INVITED COMMENTARY

HP TODAY

HYDROLOGICAL PROCESSES
Hydrol. Process. **26**, 1905–1908 (2012)
Published online 10 April 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hyp.9261

On virtual observatories and modelled realities (or why discharge must be treated as a virtual variable)

Keith Beven,[1,3,4]* Wouter Buytaert[2] and Leonard A Smith[3,5]

[1] *Lancaster Environment Centre, Lancaster University, Lancaster, UK*
[2] *Civil and Environmental Engineering, Imperial College, London, UK*
[3] *Centre for Analysis of Time Series, London School of Economics, London, UK*
[4] *Department of Earth Sciences, Geocentrum, Uppsala University, Uppsala, Sweden*
[5] *Pembroke College, Oxford, UK*

*Correspondence to:
Keith Beven, Environmental Science, Lancaster University, LA1 4YQ, Lancaster, UK. E-mail: k.beven@lancaster.ac.uk

Received 25 November 2011
Accepted 29 November 2011

Out there in the cloud, there is more computing power, and there are more databases, more images, more models and more model output than have ever existed before. There also are a variety of projects in different countries to provide ways of making all that information more readily available at different scales and to different types of users, such as the Natural Environment Research Council-funded pilot Virtual Observatory project in the UK, the Earth Cube initiative of the US National Science Foundation, and the Global Earth Observation System of Systems. There also are calls for hyper-resolution earth system science models at the global scale, analogous to virtual observatories, as a way ahead in predicting global change (Wood *et al.*, 2011).

It therefore seems worthwhile to reflect on the nature of all that activity in producing a virtual observatory as a representation of our understanding and observations of the real world. In effect, although a virtual observatory might serve simply to facilitate access to existing observations, there also will be a strong driver to blend those observations with simulation models. A virtual observatory will then also serve to manufacture virtual observations based on model simulations, either at places where observations have not yet been made or at times in the past or future where making additional observations is not actually possible and, of course, at places where the actual observations are judged noisy, unreliable, or incoherent.

In fact, the distinction between real and virtual observations is already rather more blurred than it should be. In hydrology, it is not commonly the case that stream discharge is a real observation. Much more often, it is derived from measurements of water level through a rating curve. The rating curve is itself a model that can be used to interpolate and extrapolate to high and low discharges beyond the range of the available measurements of discharge, with the possibility of making false extrapolations. The form and parameters of that model may be more or less robust and stationary depending on the characteristics of the site (e.g. Herschy, 2009; Westerberg *et al.*, 2011). Similar considerations apply to many of the variables used by hydrologists, including catchment inputs interpolated from point rain gauges or estimates of rainfall inferred from radar reflectivity, and variables derived from remote sensing digital numbers through some interpretative model (that will have its own uncertain parameters). Thus, many variables are already treated as observations even if they are model derived or uncertain estimates of the real variables.

This should surely be considered bad practice. Model-derived variables are not observations. They are virtual observations that should be clearly distinguished from what is actually observed. It is then only a small step to simulations of variables that are not directly observable (or have not been observed) using simulation models. There will then be the possibility of confusing virtual variables and direct observations. The ability for a user to distinguish one from the other will fade away as ways of visualizing the outputs from the virtual observatory become more and more sophisticated.

In fact, it will generally be much easier to visualize virtual variables in theee and four dimensions than observation-derived variables because the observations are limited in both space and time (or are not necessarily the hydrological relevant variables), whereas the virtual observations can appear to be complete in space and time. However, a better visualization does not necessarily mean better information content when it comes to making decisions (see Beven and Cloke, 2012); a prettier picture may not provide deeper insight and might actually be misleading, particularly when uncertainties are high.

So how do we try to ensure that virtual observatories help to improve decision making rather than providing misleading virtual information? When does the unavoidable error in the virtual information become misinformation or disinformation? Clearly, some model-derived or simulated variables might be expected to be more robustly estimated than others, but it is quite possible that virtual information could be misleading because of all the uncertainties that arise in the modelling process (see Beven, 2006, 2009), including uncertainties, incommensurabilities and inconsistencies in the available observations themselves (e.g. Beven and Westerberg, 2011). In as much as the observations also are imprecise, observational error will mix with epistemic representational error, and the resulting product will infect the entirety of the virtual observational space. Unlike those forms of observational error that can be considered as aleatory, there are no techniques corresponding to confidence intervals/error bars once the error goes viral in this way. Deep questions regarding 'simple' operations, like subtracting a virtual observation from an actual observation, led Lorenz to coin the word 'subtractable' (Lorenz 1985, Smith 2006) in the context of evaluating forecasts. Worse, in combining virtual and observed variables, the observational errors and gaps can be easily obscured if not made truly invisible. This also should be considered to be a bad practice.

A virtual observatory can be (at best) an approximate description of the real system under study. So the question is for which purposes can we expect this approximation to be adequate and for which will it be significantly misinformative. There are very many different types of purpose for which such a system might be useful in catchment management decisions. That naturally leads to a further question of how to define whether a model should be considered 'adequate' in making predictions about the future that might be used to inform such decisions, especially when there is necessarily epistemic uncertainty about the boundary conditions (and also the process representations) for such predictions into the future (see Parker 2010). Of course, what proves adequate for one decision maker will not prove adequate for another, leading to a variety of competing virtual worlds without a clear indication of which might be the most useful for a given purpose.

This is not a question that has been widely discussed in the literature. There are many studies that have simply taken available models, generally with some calibration against past data, and used them for predicting the impacts of future change. However, the best available model (or models) might not necessarily be fit for purpose for such applications (e.g. Smith, 2000, 2006; Beven, 2010, 2011). Again, some tests of adequacy are required, at least in representing the past and present even if we cannot fully test adequacy in evaluating the impacts of change. The virtual observatory will need to convey that assessment of adequacy to the users and decision makers in some way. There is no tradition of doing so for hydrological variables, even for the estimation of discharges (although this is starting to change).

The question of how to calibrate or condition a model or models based on past data, and how to represent their uncertainties, has been extensively discussed in the literature. Therefore, it might seem surprising that there is not already a consensus about defining an adequate model or (ensemble of models) but only a competing range of methodologies (BATEA, DREAM, GLUE and others). This is in part because of a lack of agreement about how to handle the wide range of uncertainties in the modelling process (e.g. Beven, 2006, 2010). Statistical methods, including Bayesian methodologies, are limited to fitness within a model class, which, assumed to be valid, then equates 'maximum likelihood' with 'fit for purpose'. Challenges arise when descriptive models, which are valuable for understanding the relative importance of various processes but which were never intended to be taken seriously in terms of their quantitative outputs because of known unknowns, are cast as providing relevant quantitative outputs that are merely uncertain. What part should such descriptive models play in a virtual observatory?

The tradition in hydrology also is to think in terms of the identification of parameters rather than testing models as adequate hypotheses of how a catchment functions, given a set of data and many sources of uncertainty. What is needed in defining whether a model is adequate is some form of hypothesis testing that allows for the fact that many of the sources of uncertainty are epistemic rather than aleatory in nature (e.g. Smith, 2006; Beven, 2010; Buytaert and Beven, 2011) while avoiding Hume's pitfall of induction (Howson, 2003). In particular, virtual observatories aim to represent everywhere, but the observations that might normally be used to assess models are not available everywhere (Beven, 2007). Thus, epistemic uncertainties are generic to the virtual observatory. The best that can be hoped is that a model can be shown to

shadow the available observations within the limits of observational error (including the observations used to create the inputs to a model) (Beven, 2006, 2010; Smith, 2006).

That is exactly why defining whether a model is adequate or fit for purpose is so difficult. Models are approximations and cannot be expected to shadow forever, but the time scales on which a model does shadow indicates the time scales on which it is conceivable to argue that we are dealing with measurement uncertainties in the inputs. On longer time scales, the issue is not uncertainty but indeterminacy, and the methodologies for hypothesis testing in the face of these epistemic issues are not well developed. So there are some really important questions to be resolved in setting up virtual observatories as modelled realities. Hypothesis testing might then need to rely on more qualitative input of information into the virtual observatory (photographs, observations by local residents, ...) or on defining critical experiments designed for hypothesis testing. A framework for hypothesis testing needs to evolve within such virtual observatories that goes beyond simply using the best available models, especially where these do not shadow the observations to within limits of observational error (as is the case for many environmental models). Models that fail such tests might not provide adequate evidence for decision making, even if they are the only predictions available.

However, this is not (only) a problem; it is an opportunity, an opportunity either to improve our methodology for using models (and the models themselves) to overcome those deficiencies or, if that is not possible within the time scale required for a decision, to come to a better decision in some other way. Such an approach is entirely consistent with a scientific methodology based on hypothesis testing and is more likely to avoid false confidence.

However, what form of hypothesis testing is possible when we fully understand that there are epistemic uncertainties in the modelling process? If, in the words of George Box, all models are wrong but some are useful, how is it possible to distinguish between the patently wrong and the useful approximation when, in general, we might expect to see a wide spectrum of performance, regardless of performance measure, and when any information available to evaluate model performance also might be subject to both epistemic and random errors?

Statistical methods for hypothesis testing are well developed, but what they offer is weak: 'rejection' or 'failure to reject' conditional on assuming that a model structure is correct and that data are subject only to random errors. This is effectively an assumption that epistemic uncertainties are negligible (or can be represented as if they were random in nature). It is difficult to see how such an assumption is tenable in modelling catchment processes.

However, what is the alternative? We cannot represent epistemic errors explicitly because if we knew how to represent them, they would no longer be epistemic. It is perhaps therefore necessary to focus on the expression of being fit for purpose with respect to past performance. What should be our expectations of a model that would be considered fit for purpose? We would wish it to have the functionality not only to be consistent with past observations but also to predict future conditions (although we cannot test the latter until the future evolves). We would not expect it to fit every past observation precisely; it is, after all, an approximation, and the input data and evaluation observations also are subject to epistemic uncertainties. However, consistency does imply adequate performance after allowing for potential errors in the available data. So how close to an observation does a prediction need to be for a model to be accepted as fit for purpose? Can the limits of acceptability be defined, given only the past performance, some available observations, and a knowledge of the time and space scales required for a particular purpose (Smith, 2000, 2006; Beven, 2010)? How far should failure on a single measure of acceptability lead to rejection of an otherwise acceptable model? That might be a rogue observation, or it might be a critical observation that would lead to re-evaluation of the model concepts.

There is the possibility of multiple representations satisfying some limits of acceptability. There also is a possibility that none of the representations will prove acceptable (e.g. Beven, 2006). Virtual observatory visualizations will need to convey such ambiguity and imprecision in a way that users can understand so that they are empowered to make informed decisions, given the limited realism of what they see before them. This will be a challenge, as it is already a challenge in presenting the results of the ensemble of available global climate models, when all the available models are subject to significant epistemic uncertainties (and often have systematic errors larger than the expected signal) (Beven, 2011; Smith and Stern, 2011). The growth of scientific computing in the second half of the 20th century admitted many instances of overconfidence in the quantification of environmental systems, which led to false precision and (undoubtedly) some poor decision making. The challenge is to avoid similar claims of over-realism in the virtual observatories of this century.

It is still the case that very few studies in catchment science have posed the question of model evaluation in this way. Yet it seems to be critical as modelling moves towards virtual observatory platforms and models of everywhere. There is some evidence that it might be important to involve stakeholders with local

knowledge into this type of framework; they will sometimes be able to identify model inadequacies (e.g. Beven, 2007; Lane *et al*., 2011). However, this is also really a collection of science problems: of how to define assumptions about input errors in setting limits of acceptability for different applications; of how to evaluate all the available models that might be consistent with those limits of acceptability even given cloud computing resources; of how to define observational programs for testing those models as hypotheses as a way of constraining the uncertainty in the simulated outcomes; and of how to use the outcomes within a decision-making framwork. Considering these questions might actually provide a way of doing hydrological science within virtual observatory representations of hydrological realities.

References

Beven KJ. 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* 320: 18–36.

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

Beven KJ. 2009. Environmental Modelling: An Uncertain Future? Routledge: London.

Beven KJ. 2010. Preferential flows and travel time distributions: defining adequate hypothesis tests for hydrological process models. *Hydrological Processes* 24: 1537–1547.

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

Beven KJ, Cloke HL. 2012. Defining Grand Challenges in hydrology: A comment on Wood *et al*. (2011) Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, W01801. DOI: 10.1029/2011WR010982.

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

Buytaert W, Beven KJ. 2011. Models as multiple working hypotheses: Hydrological simulation of tropical alpine wetlands. *Hydrological Processes* 25: 1784–1799.

Herschy RW. 2009. Streamflow Measurement, 3rd edn. Taylor and Francis: Abingdon.

Howson C. 2003. Hume's Problem: Induction and the Justification of Belief. Clarendon Press: Oxford.

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

Lorenz E. 1985. The growth of errors in prediction. In Turbulence and Predictability in Geophysical Fluid Dynamics, Ghil M (ed.). North Holland: Amsterdam; 243–265.

Parker WS. 2010. Whose Probabilities? Predicting Climate Change with Ensembles of Models. *Philosophy of Science* 77(5): 985–997.

Smith LA. 2000. Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems. In Nonlinear dynamics and statistics, Mees AI (ed.). Birkhauser: Boston; 31–64.

Smith LA. 2006. Predictability past predictability present. In Predictability of weather and climate, Palmer T, Hagedorn R (eds.). Cambridge University Press: Cambridge, UK; 217–250.

Smith LA, Stern N. 2011. Uncertainty in science and its role in climate policy. *Philosophical Transactions of the Royal Society A* 369: 1–24.

Westerberg I, Guerrero J-L, Seibert J, Beven KJ, Halldin S. 2011. Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrol Process* 25: 603–613, DOI: 10.1002/hyp.7848.

Wood EF, Roundy JK, Troy TJ, van Beek R, Bierkens M, Blyth E, de Roo A, Doell P, Ek M, Famiglietti J, Gochis D, van de Giesen N, Houser P, Jaffe P, Kollet S, Lehner B, Lettenmaier DP, Peters-Lidard C, Sivapalan M, Sheffield J, Wade A, Whitehead P. 2011. Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research* 47: W05301. DOI: 10.1029/2010WR010090.