# Comment on 'Of data and models'

Jim W. Hall

**Jim W. Hall**
Department of Civil Engineering,
University of Bristol,
Queen's Building,
University Walk,
Bristol BS8 1TR
UK
Tel: +44 117 928 9763
Fax: +44 117 928 7783
E-mail: *Jim.Hall@bristol.ac.uk*

The author is to be applauded for probing how data and models are employed in the practice of hydroinformatics. In particular, his approach from the point of view of an engineer who wishes to predict the consequences of unobserved interventions in the aquatic environment is welcome. The purpose of this comment is to scrutinize some aspects of the author's reasoning, which on the whole he justifies on the basis of his admirable experience and intuition, in order to establish whether or not it can be placed on a more sound theoretical footing by recourse to modern theories of uncertainty.

I would like to begin by suggesting a change in terminology, a suggestion that, as I hope I will demonstrate, is based on more than pedantry. The author refers to 'deterministic' models as the opposite of 'data-driven' models. However, the distinguishing feature of the class of models he refers to as 'deterministic' is not their purported determinism, but rather that they are based on general and accepted physical principles, such as conservation of mass and momentum. In other words, they are physics-based or mechanistic models. By suggesting that physics-based models are not necessarily deterministic I am not entertaining the ontological question as to whether or not reality is deterministic. I am merely making the practical point that because a model is physics-based, the modeller does not have to employ it in a deterministic mode. Probabilistic, fuzzy or more generally imprecise quantities, not just point-valued quantities, may all be propagated through physics-based models (Hall 2003). Moreover, there is no reason why, even with point-valued inputs, a physics-based model need generate a unique point-valued output—in other words it may enact a multi-valued mapping (Dempster 1967). Recognizing that physics-based models need not be deterministic is more than a mere technicality. It unlocks uncertainty theory that can provide a theoretical basis for the modified modelling paradigm the author proposes.

The author sets great store by the assertion that mechanistic models are '*compelling*, *appealing*, *simple* and *related to the physics*' (his italics). These are indeed persuasive and widely held justifications for a hydraulic engineer to believe a model's predictions. However, the psychological appeal of causality may be a double-edged sword (or crutch), which perhaps is what Bertrand Russell had in mind when he offered to purge the word 'causation' from the language of science. Physics-based models are appealing because of their generality, but their generality will always, in theory and perhaps also in practice at scales of relevance to the hydraulic modeller, be bounded, so in principle it could be argued that there is nothing to distinguish physics-based and data-based models other than the range of their generality. Certainly there is a need to be more explicit about the range of applicability of any model. Hydroinformatics tools could do more to capture this important information and present it to users who unwittingly depart from that range.

Physics-based appeal is not the only reason the author cites for believing a model. Empirical verification is rightly

also paid considerable attention. But, it is argued, in practical hydraulic modelling studies, there may be precious few verification data. True, though perhaps the case is overstated–for example the flood outlines that the author suggests are impossibly costly to obtain are now available from airborne Synthetic Aperture Radar observations (see for example Aronica *et al.* 2002), occasionally with several images from a single flood, though seldom on the rising limb of the hydrograph. A responsible engineer will establish a degree of belief in a model on the basis of a body of *evidence*, which may take several forms. It will include verification data, analogous cases, physics-based reasoning and scrutiny of the quality of the process of model construction and use (how competent were the practitioners, how much time did they dedicate to the task, etc.). The mathematical theory of evidence provides mechanisms for representing, combining and generating inferences with uncertain evidence, which can formalize the reflective behaviour of responsible practitioners that the author refers to. This formalization is necessary because the integrated modelling activities that modern hydroinformaticians are engaged in have become too complex for any single individual to sensibly reason about without computer support. Hydroinformaticians operate in multi-disciplinary teams, engaging with a diverse range of stakeholders. Externalizing and formalizing reasoning about model dependability can help to improve communication and cultivate richer, less ambiguous shared understanding.

Data-driven approaches may be unavoidable, and the author fails to acknowledge that they have a respectable tradition of being used to describe the long-term characteristics of weather-related phenomena. Not only are (entirely data-driven) statistical methods more or less universally used to describe long term rainfall, offshore waves and sea surface elevations at a site, it is also customary to use extreme value theory to extrapolate beyond the observed data in order to estimate the severity of very rare events. Naturally great care is required, for example in assessing the stationarity and representativeness of the available data, yet there is seldom a physics-based alternative. The author overlooks the practice of extreme value statistics when he anathematizes extrapolation of data-based models.

The author is in danger of misrepresenting some modelling activities, perhaps in an attempt to endow them with a respectability that may not in fact be warranted, by forcing them into the physics-based paradigm. For example, he repeatedly refers to the empirical observation of Manning's $n$, asserting that $n$ 'can be estimated with sufficient accuracy . . . on the basis of engineering experience'. He doesn't probe the nature of that mysterious 'engineering experience' because of course to do so would reveal that it has built up from many applications of the type of calibration procedure that he goes on to decry! It would be more honest to acknowledge that there is a populous middle ground between the poles of purely physics-based and purely data-based approaches. Surely this is what Babovic *et al.* (2001) had in mind when they wrote '. . . we strongly believe that the most appropriate way forward is to combine the best of two approaches–theory-driven, understanding-rich, with data-driven modelling processes'. This hybrid position is explicit in Peter Young's (1998, 1999) data-based mechanistic models, in the physical interpretation of the stochastic models of Hall *et al.* (2002) and in the linguistic interpretation of fuzzy rules learnt from data (Lawry *et al.*, 2004), all of which might be regarded as different types of 'grey-box' models. Recognition of this middle ground provides the opportunity to avoid the outbreak of war, which the author warns against, between proponents of data-driven and physics-based approaches. It does not merely represent an unsatisfactory staging post on the quest for a reductionist holy grail of purely physics-based approaches. As Beven (2002) has argued there will always be sub-grid-scale processes that we are incapable of exactly measuring at an appropriate scale.

An important feature of the author's proposed modified modelling paradigm is that it is recognized that there will be a window of uncertainty around a model prediction, implying that model predictions should, in general, be regarded as being *imprecise*. This is quite different to generating probabilistic predictions, where the assumption (in both the Bayesian and the frequentist paradigms) is that as more data are acquired the distribution of model predictions should converge to the distribution of the data. Rather, it is an acknowledgement that given available evidence it is not possible to

specify a precise outcome, or indeed a precise distribution of outcomes. The imprecise probabilistic approach has several desirable side-effects. To generate imprecise predictions requires quite thorough exploratory analysis of model space, which invariably yields insights into model behaviour that would not have been revealed by a handful of deterministic model runs. It is accepted that data do not necessarily appear in a precise format and that in situations of information scarcity we will wish to make use of data whatever their format. The author refers to descriptive observations in hydrology: 'the rain is extremely heavy and I can hear a horrible noise made by a stream upstream in the hills'. These linguistic observations, which lend themselves to representation with fuzzy sets, can be readily admitted into the imprecise formalism. More fundamentally, as I have argued previously in this journal (Hall 2003), the imprecise approach captures the relationship identified by Karl Popper between truth and information content. The less precise (i.e. the less informative) a statement is, the less prone to falsification. Precise (deterministic) models are attractive because they have very high information content, yet on the other hand are very likely to be falsified, in that they will not coincide exactly with future observations. We can guard against falsification by making appropriately imprecise statements. The language of imprecision enables effective communication of uncertainty to decision-makers. In some cases they will be able to make a robust decision regardless of the uncertainty, whilst in other cases only a set of possible options will be demonstrable as being more desirable than the remainder. This approach is quite different to normative probabilistic decision theory and is much more closely akin to what Frank Knight (1921) thought of as decision-making under uncertainty.

The foregoing raises some considerable challenges for the practice of hydroinformatics. The exploratory analysis of uncertainty and decision robustness needs to become routine. The success of simple spreadsheet add-ins for Monte Carlo simulation demonstrates, at a rather naïve level, the demand for this type of approach. The conversion of hydroinformatics tools from the deterministic paradigm to deal with probabilistic, fuzzy and more general uncertainty structures is long overdue.

Professor Cunge, in his inspiring paper, has also demonstrated the need for computer-based support for reasoning about the evidence for belief in every model application. The model user should be able to access knowledge about the bounds of model applicability, verification and performance in analogous cases, as well as recording their own insights. Promising responses to these challenges are already under development at the University of Bristol (MDF 2003). Professor Cunge's important paper has given their development renewed justification and impetus.

## REFERENCES


Aronica, G., Bates, P. D. & Horritt, M. S. 2002 Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. *Hydrological Processes*. **16**, 2001–2016.

Babovic, V., Canizares, R., Jensen, H. R. & Klinting, A. 2001 Neural networks as routine for error updating of numerical models. *J. Hydraulic Engineering, ASCE*. **127**(3), 181–193.

Beven, K. 2002 Towards a coherent philosophy for modelling the environment. *Proc. Royal Society London*. A, 458, 2465–2484.

Cunge, J. 2003 Of data and models. *J. Hydroinformatics* **5**(2), 75–98.

Dempster, A. P. 1967 Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*. **38**, 325–339.

Hall, J. W., Meadowcroft, I.C., Lee, E.M. & van Gelder, P. H. A. J. M. 2002 Stochastic simulation of episodic soft coastal cliff recession. *Coastal Engineering*. **46**(3), 159–174.

Hall, J. W. 2003 Handling uncertainty in the hydroinformatic process. *J. Hydroinformatics*. **5**(4), 215–232.

Knight, F. H. 1921 *Risk, Uncertainty and Profit*. Houghton Mifflin, Boston.

Lawry, J., Hall, J. W. & Bovey, R. 2004 Fusion of expert and learnt knowledge in a framework of fuzzy labels. *International Journal of Approximate Reasoning* (in press).

MDF 2003 Model Description Framwork web site: *http://mdf.hydoinformatics.org.uk*

Young, P. 1998 Data-based mechanistic mdelling of environmental, ecological, economic and engineering systems. *Environmental Modelling and Software*. **13**, 105–122.

Young, P. 1999 Data-based mechanistic modelling, generalised sensitivity and dominant mode analysis. *Computer Physics Communications*. **117**, 113–129.