmark for? Benchmarking is really a matter of trying to assess the confidence we might have in a model or models as fit-for-purpose. But fitness-for-purpose will depend on the purpose. We should expect that different model structures or parameter sets might be more or less suitable for different types of application, including the utility of data assimilation in real-time. Thus, the first step in any benchmarking exercise should be deciding on the purpose (see Figure 2). Different purposes might require different types of evaluation (N-step ahead predictions for forecasting; flood peaks for evaluating future change in flood hazard; annual exceedance probabilities for flood frequencies; flood inundation patterns for distributed models; .....) but all benchmarking evaluations will need to allow for the uncertainties in the observations.

This would be easier if we could safely assume that the uncertainties involved in model forcings and evaluation could be considered as aleatory and treated as stochastic variables. In that case the power of formal statistical methods for hypothesis testing can be used. This is not the case, however. As well as the rating curve extension problem there are other sources of epistemic uncertainty in the modelling process. Probably the most important is the question of estimating catchment rainfalls, either at the catchment scale or in some distributed form, from the limited rain gauge and uncertain radar data that might be available. This is an epistemic uncertainty problem, with the expectation that the uncertainty might vary in both time and space in rather arbitrary ways.

This then suggests that some alternative to statistical hypothesis testing might be needed for any benchmarking exercise. One approach is a logical extension of the expectation that there might be equifinality of model structures and parameter sets for different types of application, hopefully with many that might be considered as fit-for-purpose. This then suggests turning the problem around to consider what models and parameter sets might be considered as not fit-for-purpose, while allowing for the uncertainties in the forcing and evaluation data (Beven, 2018, 2019). Beven and Lane (2019, 2022) discuss the principles upon which such a rejectionist or model invalidation approach might be based, including the principle of defining limits of acceptability for a model to be considered as fit-for-purpose prior to any model runs being made. Of course, because this involves a consideration of epistemic sources of uncertainty, the definition of such limits of acceptability might require an input of expert judgment (though see Beven and Smith, 2015, Beven, 2019, and Beven et al., 2022a,b for examples of doing so based on historical event runoff coefficients that is applicable in catchments without significant baseflow). Particularly for the evaluation of distributed models, such judgments or feedback from local stakeholders might be needed to decide if models are getting the right results for the right reasons when distributed evaluation data are not available (Beven, 2007; Beven and Lane, 2022; Beven et al., 2022b).

**Benchmarking and fitness-for-purpose in predicting the future**

One of the implications of taking such an approach is that all the models tried might be rejected (see, for example, Brazier et al., 2000; Page et al., 2007; Choi and Beven, 2007; Dean et al., 2009; Hollaway et al., 2018b). As I have written many times before, this is, of course, a good thing in that it forces a re-evaluation of some sort. This could be a re-evaluation of model structures, of how the model parameters are sampled, of the consistency of the available observations, or of the range for the limits of acceptability. Since it will always be possible to extend the limits arbitrarily to ensure that not all the models are rejected, it is important that the assumptions on which the limits are based be clearly stated. We can extend this to the requirement that there should be an audit trail to justify and record all the assumptions associated with any benchmarking study, that then allows those assumptions to be revisited later (Beven and Alcock, 2012; Beven and Lane, 2022). The CURE uncertainty estimation toolbox, for example, has a facility for producing such an audit trail as an output from an analysis (Page et al., 2023).

In setting limits of acceptability, we are necessarily constrained to using evaluations based on past events and historical time series (unless doing so on a purely subjective basis as to what might be considered fit-for-purpose). Beven and Lane (2022) suggest 8 principles for setting limits of acceptability, including where this might involve expert elicitation. However, in many cases the reasons for using a hydrological model are to predict what might happen under future conditions. This could be an expected change in the inputs projected by a climate model, or a change in catchment characteristics as a result, for example, of natural flood management measures, deforestation or urbanisation. In the case of changes in inputs, the value of evaluations based on historical data will depend on the range of past conditions monitored (see Wi and Steinschneider, 2022, for an example using a deep learning model). If future conditions, especially the extremes are expected to be outside the range of past behaviours, then both process-based and data-based or machine learning models might be limited in their abilities to predict such changes outside any training data (e.g. Beven, 2020). In the case of changes in catchment characteristics, the training data might again not include examples of such changes. We then either have to transfer information from catchments where similar changes have occurred or make subjective judgments about changes in parameter values. This can work (e.g. Buytaert and Beven, 2009) but might not work consistently. Where catchments have been monitored over periods of such change, then evaluations of predictions of such change could be assessed. If acceptable models are found, this can give increased confidence in applications elsewhere.

It is clear that the types of limits of acceptability that might be used in model evaluation, and the way in which they might be defined before making model runs will very much depend on the purpose for which a model might be used. Taking each of the vertical pathways in Figure 2, for example, it will be appreciated that what is required for N-step ahead real time forecasting will be different to the use of a catchment model for continuous simulation flood frequency simulation, or for the prediction of future catchment change, distributed inundation predictions for planning purposes, or for tracer or water quality variables. What figure 2 provides, however, is a common framework for assessing model performance in a way that can allow considerations of data uncertainties (and more subjective evaluation measures) to be incorporated in a consistent and thoughtful way. It provides an alternative to considering benchmarking in terms purely of relative values of performance indices, that in the past have often ignored the effects of observational errors on model performance (but which might also include some additional dimensions of ease of understanding and use and costs of application). In this respect we should learn from the poor performance of both machine learning methods and conceptual hydrological models in some catchments to really think about what might be considered as fit-for-purpose for a particular application.

Of course, if it is necessary to reject all the models that are tried for a particular purpose in a particular catchment of interest, it should be the start of a learning process (as shown in Figure 2). This could be learning about the failings of a particular model structure, though it may often be difficult to understand why a model has failed, especially in the case of a machine learning model. In many cases it will be a result of providing the modelling process with inadequate or inconsistent data. Machine learning, for example, should be able to deal with data that have consistent errors (Beven, 2020). The fact that it still seems to provide poor results on some catchments (e.g. Frame et al., 2023) would certainly suggest that there are inconsistent errors or forms of disinformation in some catchment datasets that limit predictive performance. While such rejections do not help a decision maker, they are important to advancing understanding of the modelling process (e.g. Beven, 2018). In extremis a decision maker could still have resort to trying to characterise the errors associated with each model run, and to allow for those errors by being precautionary in her decisions. Still better, of course, would be to understand just why models might fail benchmarking tests.

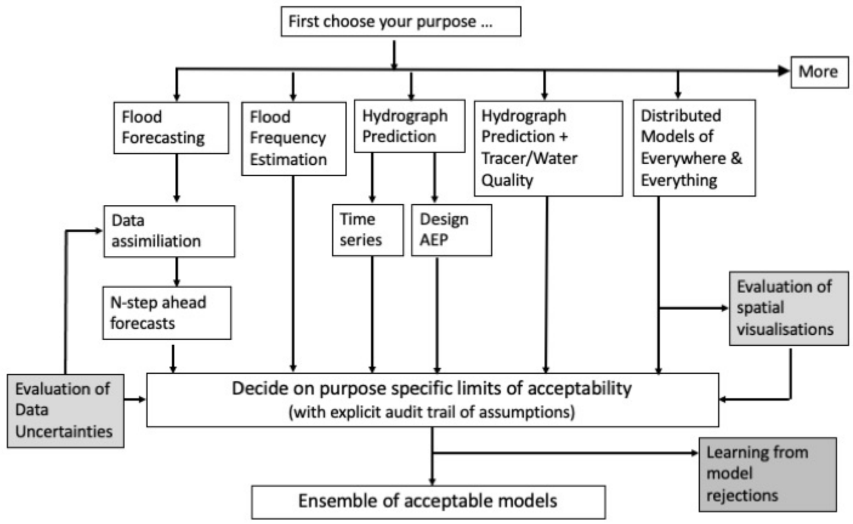


Figure 2. A framework for model benchmarking for different purposes. Light shading indicates the need for decisions about how temporal ands spatial observations and their uncertainties are used to define limits of acceptability. Learning from model rejections indicates an area of research that is largely unexplored (though intrinsic to most model development).

**References**

Abramowitz, G.,, 2012: Towards a public, standardized, diagnostic benchmarking system for land surface models. Geosci. Model Dev., 5, 819–827, doi:10.5194/gmd-5-819-2012.5, 819–827, doi:10.5194/gmd-5-819-2012.

Best, M.J., Abramowitz, G., Johnson, H.R., Pitman, A.J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P.A., Dong, J. and Ek, M., 2015. The plumbing of land surface models: benchmarking model performance. *Journal of Hydrometeorology* , *16* (3), 1425-1442.

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, 2018, On hypothesis testing in hydrology: why falsification of models is still a really good idea, WIRES Water, 5(3), e1278, DOI: 10.1002/wat2.1278.

Beven, K. J., 2019, Towards a methodology for testing models as hypotheses in the inexact sciences, Proceedings Royal Society A, 475 (2224), 20180862, doi: 10.1098/rspa.2018.0862

Beven, K. J., 2020, Deep Learning, Hydrological Processes and the Uniqueness of Place, *Hydrological Processes,* 34(16), 3608-3613, doi: 10.1002/hyp.13805

Beven, K.J. and A.M. Binley (1992), The future of distributed models: model calibration and uncertainty prediction, *Hydrological Processes* , 6, 279-298.

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

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., and Smith, P. J., 2015, Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models, *ASCE J. Hydrol. Eng* ., *20* (1), p.A4014010, doi: 10.1061/(ASCE)HE.1943-5584.0000991.

Beven, K. J. and Lane, S., 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. 145-171.

Beven, K. J. and Lane, S., 2022. On (in)validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for purpose. *Hydrological Processes* , 36(10), e14704,*https://doi.org/10.1002/hyp.14704*.

Beven, K. J., Lane, S., Page, T., Hankin, B, Kretzschmar, A., Smith, P. J., Chappell, N., 2022a, On (in)validating environmental models. 2. Implementation of the Turing-like Test to modelling hydrological processes, *Hydrological Processes* , 36(10), e14703, https://doi.org/10.1002/hyp.14703.

Beven, K. J., Page, T., Hankin, B, Smith, P.J., Kretzschmar, A., Mindham, D., Chappell, N., 2022b, Deciding on fitness-for-purpose - of models and of natural flood management, *Hydrological Processes, 36* (11), e14752, https://doi.org/10.1002/hyp.14752.

Birkel, C. and Soulsby, C., 2015. Advancing tracer-aided rainfall–runoff modelling: A review of progress, problems and unrealised potential. *Hydrological Processes* , *29* (25), 5227-5240.

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.

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

Cavadias, G. and Morin, G., 1986. The combination of simulated discharges of hydrological models: Application to the WMO intercomparison of conceptual models of snowmelt runoff. *Hydrology Research* , *17* (1), 21-32.

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 (3-4): 316-336

Clarke, R.T., 1999. Uncertainty in the estimation of mean annual flood due to rating-curve indefinition. *Journal of Hydrology* , *222* (1-4), 185-190

Costa, J.E. and Jarrett, R.D., 2008. *An evaluation of selected extraordinary floods in the United States reported by the US Geological Survey and implications for future advancement of flood science* (No. 2008-5164). US Geological Survey.

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), 6135-6150.

Coxon, G., Freer, J., Westerberg, I.K., Wagener, T., Woods, R. and Smith, P.J., 2015. A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water resources research* , *51* (7), 5531-5546.

Crochemore, L., Perrin, C., Andréassian, V., Ehret, U., Seibert, S.P., Grimaldi, S., Gupta, H. and Paturel, J.E., 2015. Comparing expert judgement and numerical criteria for hydrograph evaluation. *Hydrological sciences journal* , *60* (3), 402-423.

Dean, S., Freer, J., Beven, K., Wade, A.J. and Butterfield, D., 2009. Uncertainty assessment of a process-based integrated catchment model of phosphorus. *Stochastic Environmental Research and Risk Assessment* , *23* , 991-1010.

Delavau, C. J., Stadnyk, T., and Holmes, T., 2017, Examining the impacts of precipitation isotope input ($\delta\ ^{18}O_{ppt}$) on distributed, tracer-aided hydrological modelling, Hydrol. Earth Syst. Sci., 21, 2595–2614, https://doi.org/10.5194/hess-21-2595-2017.

Domeneghetti, A., Castellarin, A. and Brath, A., 2012. Assessing rating-curve uncertainty and its effects on hydraulic model calibration. *Hydrology and Earth System Sciences* , *16* (4), 1191-1202.

Environment Agency. (2010). Benchmarking of 2D hydraulic modelling packages (Report No. SC080035/SR2), Environment Agency, Bristol, UK.

Environment Agency. (2013). Benchmarking the latest generation of 2D hydraulic modelling packages (Final Technical Report Project SC120002). Environment Agency, Bristol, UK.

Environment Agency. (2022). Flood Hydrology Roadmap. (Report No. FRS18196/R1). Environment Agency, Bristol, UK.

Frame, J., Ullrich, P., Nearing, G., Gupta, H. and Kratzert, F., 2023. On strictly enforced mass conservation constraints for modeling the rainfall-runoff process. *Hydrological Processes,* 37(3), e14847, https://doi.org/10.1002/hyp.14847

Gauch, M., Kratzert, F., Gilon, O., Gupta, H., Mai, J., Nearing, G., Tolson, B., Hochreiter, S. and Klotz, D., 2022. In Defense of Metrics: Metrics Sufficiently Encode Typical Human Preferences Regarding Hydrological Model Performance, eartharxiv,*https://doi.org/10.31223/X52938*

Georgakakos, K.P. and Smith, G.F., 1990. On improved hydrologic forecasting—Results from a WMO real-time forecasting experiment. *Journal of Hydrology* , *114* (1-2), 17-45.

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.

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

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. 2018a, A method for uncertainty constraint of catchment discharge and phosphorus load estimates. *Hydrological Processes* . 32:2779- 2787. https://doi.org/10.1002/hyp.13217

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., 2018b, 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.

Knoben, W.J., Freer, J.E. and Woods, R.A., 2019. Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences* , *23* (10), 4323-4331.

Kollet, S., Sulis, M., Maxwell, R.M., Paniconi, C., Putti, M., Bertoldi, G., Coon, E.T., Cordano, E., Endrizzi, S., Kikinzon, E. and Mouche, E., 2017. The integrated hydrologic model intercomparison project, IH-MIP2: A second set of benchmark results to diagnose integrated hydrology and feedbacks. *Water Resources Research* , *53* (1), .867-890.

Krueger, T., Quinton, J.N., Freer, J., Macleod, C.J., Bilotta, G.S., Brazier, R.E., Butler, P. and Haygarth, P.M., 2009. Uncertainties in data and models to describe event dynamics of agricultural sediment and phosphorus transfer. *Journal of Environmental Quality* , *38* (3), 1137-1148.

Lane, R.A., Coxon, G., Freer, J.E., Wagener, T., Johnes, P.J., Bloomfield, J.P., Greene, S., Macleod, C.J. and Reaney, S.M., 2019. Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain. *Hydrology and Earth System Sciences* , *23* (10), 4011-4032.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G. and Dadson, S.J., 2021. Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences* , *25* (10),.5517-5534.

Liu, Y, Freer, JE, Beven, KJ and Matgen, P, 2009, Towards a limits of acceptability approach to the calibration of hydrological models: extending observation error, *J. Hydrol.* , 367:93-103, doi:10.1016/j.jhydrol.2009.01.016.

Mai, J., Shen, H., Tolson, B.A., Gaborit, É., Arsenault, R., Craig, J.R., Fortin, V., Fry, L.M., Gauch, M., Klotz, D. and Kratzert, F., 2022. The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL). *Hydrology and Earth System Sciences* , *26* (13), 3537-3572.

Maxwell, R.M., Putti, M., Meyerhoff, S., Delfs, J.O., Ferguson, I.M., Ivanov, V., Kim, J., Kolditz, O., Kollet, S.J., Kumar, M. and Lopez, S., 2014. Surface-subsurface model intercomparison: A first set of benchmark results to diagnose integrated hydrology and feedbacks. *Water resources research* , *50* (2), 1531-1549.

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

McMillan, H.K., Westerberg, I.K. and Krueger, T., 2018. Hydrological data uncertainty and its implications. *Wiley Interdisciplinary Reviews: Water* , *5* (6), p.e1319.

Nearing, G.S., Mocko, D.M., Peters-Lidard, C.D., Kumar, S.V. and Xia, Y., 2016. Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *Journal of hydrometeorology* , *17* (3), 745-759.

Nearing, G.S., Ruddell, B.L., Clark, M.P., Nijssen, B. and Peters-Lidard, C., 2018. Benchmarking and process diagnostics of land models. *Journal of Hydrometeorology* , *19* (11), 1835-1852.

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.

Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B. and Nearing, G., 2017. Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology* , *18* (8), 2215-2225.

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., Ramos, M.H., Cloke, H.L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A. and Salamon, P., 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology* , *522* , 697-713.

Perrin, C., Michel, C. and Andreassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal* *of hydrology* , *242* (3-4), 275-301.

Sittner, W.T., 1976. WMO project on intercomparison of conceptual models used in hydrological forecasting. *Hydrological Sciences Journal* , *21* (1), 203-213.

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), 2239-2259.

Smith, M.B., Seo, D.J., Koren, V.I., Reed, S.M., Zhang, Z., Duan, Q., Moreda, F. and Cong, S., 2004. The distributed model intercomparison project (DMIP): motivation and experiment design. *Journal of Hydrology* , *298* (1-4), 4-26.

Smith, M.B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., Cui, Z., Mizukami, N., Anderson, E.A. and Cosgrove, B.A., 2012. The distributed model intercomparison project–Phase 2: Motivation and design of the Oklahoma experiments. *Journal of Hydrology* , *418* , 3-16.

Smith, M.B., Koren, V., Zhang, Z., Moreda, F., Cui, Z., Cosgrove, B., Mizukami, N., Kitzmiller, D., Ding, F., Reed, S. 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.

Stevenson, J.L., Birkel, C., Neill, A.J., Tetzlaff, D. and Soulsby, C., 2021. Effects of streamflow isotope sampling strategies on the calibration of a tracer-aided rainfall-runoff model. *Hydrological Processes* , *35* (6), p.e14223.

Westerberg, I., Guerrero, J.L., Seibert, J., Beven, K.J. and Halldin, S., 2011. Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes* , *25* (4), 603-613.

Westerberg, I.K., Sikorska-Senoner, A.E., Viviroli, D., Vis, M. and Seibert, J., 2022. Hydrological model calibration with uncertain discharge data. *Hydrological Sciences Journal* , *67* (16), 2441-2456.

Wi, S. and Steinschneider, S., 2022. Assessing the physical realism of deep learning hydrologic model projections under climate change. *Water Resources Research* , *58* (9), p.e2022WR032123.