C. R. Geoscience 344 (2012) 77–88

Contents lists available at SciVerse ScienceDirect

Comptes Rendus Geoscience

www.sciencedirect.com

Hydrology, environment

Causal models as multiple working hypotheses about environmental processes

Les modèles causals comme hypothèses multiples des processus environnementaux

Keith Beven [a,*,b,c,d]

[a] *Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK*
[b] *Department of Earth Sciences, Geocentrum, Uppsala University, Uppsala, Sweden*
[c] *Centre for the Analysis of Time Series (CATS), London School of Economics, London WC2A 2AE, UK*
[d] *ENAC, laboratoire d'écohydrologie, EPFL, Lausanne, Switzerland*

ARTICLE INFO

Article history:
Received 14 December 2011
Accepted after revision 17 January 2012
Available online 9 March 2012

Presented by Ghislain de Marsily

Keywords:
Environmental prediction
Model evaluation
Limits of acceptability
Modelling philosophy
Critical experiments
Epistemic uncertainties

Mots-clés :
Prédiction environnementale
Évaluation de modèles
Limites d'acceptabilité
Philosophie de la modélisation
Expérimentations critiques
Incertitudes épistémiques

ABSTRACT

The environmental modeller faces a dilemma. Science often demands that more and more process representations are incorporated into models (particularly to avoid the possibility of making missing process errors in predicting future response). Testing the causal representations in environmental models (as multiple working hypotheses about the functioning of environmental systems) then depends on specifying boundary conditions and model parameters adequately. This will always be difficult in applications to a real system because of the heterogeneities, non-stationarities, complexities and epistemic uncertainties inherent in environmental prediction. Thus, it can be difficult to define the information content of a data set used in model evaluation and any consequent measures of belief or verisimilitude. A limit of acceptability approach to model evaluation is suggested as a way of testing models, implying that thought is required to define critical experiments that will allow models as hypotheses to be adequately differentiated.

RÉSUMÉ

Le modélisateur de systèmes environnementaux doit faire face à un dilemme. La science demande souvent que de plus en plus de représentations des processus soient incorporées dans les modèles (particulièrement pour éviter la possibilité de commettre des erreurs d'omission, lors des simulations). L'évaluation de représentations causales dans les modèles environnementaux (comme hypothèses multiples en ce qui concerne le fonctionnement des systèmes environnementaux) sous-entend la spécification des conditions aux limites et des valeurs des paramètres du modèle. Ceci sera toujours difficile pour les applications à des systèmes réels en raison des hétérogénéités, des non-stationnarités, des complexités et des incertitudes épistémiques inhérentes aux simulations environnementales. Il est ainsi ardu de définir l'information dans un jeu de données utilisé pour l'évaluation de modèles et les mesures de croyance ou de verisimilitude qui en découlent. Une méthodologie fondée sur des limites d'acceptabilité est suggérée pour l'évaluation des modèles, mais ceci implique qu'il faille définir des expérimentations critiques qui permettront de différencier suffisamment les modèles entre eux, considérés comme étant des hypothèses.

[*] Corresponding author.
E-mail address: k.beven@lancaster.ac.uk.

1631-0713/$ – see front matter © 2012 Published by Elsevier Masson SAS on behalf of Académie des sciences.
doi:10.1016/j.crte.2012.01.005

1. Introduction

There are many decisions in the domain of environmental science that are perceived as requiring quantitative predictions to enable ranking of different management scenarios. This implies proposing causal models about environmental processes in a context of open environmental systems subject to significant uncertainties. While some predictions of this type have become routine (such as in weather forecasting) in many other cases, the act of prediction comes close to crossing the border between science and the "trans-science" of Weinberg (1972). Certainly, the limitations of such predictions are not always well understood by those who wish to use them to inform decisions.

Quantitative predictions require models that are intended as formal representations of causal (process) relationships, even if such relationships are expressed in stochastic or fuzzy form. These formal representations, however, are often gross simplifications of the *perceptual model*[1] of the relevant processes and causalities. This is the case even where the formal model is intended to represent our current understanding of the physics (and/or geochemistry and/or ecology). There are always complexities, non-stationarities and heterogeneities that we can perceive as possibly having a significant impact on the system response but which are not included in the formal model because there is no agreed mathematical representation, or which are treated as a simplified "sub-grid" parameterisation in a larger scale model (as, for example, in atmospheric models used from weather forecasting to climate change predictions).

These approximate model representations are then supplied with incomplete and imperfect initial and boundary condition data to generate quantitative predictions. Any single run of the most detailed global climate model is of this type. It includes causal relationships (expressed as partial differential equations) to represent the fluxes of mass and energy in the atmosphere and oceans. It also includes many different *approximately* causal parameterisations to allow for sub-grid processes (energy dissipation, land surface to atmosphere fluxes of vapour and heat, cloud formation, rain generation, aerosol dispersion...). It is normally run deterministically, to produce a single set of outputs (though recently limited ensemble predictions have been run by the Hadley Centre as part of the UKCP09 product[2] with realisations generated using a "stochastic physics"; ensemble predictions are also being used in other areas of environmental modelling to assess prediction uncertainties, Beven, 2009).

This is one common feature of causal models in the environmental sciences[3]. Approximate implementations, driven by approximate boundary and auxiliary conditions will necessarily produce approximate results and different implementations (different model structures, parameter sets or ways of defining initial and boundary conditions) will produce different results. These different implementations can then be considered as multiple working hypotheses about the way the system functions (note that a similar argument applies to stochastic or fuzzy representations, even if the resulting predictions are not deterministic). The issue that then arises is whether, given the different sources of uncertainty in the modelling process, we can differentiate one hypothesis from another in trying to refine the science (and consequently have more faith in the model predictions). We can again use the prediction of global climate change as an example. The IPCC (2007) (Inter-governmental Panel on Climate Change) reported on the outcomes from a number of different deterministic global climate models that could be considered as different representations (multiple working hypotheses) about how the global climate system works. The different implementations produce different results (and for predicted variables such as rainfall, different implementations can produce very different results in some parts of the globe). In this case, rather than decide on whether one model hypothesis is more realistic than another, the range of outcomes from different models has been presented. They have also been presented as conditional scenarios because of the real epistemic uncertainty about future emission forcing as well as the approximate nature of the models themselves.

The concept of testing causal models of environmental systems as hypotheses raises a number of interesting philosophical issues that will be discussed in what follows. In particular we will consider the paradox of model complexity; the relationship between degrees of belief and verisimilitude; defining belief as likelihood for models of complex environmental systems in the face of epistemic uncertainties; the compromise between acceptance and rejection of models, and the role of critical experiments in testing competing models as hypotheses of system response.

2. The paradox of model complexity

In predicting the impacts of future change on environmental systems, it is necessary to be careful about defining the process representations relevant to the expected future states of the system. A model (as hypothesis) that does well in predicting the response of a system under current conditions might not do well in predicting the future

[1] The term perceptual model is used here to indicate a qualitative set of perceptions of how an environmental system works. As such it can include complexities and ambiguities that cannot be easily incorporated into a formal model producing quantitative predictions (see Beven, 2012).

That is not to say that the perceptual model is itself an adequate appreciation of the complexity of the real system which may not yet be properly understood even qualitatively.

[2] See http://ukclimateprojections.defra.gov.uk/content/view/868/531/index.html, Murphy et al., 2007; Rougier and Sexton, 2007).

[3] Note that I am using causal here in the sense as being defined by process descriptions that purport to represent causal linkages rather than models developed directly from observational data. It is worth noting, however, that an argument can be made that models developed directly from observational data might be more robust in prediction than deductive models constrained by prior conceptions (see for example, Young, 1998, 2001, 2003, 2011; Young and Beven, 1994; Young and Ratto, 2009).

conditions. This might be because the boundary conditions change to induce a different type or range of response beyond that for which the model has been calibrated. It might be because the future change invokes changes in the effective parameter values in the model (without a significant change in model structure). It might be because the future changes the nature of the process mechanisms involved in a way that implies the need for a change in model structure.

Thus there is a *deductive* argument to include as many processes as perceived as being important under both current and future conditions in a predictive model to avoid model failure as a result of missing processes. Most modellers want to be realist in the sense of defining a "correct" model representation of the system under study (see discussion of Beven, 2002a). The missing processes argument also fulfils the purpose of models as a way of providing a framework for formalising scientific understanding. Different scientists might well make different choices about what processes are important, and about how those processes might best be represented. While some processes are certainly better understood than others, there are enough uncertainties in describing the processes that, as in the case of global climate models, such choices produce a multitude of model implementations (or, effectively, competing hypotheses about system function).

However, the introduction of more processes means that because many process representations are based on *inductive* empirical functions, there are more model parameters to be identified. The deductive argument then conflicts with the much discussed inductive problem of fitting a parsimonious model to represent a set of data while avoiding over-fitting. This is particularly the case when functional representations that have been developed at small scales or laboratory scales are used at larger time and space scales for practical applications. A hydrological example is the widespread use of Darcy-Richards equation to describe unsaturated flow in soils where the nature of the unsaturated flow relationships results from the particular conditions of the laboratory experiments carried out by Richards (1931) in which increasing air pressure was used to reduce the saturation of the soil. This ensures that the larger pores stay empty, something that is not necessarily the case in field soils. The representation is based on the wrong experiment and the wrong physics (though Richards never claims in his paper that his equilibrium process representation will hold generally for real soils). In addition, heterogeneity of soil characteristics in the field means that, given the non-linearity of the unsaturated flow relationships, even if local fluxes are additive, the local parameters do not average linearly. Thus, the parameter values determined at the small scale are then not necessarily appropriate as *effective* parameter values that are required in a model to get good predictions at larger scales, compensating for small scale heterogeneities and other model and boundary condition deficiencies. Because in many cases the process representations are nonlinear, this can be a problem even for some of the most accepted process equations when there are any sub-grid small-scale heterogeneities or temporal non-stationarities in the real system.

Another example is the widespread use of the Manning equation to describe velocities in river flows. What is now known as the Manning equation was rejected by Manning in his original analysis (Manning, 1891) in favour of something more complicated (and dimensionally correct). It was also based on an analysis of what is called uniform flow (in which velocity vectors do not change in the downstream direction or in time) but is now widely applied in hydraulic models to represent friction losses for non-uniform, gradually-varying flows. Furthermore, literature values of the Manning roughness coefficient are normally back-calculated from measurements in single river cross-sections. What is needed for prediction is normally a roughness representing a whole river reach (with all its geometrical complexity and heterogeneities. Again, a local representation is being used as a larger scale parameterisation and *effective* values of the parameters are required.

In both these hydrological cases (and similar cases in other branches of environmental modelling), the use of such empirical functional relationships would not necessarily be a problem if there were measurement techniques available to evaluate the process equations and model predictions at the scales at which the models are being applied (at least under current conditions: the determination of effective parameter values for future conditions would still be a problem). In nearly every practical application of environmental models, however, this is not the case. It is then necessary to infer effective parameter models by a process of calibration or history matching, in which case the possibility of interaction between process representations in providing good fits to the available observations assumes greater significance. This means that model testing will have limited power. Not only may it be difficult to differentiate between different model structures as hypotheses but it may also be difficult to differentiate between different sets of effective parameter values within a model structure, a problem that is compounded by both aleatory and epistemic uncertainties in the input or forcing data used to run the model. This is what Beven (Beven, 1993, 2006a; Beven and Freer, 2001) has called the equifinality problem (after Von Bertalanffy, 1968).

Equifinality, as used in this sense, is intended to express the concept that there may be many different competing model structures and parameter sets within a model structure that might give equally acceptable results when compared with observations (Beven, 2006a). The models considered as acceptable or behavioural might, however, give quite different predictions. While the concept is similar to the use of the terms ambiguity, non-identifiability, and non-uniqueness, it has been chosen to reflect the fact that this is a generic problem in environmental modelling, not simply a question of underdetermination in determining *the* model of the system. Equifinality implies that deciding on an "optimal" model in calibration may be a poor strategy, even if residual errors are accounted for in some statistical framework.

The paradox of model complexity then arises because the greater the process understanding that is included in the model, the more parameter values that must be defined. Thus, the greater the potential for equifinality of models where the available observations cannot differentiate between different models as hypotheses. Even a hypothetically “perfect” causal model of the processes (which would necessarily have a large number of degrees of freedom in calibration) would not be immune to this paradox because in any practical application, it is necessarily forced by imperfect input data and compared with imperfect observations of system of response.

So we wish to make models more complex in order to make their predictions more realistic, but we might find it difficult to differentiate between models as hypotheses in practical applications. In environmental modelling, however, there is no possibility of ultimate resort to the Popperian solution of falsification since, as our models are necessarily approximate and driven by necessarily approximate boundary conditions, they can all be falsified if we look closely enough (Beven, 1993, 2002a; Cartwright, 1999; Morton, 1993; Oreskes et al., 1994). There may also be important understanding to be gained from model failures (see Andréassian et al., 2010).

And yet, even if our models are wrong in detail, they might well reflect the dominant causal characteristics of the system and consequently be useful in making predictions for decision makers, even if those predictions are uncertain. They might, in a classic instrumentalist view, be “fit for purpose” in the sense of providing predictions of future observations within some limits of acceptability or statistical consistency (e.g. Hitchcock and Sober, 2004; Sober, 1999, 2002; but see also Mikkelson, 2006) noting that success in this sense will also depend on assumptions about future boundary conditions as well as model structure and parameter values as hypothesis. The question then is how we can tell if one model (as hypothesis) is more or less fit for purpose than another in the face of the different sources of uncertainty in the modelling process. This raises issues of degrees of belief and verisimilitude.

3. Degrees of belief and verisimilitude in causal environmental models

Testing models as hypotheses implies changing degrees of belief (often expressed as probabilities or possibilities) in a particular hypothesis or set of hypotheses as evidence is gathered. Testing hypotheses in the face of uncertainty is traditionally the domain of statistical theory and there are a variety of well-established methodologies for hypothesis testing in statistics when the errors can be reduced to having only random characteristics. Some have argued that probability theory is the *only* way of rigorously taking account of uncertainties (e.g. O’Hagan and Oakley, 2004).

There has been extensive discussion about the use of the Akaike (or other) information criterion as a way of determining appropriate model complexity and avoid over-fitting (e.g. De Vito, 1997; Forster and Sober, 1994; Hitchcock and Sober, 2004). Bayesian frameworks also allow prior judgements about model structures where modelling errors can be reduced to random characteristics (Goldstein and Rougier, 2006; Kennedy and O’Hagan, 2001; but see also discussions by Dowe et al., 2007; Forster, 1995; Howson, 2003). The same methods provide ways of estimating the probabilities of an outcome conditional on the model predictions that can be used in estimating prediction uncertainties for decision makers.

Bayesian methods are increasingly used in environmental modelling (see, for example, Beven, 2009; Clark, 2006), but practical experience suggests that the assumption that models can be evaluated in terms of random errors conditional on the model being considered correct is rarely valid in environmental modelling. They are often used, *as if* the modelling errors have random characteristics, even when the stochastic assumptions may be difficult to justify (e.g. Rougier and Sexton, 2007, for just one example) and it is more likely that the total model error is structured in non-stationary and non-statistical ways as a result of epistemic uncertainties arising from both model structure error and input error (see the discussion of non-ideal cases in Beven, 2006a and practical examples in Beven and Westerberg, 2011; Beven et al., 2011).

It is important to understand what is being suggested here. It is well known that both deterministic chaotic and stochastic (Hurst/Kolmogorov) processes can generate non-stationary characteristics from rather simple mathematical mechanisms that are well defined and stationary in the long term (e.g. Koutsoyiannis, 2011; Schertzer et al., 2010). Within a statistical framework, it is also perfectly possible to add components to the error model that account for structures such as heteroscedasticity, auto- and spatial correlation or functional non-stationarities (such as a sinusoidal seasonal variation) with the aim of leaving an aleatory residual component. Such components are not causal in themselves (although they might later be interpreted to modify one or more process representations) but rather describe data or model inadequacies in a functional way (e.g. Kennedy and O’Hagan, 2001). Here I am suggesting something rather different, an *expectation* that arbitrary epistemic uncertainties will result in a variability in model errors that is not consistent (in time or space) and that is also not random in the sense of being convergent to some asymptotic distribution of such errors *within the time frame or spatial domain of a modelling application*. Given very large data series, some more general behaviour for these errors might be identified but in real applications *we do not and cannot have the knowledge* of any underlying structure that might be provided by such quantities of data. That is why such errors can have the appearance of being arbitrary and must be treated as epistemic.

To give a hydrological example from rainfall-runoff modelling, even if we could be sure of having a realistic model of runoff generation processes, epistemic uncertainties arise in both the forcing rainfall data and the river discharge outputs that might be used in model evaluation. This is because raingauge networks are sparse, such that particularly for strongly convective rainfalls, the total volume of inputs might be poorly estimated in a way that is storm type, position and movement dependent. Such errors cannot be easily represented by simple statistical error models (albeit that there have been attempts to do so using geostatistical methods with inadequately estimated variograms). Similar arguments can be made for rainfalls estimated from radar systems. They are then processed nonlinearly though the rainfall-runoff model structure (all hydrological models must reflect the nonlinear effects of antecedent wetness on runoff generation) with an expectation that this nonlinear transformation will lead to non-stationarity in any heteroscedasticity or autocorrelation in the residuals. The model output is then compared with the river discharge measurements, except that, in general, it is not river discharge that is measured. It is water level in the river that is measured. This is then converted to discharge by means of an empirical rating curve based on point velocity measurements at the site. It is much more difficult to do such measurements at higher flows, particularly at flood levels. Thus the rating curve measurements are often extrapolated to higher discharges by fitting some site-specific functional form for the rating curve (alternatively hydraulic calculations are sometimes used but require some rather strong assumptions about how the effective roughness coefficient will change with level). This means that the extrapolated discharge data used to evaluate model predictions might also be subject to epistemic errors (e.g. Beven et al., 2011, 2012; Krueger et al., 2010; Mathevet and Garçon, 2010; Westerberg et al., 2011). Similar examples can be cited from other environmental domains, including commensurability problems of comparing point measurements to spatial integral model predicted variables over some landscape element (e.g. Beven, 1989, 2006a).

Such epistemic uncertainties are then a strong constraint on the use of formal statistical likelihoods in model evaluation. Under the assumption that the errors (after identifying appropriate model inadequacy functions as necessary) are random, formal likelihoods will normally lead to a very strong over-differentiation between models that actually have somewhat similar performance in terms of error variance and bias. This is a direct result of assumptions about the information content in a series of residuals as if they arise from purely random effects. In nearly all environmental modelling applications, this will not be the case: the residuals will have complex non-stationary structure and persistence as a result of the nonlinear processing of input errors and model structural errors (see Beven, 2006a; Beven et al., 2008).

This clearly has important implications for differentiating between models as causal hypotheses. If all models can be falsified but some might be useful in prediction how can we assess an appropriate ‘‘degree of verisimilitude’’ (in Popper's phrase) in making some allowance for epistemic uncertainties? This effectively poses a problem of compromise between rejection (all models can be falsified) and acceptability (but some may be fit for purpose in having utility in prediction). This compromise is fundamental in the development of theories of verisimilitude (see review by Niiniluoto, 1998) but it would seem to be difficult to make such a compromise objective in practice if formal statistical likelihoods are not valid in the face of epistemic uncertainties (although Niiniluoto, 1987, presents many different forms of similarity measures). Fitness for purpose then becomes a question of likelihood as subjective belief, conditioned on what the modeller considers to be important in terms of performance (see the discussion of this difficulty in respect of hydrological modelling in section 5 below).

4. Evaluation of causal model likelihoods

Howson and Urbach (1993) and Press and Tanur (2001) have illustrated the ways in which science can be interpreted as working in subjective Bayesian ways, combining prior beliefs with (sometimes selected) evidence to derive some posterior ranking of possible hypotheses. The subjectivity in this process is not a problem within a Bayesian framework. The original formulation of Bayes equation (Bayes, 1763) is couched in terms of combining prior likelihood beliefs about hypotheses with evidence expressed in terms of some odds expressing support for different hypotheses. There is no intrinsic reason why those odds should not be expressed subjectively, as well as any prior beliefs. Formal likelihood theory, however, and particularly attempts to formulate an objective Bayes theory, have attempted to quantify the information content of model residuals so as to minimise subjectivity. This then introduces assumptions about the random nature of model residuals in ways that may not be justified in real applications where, as we have noted earlier, the non-stationarity of residual properties may be almost guaranteed by the nonlinear model dynamics. Over-strong assumptions about the randomness of residuals will then lead to over-conditioning of the likelihood surface and overestimation of the likelihood ratios that might be used to differentiate one model as hypothesis from another (see, for example, Beven et al., 2008, where this is demonstrated even for a hypothetical case free of any model structural error).

In principle, any assumptions about the model residuals used to formulate a likelihood function can be checked (for good practice in this respect, see for example Engeland et al., 2005). This is, however, rarely done in actual applications and even when it is done, it is rare to question more than the summary statistics of mean bias, residual variance and low order autocorrelation for some model deemed as having maximum likelihood under the chosen error assumptions. Where residuals are clearly non-Gaussian, then they are often transformed (by using, for example, Box-Cox or meta-Gaussian transforms) so as to still take advantage of a standard (simple) likelihood formulation.

There is no doubt that in some cases such formal likelihoods might be wrong but still a useful approximation in giving models that give good predictions high likelihood, but in other cases there is a danger that over-conditioning may result in making a Type II error in giving a likelihood of close to zero to a model that would perform well in prediction just because of epistemic errors in the forcing data. If models are being tested as hypotheses this is tantamount to rejecting such a model. It is easily demonstrated that, by the very nature of formal statistical likelihood functions, two models with small differences in error variance and bias can have differences of many orders of magnitude in likelihood as more data are added in conditioning. Such enormous differences in likelihood seem totally unrealistic given that some of the sources of such errors are epistemic.

That is not necessarily an argument for *not* using a Bayesian statistical approach for such inference, only that the way in which likelihood functions are developed should be the subject of more study for cases where epistemic errors are significant (which might be *all* real-world environmental modelling applications, Beven, 2006a). More realistic likelihood functions should then reflect the real information content of a residual series if it has complex non-stationary characteristics arising from the interaction of input errors, model structural errors, and evaluation observation errors (which, while independent of the model, might also be structured in complex ways as a result of the physical nature of the observation, e.g. errors in discharge estimates when a river goes overbank in a flood). Since there is no easy way to separate out the effects of these different sources of error on the residual series (see discussion in Beven, 2005), there is also no easy way to estimate the real information content in the residuals and consequently an appropriate likelihood formulation.

Beven (2006a) has therefore suggested an alternative, non-parametric, approach to the problem of evaluating model likelihoods that is consistent with the equifinality thesis introduced above. This approach is based on specifying limits of acceptability before running the model. The approach has been applied in rainfall-runoff modelling (Freer et al., 2004; Liu et al., 2009), hydraulic modelling (Pappenberger et al., 2007), water quality modelling (Dean et al., 2009; Page et al., 2007) and mixing models (Iorgulescu et al., 2007). Ideally, such limits of acceptability can be defined for all the individual observations with which a model will be compared in testing. Models that provide results within the limits of acceptability will be retained for use in prediction, those that do not will be rejected. Within the range of the limits of acceptability, likelihood measures can also be specified that, when combined over all observations (not necessarily multiplicatively as in Bayes), can be used to weight the predictions of the individual models in the acceptable set. This represents an extension of the Generalised Likelihood Uncertainty Estimation (GLUE) methodology first introduced by Beven and Binley (1992).

This approach replaces the need to make assumptions about the random structure of the residuals in the formal likelihood method, to a need to specify limits of acceptability. The starting point in doing so is the uncertainty in the observations with which a model will be compared, since we should not expect a model to provide predictions with greater accuracy than the expected observation errors. This is not in itself sufficient, however, since a good model might still be rejected on the basis of observation error alone if it is driven with poor input data. Thus, the limits of acceptability need to reflect the possibility of different sources of input error in some way. Such errors may also not be simple or random in nature. In hydrological models, for example, two types of inputs are required–precipitation inputs (both rain and snow where appropriate) and estimates of evapotranspiration loss back to the atmosphere (often derived from weather measurements at a point or remote sensing images by another interpretative model). These might also be subject to multiple sources of uncertainty and exhibit quite different structures. This is true even, for example, of simple remote sensing images of surface temperature (e.g. Santer et al., 2011). As in the case of rating curves to quantify discharge discussed earlier, this role of interpretative models in defining "observations" is often neglected in model applications (e.g., Beven et al., 2012).

While the limits of acceptability approach does encourage the consideration of all these different sources of uncertainty in the modelling process, there does not seem to be any easy way around making an assessment of the effects of input error on limits of acceptability for applications where there can be no independent assessment of the nature of input errors (which again must be very nearly all real applications of environmental models). In principle, given enough information, we could evaluate such errors objectively but we do not have that information so these errors are certainly epistemic and should not be treated *as if* aleatory. We should then *expect* that such errors might have non-stationary characteristics but the only evidence that a particular part of an input data series might be in error is normally that a model that is acceptable elsewhere does not produce good results following that particular input. Such evidence is, of course, highly conditional on the possibility that the model itself might be in error and would not in any case produce good results in that particular period. Testing models as hypotheses within this context would appear to be challenging and requiring further methodological developments (see, for example, the vagueness measures suggested in Lawry, 2006, 2008).

5. Compromise between model acceptance and model rejection: a hydrological example

Hydrology is a difficult science. It deals with nonlinear catchment (and global) systems subject to naturally highly variable (and non-stationary) boundary conditions. It does not have measurement techniques to assess all the relevant variables at the different scales of interest and consequently tends to use causal theory developed at laboratory scales as if it applied at larger scales even though there are reasons to doubt the validity of this approach (see, for example, Beven, 2006b and discussion of the Darcy-Richards and Manning equations above). It has to deal with the fact that every catchment is unique in its detailed characteristics and therefore any causal model will have some parameters that need to be adjusted or calibrated to allow for this uniqueness (Beven, 2000; Beven et al., 2002; De Marsily, 1994; Ganoulis, 1996). In these features it is quite representative of the environmental sciences in general (see the general discussions in Beven, 2009; Morton, 1993). The fundamental equation in hydrology, that can be applied at any scale, is the water balance equation. The water balance equation can also be supplemented by energy balance and momentum balance equations (e.g. Reggiani et al., 1998, 1999, 2000). Since these are forced input-output systems, there is an implied causality in these balance equations, since changes in state variables and outputs of the system are a direct response to the input terms. Thus, for the water balance for a catchment area, A,

$$A\,P\,\Delta t = Q\,\Delta t + A\,E\,\Delta t + G\,\Delta S + A\;\;\Delta S \tag{1}$$

where P is precipitation input per unit area (units of depth per unit time) over a time step Δt, Q is the measured discharge from the catchment (units of volume per unit time), E is loss to the atmosphere due to evapotranspiration per unit area (units of depth per unit time), G is unmeasured subsurface outflows (units of volume per unit time) and ΔS is the change in bulk storage (units of depth per unit area).

Thinking about each of the terms in this most basic of causal equations, some of the reasons why hydrology is a difficult science become immediately apparent. It is well known that the effective catchment area measured at the land surface may not be the same as the subsurface catchment area. The precipitation inputs can only be measured with some error (generally greater for snow inputs than rainfalls). Some inputs (such as the deposition of water on vegetation in mountain clouds) are not easily measured at all.

For the output terms, river discharge is generally not measured continuously but, as noted earlier, measurements of water level are converted to discharge using a “rating curve” developed from a small number of discharge measurements. It therefore also is associated with some (potentially epistemic) error, especially for the highest and lowest flows. Actual rates of evapotranspiration are controlled by weather variables (net radiation, temperature, humidity, wind speed) but also by the availability of water in the soil and by plant physiology in complex ways. There are reasonable measurement techniques at a point (by eddy correlation methods) or transects (by laser sounding or scintillometry) but not for spatially integrated values over a catchment area. Potential outputs via deeper subsurface flow pathways cannot be measured and are therefore normally assumed zero, but this might also introduce an error. Finally, there are techniques for point measurements of changes in storage in the soil profile, or remote sensing methods that provide estimates of the storage in the uppermost soil (at least in optimum circumstances) but these are subject to their own uncertainties and are not sufficient to get an adequate estimate for the spatially integrated change in storage.

Thus, even the most fundamental equation in hydrology is subject to significant error in each of its terms. If the water balance equation would be proposed as a hypothesis in hydrology it clearly cannot be proven without allowing for such uncertainties (Beven, 2005; see also the arguments of Cartwright, 1999, with respect to Newton's 2nd law of motion as applied to real problems). Similar arguments can be applied to the energy balance and momentum balance equations. Most causal models, however, assume that these balances are met de facto (as in the catchment theory outlined in Reggiani et al., 1998, 1999, 2000), although some operational models do allow parameters to be calibrated that modify some of these terms (e.g. bias correction in the inputs either for a whole period or on an event by event basis, or some “deep percolation rate” parameter to allow for the term G and errors in other terms that might lead to inbalance). Such modifications are not, however, based on directly on measurements and this will then allow trade-offs between ways of satisfying the water balance constraint (e.g. Le Moine et al., 2007).

Yet, so far, we have not even started to think about representing the causal processes that control the dynamic response of the system: i.e. how, given a particular state of the system as a pattern of water storage, the inputs will lead to changes in the output boundary fluxes of Q, E and G. This is what Reggiani et al. call the closure problem (see also Beven, 2006b, 2012), and it is essentially where the very many available hydrological models differ in their underlying equations and dynamics. Why are there so many hydrological models available? Precisely because each has some adjustable parameters that can be calibrated in each application and there is enough uncertainty in the terms of the water balance equation that no hydrologist expects perfect agreement between predictions and observations.

But this then raises an interesting question about testing models as causal hypotheses about the catchment system response. Is it indeed possible to say if one model is better than another in a useful way, given the uncertainties outlined above? This is clearly a problem that is not limited to hydrology, but will also hold in many different environmental applications, and particularly applications such as sediment transport, water quality and aquatic ecology that depend on predictions of water fluxes.

To take a very simple example, in considering the response of a catchment system we could propose to test a hypothesis that G is zero because a particular catchment is underlain by a relatively impermeable bedrock (though even impermeable bedrock is sometimes associated with fracture lines that can provide sufficient storage of water and permeability to support water uptake by tree roots, wells for local water supplies, and which might provide flow pathways for water out of a catchment). We can implement models with and without G but the paradox of complexity applies. We will need some representation of G (even if only a simple linear function of storage) with one (or more) parameters to be calibrated. It is very difficult to assess the values of such parameters *a priori* for any arbitrary application. Thus we would need to assess whether this additional parameter adds to how well the model can predict the catchment outputs. However, since the other terms in the balance equation are not known with any certainty, it is unlikely that any clear improvement will be found. Indeed we can go further and say that any model that does fit a calibration data set exceedingly well should be considered suspect, since it may be fitting to the errors in the data (the classic problem of over-fitting an over-parameterised model or a high order polynomial function).

The question therefore is how to proceed in the face of these difficulties, which can be posed in terms of how to assign relative probabilities (if only as measures of belief) to different causal process representations as a compromise between model acceptance and model rejection.

One approach is to work by deduction alone. This is equivalent to assigning only prior probabilities and in particular, a prior probability of 1 to a community consensus model and a prior probability of zero to all other representations. There is (as yet) no accepted example of such a consensus model at the catchment scale in hydrology, although there have been suggestions for community land surface models and community catchment modelling systems (the later allowing the user to choose from a menu of process representations). The theoretical framework of Reggiani et al. (*op. cit.*) was intended to be a step towards an agreed framework, and many hydrological modellers still hold to the blueprint outlined by Freeze and Harlan (1969) despite its demonstrated deficiencies (Beven, 1989, 2002b).

This particular example does not necessarily, of course, prove a general rule but is illustrative of the problem of deduction in these circumstances. How complex would such a model have to be to achieve a consensus; should different users be allowed to make choices of different subcomponents; how complex would it need to be to be generally fit for purpose; and how far is it possible to estimate the effective values of the parameters deductively in any particular application? This leaves open therefore the possibility of different consensus deductive model formulations (and different parameter sets within those formulations) depending on the group of scientists involved. Should instead, as suggested by one referee of this paper, model structures be subjected to a randomised experimental design (as might be possible within the Imperial College Rainfall-Runoff Modelling Toolbox of Wagener et al., 2004, the FUSE system of Clark et al., 2008, and the FLEX system of Fenicia et al., 2008) rather than relying on the causal reasoning of the scientist? But, given epistemic uncertainties, there is still no reason why models that are the most successful in calibration should continue to be the most successful in prediction. Similar arguments apply to models derived by induction (see footnote 3 above). The question of how models should be assessed in terms of probability as belief remains. Again this does not seem to be just a problem in hydrology, but applies to many forms of modelling of environmental processes.

There are two obvious ways of assessing probability as belief. The first would be to rely on expert opinion. But this will then tend towards circular reasoning, because of the difficult of finding scientists who are not committed to one modelling paradigm or another. The second would be to test the models as hypotheses against available data for a range of different circumstances. But, in fitting historical data, this is then subject to the paradox of complexity and the uncertainty of measurements as discussed above.

6. A third approach

A third approach has not often been tried in this area of science; to analyse the assumptions of different models as hypotheses and try to find critical experiments that might distinguish between different model structures or parameter sets (e.g. Beven, 2002a,b). Hypothesis testing is a classic approach in science; so why has it not been more widely used in environmental modelling? It would seem that there may be a general perception that any such critical experiments may not be feasible given the available measurement techniques and resources. Simply having more detailed point measurements in space may not be sufficient since we can allow for such variability in many model representations by allowing auxiliary conditions (initial states, local parameter values) to also vary in space (see Morton, 1993). An example is the matching of spatial water table measurements by allowing local soil hydraulic parameter values (e.g. Blazkova et al., 2002; Lamb et al., 1998). In this case, the assumption of homogeneous soil hydraulic parameters (which would be necessary *without* the availability of spatial observations) would result in the model predicting water tables less accurately at most observation points. Forster and Sober (1994) argue that any changes in such auxiliary conditions need to be justified by a sufficient gain in predictive power where, they suggest, sufficient might be measured statistically in terms of Akaike information. The use of such measures of information presupposes, however, that model residuals have a simple statistical structure which, as detailed more fully above, will not be the case for most environmental models subject to epistemic uncertainties in both forcing data and structure (see for example, Beven, 2010).

And yet, this still seems to be a positive way of approaching the problem of testing models as hypotheses even if it may be that rather than falsifying one hypothesis relative to another it will only be possible to change the relative probabilities, possibilities or vagueness measures (as belief) in one model relative to another by whatever means (see for example, Iorgulescu et al., 2007; Liu et al., 2009; Pappenberger et al., 2007). This will then produce a set of prior likelihoods for prediction for all those models that survive such an evaluation (the set of behavioural models within the GLUE methodology, Beven, 2006a, 2009, 2010). Note that this applies equally to models being evaluated at local scales and those being tested for performance at regional or global scales.

7. Causal models, hypothesis testing, and philosophy

Thus, the environmental modeller faces a dilemma. The science often demands that more and more process representations are incorporated into models (particularly to avoid the possibility of making missing process errors in predicting future response). Testing the causal representations in environmental models (as hypotheses about the functioning of environmental systems) then depends on specifying boundary conditions and model parameters adequately. This can be done in the laboratory, but this does not guarantee equal validity when applied to the real system because of the heterogeneities, non-stationarities, complexities and epistemic uncertainties inherent in environmental prediction. Thus, it can be difficult to define the information content of a data set used in model evaluation (where the resulting model residuals may not have simple statistical structure) and therefore to define adequate hypothesis tests with respect to observations made under current conditions. Predicting future change will necessarily be that much more difficult. The future inevitably involves some unknown unknowns (and, as the theory of nonlinear dynamics suggests, the possibility of mode switches in behaviour).

This disjunctive nature of the possibility of prediction is an instance of Hume's problem of induction (see the discussion of Howson, 2003) most recently popularised by Taleb (2010) as the "black swan" problem. Here, time is of the essence (both in the past and future). We have a set of models that can be shown to be (more or less) behavioural given current observations, even if we suspect that they cannot *all* be realistic when more than one model or parameter set is consistent with observations and when the specification of past boundary conditions is uncertain (we cannot go back and check past estimates of boundary conditions and forcing data, but model calibration is always conditional on these estimates). We would also want our models to be instrumentally behavioural in future but know very well that success in this sense will depend heavily on any future estimation of the boundary conditions, boundary conditions that might be quite different (or with different error characteristics) than seen in model calibration. This issue was first raised in hydrology by Stephenson and Freeze (1974); discussed later by Konikow and Bredehoeft (1992) and Anderson and Woessner (1992); then considered by professional philosophers in Oreskes et al. (1994). We also cannot be sure that such nonlinear systems will not exhibit mode changes.

As with any inference from induction, we will not know if a model is instrumentally behavioural or not until the future evolves. This suggests that, as a result of epistemic uncertainties, we should *expect* surprises in prediction. This might include following good modelling practice in carrying out a split record test model evaluation (Klemeš, 1986; Refsgaard, 1997; Refsgaard and Henriksen, 2004) since the epistemic errors of the forcing data and observations used in model calibration might be quite different to those in the additional evaluation period.

A common strategy to deal with this is to treat future predictions only as uncertain scenarios. We are not sure of these future boundary conditions (and all those for global change models were estimated before the 2008 global financial crisis); they are presented as potential future scenarios of unknown probability. Many of these scenarios may not turn out to be realistic as the future evolves but the model itself does not need to be questioned if it has already been declared to be fit on the basis of consistency with past observations according to some conditions of acceptability (perhaps relaxed in the case of current climate models, although we can be sure that the models used in the next IPCC5 report will also have evolved further). At any point in time, therefore, the focus has to be on (conditional) hypothesis testing with respect to past observations, but we should bear in mind that what until now has appeared to be behavioural, might turn out to have been in some important sense non-behavioural all along.

The aims of the environmental modeller are both realist and instrumentalist in nature. They are realist in the sense of wishing to have some verifiable representation of how the real system works. They are instrumentalist in the sense of wishing to make predictions that will be consistent with future observations. It has been seen in the above discussion, however, that neither aim is readily achievable in practice. Thus, any degree of belief in a current model as a true representation of reality might be low, but still (consensually or empirically) stronger that that of other possible model representations. It has therefore been suggested that most environmental modellers hold a form of "pragmatic realist" philosophy while being predominantly instrumentalist in practice (Beven, 2002a). A pragmatic realist knows very well that models are code running on silicon processors but would like to think that the variables in a model represent real water, or real populations, or real chemical masses, or real atmospheric gases. In particular, we would like to think that as process understanding improves, the models will get closer to representing these real variables.

The difficulties of testing models as hypotheses outlined above seem to impose some strong constraints on how far this process can proceed since in postulating and testing improvements to model structures based on improved understanding it might be very difficult to show that a particular model performs better than another given the uncertainties in inputs and evaluation observations. This will remain the case unless critical measurements can be made that can specifically differentiate between one model as hypothesis and another (or at least allow the probabilities/verisimilitude measures associated with particular models to be modified in a way that reflects the limited information content in the uncertain boundary conditions and observations with which predictions will be compared).

It would appear, therefore, that both realist and instrumentalist ambitions may be thwarted in real applications. It may be that we can only aspire to the more limited ambition of assessing degrees of verisimilitude (see Section 3 above) to the available observations, and that any such assessment is likely to be rather subjective, whether done by a consensus of experts or by the choice of some quantitative likelihood measure (albeit that it is not clear how to represent the information content of the available observations or model residuals and therefore what is an appropriate belief measure on which to base any assessment of verismilitude). That is how environmental modelling current works in practice, whatever the philosophical leanings of its practitioners. This leaves many environmental modelling activities verging on "trans-science" (see Philip, 1980). Indeed, while there is a demand to be internal consistent in this type of pragmatic realism, there should be no expectation (at least, as yet) of bivalent correspondence with reality. While Taylor (2006, see also Groff, 2004) suggests that this is, in effect, an anti-realist stance, that is not how it is thought of by practitioners. It is certainly not evident in the way that many models are marketed for real applications (that may not consider any representation of uncertainty).

This is the conclusion of my active engagement in the process of environmental modelling that has involved consideration of multiple sources of uncertainty and complex error structures in practical applications. It is a conclusion that many environmental modellers prefer not to think about, in part because it inevitably undermines confidence in future predictions, especially when such predictions are being used as evidence for framing future policy (as in the case of the IPCC global climate predictions). In such circumstances, it is generally better to argue that the predictions provide the best scenarios currently available and that we expect to be able to improve them further (in terms of both realism and instrumentalist performance) as both understanding and computer power increases. The arguments here are not incompatible with that view but suggest that a framework is needed for taking account of different sources of uncertainty in testing models as causal hypotheses, and in encouraging the thoughtful search for critical experiments as a way of doing science in this important domain. Otherwise decisions might (or should) be made in other ways (see Beven, 2011).

8. A final reflection on the *as if* issue.

One of referees on this paper (quite rightly) asked what is the problem of treating modelling errors *as if* they are aleatory in nature? Treating a variable as if it is a (potentially complex) stochastic process is convenient, even if we know very well that the ultimate causes of the variability might not be random in nature. And, as noted above, such an approach can be objective in that the assumptions that are made about that variability can be checked for validity. He then contrasted such an approach with his perception of the subjectivity of the limits of acceptability approach suggested here.

I think that there are two issues that are compounded in these comments. One is the issue of how assumptions about error structures lead to formal likelihood functions when the errors are treated as aleatory. Thus, I would like to stress again that doing this in the standard way will overestimate the information content of a series of model errors when epistemic errors are important and therefore lead to over-conditioning of the model space. As noted earlier this is not necessarily an argument against using a Bayesian framework, but rather an indication that the specification of likelihood functions should be revisited. Howson (2003), for example, argues that Bayes provides the only useful philosophical framework for dealing with the problem of induction (without being specific about the definition of likelihood) while Tarantola (2005, 2006) points out that the L^2 norm originated by Gauss that is commonly used in formulating likelihood functions is only a choice that became popular because of its analytical convenience in the 19th and early 20th centuries. Other norms could be used that might be less sensitive to the effects of individual large epistemic errors (see, for example, the plot in Beven and Westerberg, 2011). However, if it is desired to reflect the role of epistemic error in reducing information content, the definition of a suitable function will then become less objective (*precisely because* of the lack of knowledge about the nature of the epistemic errors), even if any assumptions can still be checked.

The second (and complementary) issue is the supposed subjectivity of the limits of acceptability approach. I am not suggesting here that the limits of subjectivity should be chosen arbitrarily but rather that they should be developed as objectively as possible *before running the model*. I do not doubt that this requires new methodological developments but I actually see only one reason why such an approach should not be objective and that (again) is *precisely because* of the lack of knowledge about the nature of the epistemic errors, particularly the input errors discussed earlier. Such an approach would also remain objective in the same sense of being able to check the assumptions on which the choice of limits has been based.

Thus, lack of knowledge will remain a constraint on being fully objective. In fact, if we start to think more deeply about the nature of epistemic errors (which is the primary purpose of this paper) we should not *expect* to be able to be fully objective. This is not a new insight. In hydrology, it was implicit in the discussion of Stephenson and Freeze (1974) in discussing the problem of validating a causal model of hillslope hydrology when there was limited information (i.e. epistemic error) about the boundary conditions and internal states. It is, however, an insight that has largely been disregarded in hydrological modelling in the decades since Stephenson and Freeze. How best to be as objective as possible remains, however, an open question. It is to be hoped that the reader might at least allow that it is a question that is still worth asking and that requires new research.

Acknowledgments

Many people have contributed to the thoughts contained in this paper over many years, many of whom have radically disagreed with the views expressed but who have made me think more deeply about the issues. I would particularly like to thank Peter Young, Jonty Rougier, Paul Smith and Ida Westerberg for discussions on these issues over a considerable period of time, I would also like to thank Ghislain de Marsily for his invitation to develop this paper for C. R. Geoscience and the four reviewers, Alberto Montanari, Nicolas Le Moine, Adam Morton and Vazken Andréassian whose thoughtful responses to the original submitted version have led to an improvement in the final presentation.

References

Anderson, M.P., Woessner, W.W., 1992. The role of the post-audit in model validation. Adv. Water Resour. 15, 167–174.

Andréassian, V., Perrin, C., Parent, E., Bárdossy, A., 2010. Editorial–The court of miracles of hydrology: can failure stories contribute to hydrological science? Hydrol. Sci. J. 55 (6), 849–856.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. Phil. Trans. Roy. Soc. Lond. 53, 370–418.

Beven, K.J., 1989. Changing ideas in hydrology: the case of physically-based models. J. Hydrol. 105, 157–172.

Beven, K.J., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Adv. Water Resour. 16, 41–51.

Beven, K.J, 2000. Uniqueness of place and process representations in hydrological modelling. Hydrol. Earth Syst. Sci. 4, 203–213.

Beven, K.J., 2002a. Towards a coherent philosophy for environmental modelling. Proc. Roy. Soc. Lond. A460 (458), 2465–2484.

Beven, K.J., 2002b. Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modelling system. Hydrol. Process. 16, 189–206.

Beven, K.J., 2005. On the concept of model structural error. Water Sci. Technol. 52, 165–175.

Beven, K.J., 2006a. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36.

Beven, K.J., 2006b. The Holy Grail of scientific hydrology: $Q_t = H(\leftarrow S \leftarrow R)A$ as closure. Hydrol. Earth Syst. Sci. 10, 609–618.

Beven, K.J., 2009. Environmental modelling: an uncertain future? Routledge, London.

Beven, K.J., 2010. Preferential flows and travel time distributions: defining adequate hypothesis tests for hydrological process models. Hydrol. Process. 24, 1537–1547.

Beven, K.J., 2011. I believe in climate change but how precautionary do we need to be in planning for the future? Hydrol. Process. 25, 1517–1520, doi:10.1002/hyp.7939.

Beven, K.J., 2012. Rainfall-runoff modelling –the primer, 2nd edition. Wiley-Blackwell, Chichester.

Beven, K.J., Binley, A.M., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6, 279–298.

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

Beven, K.J., Westerberg, I., 2011. On red herrings and real herrings: disinformation and information in hydrol. inference. Hydrol. Process. 25, 1676–1680, doi:10.1002/hyp.7963.

Beven, K.J., Musy, A., Higy, C., 2002. Tribune libre : l'unicité de lieu, d'action et de temps. Rev. Sci. Eau 14, 525–533.

Beven, K.J, Smith, P.J., Freer, J., 2008. So just why would a modeller choose to be incoherent? J. Hydrol. 354, 15–32.

Beven, K., Smith, P.J., Wood, A., 2011. On the colour and spin of epistemic error (and what we might do about it). Hydrol. Earth Syst. Sci. 15 (2011), 3123–3133, doi:10.5194/hess-15-3123-2011.

Beven, K.J., Buytaert, W., Smith, L.A., 2012. On virtual observatories and modeled realities (or why discharge must be treated as a virtual variable). Hydrol. Process., doi:10.1002/hyp.9261.

Blazkova, S., Beven, K.J., Tacheci, P., Kulasova, A., 2002. Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): the death of TOPMODEL? Water Resour. Res. 38, W01257, doi:10.1029/2001WR000912.

Cartwright, N., 1999. The dappled world: a study of the boundaries of science. Cambridge University Press, Cambridge, UK.

Clark, J.S., 2006. Why environmental scientists are becoming Bayesians. Ecol. Lett. 8, 2–14.

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H.V., Wagener, T., Hay, L.E., 2008. Framework for Understanding Structural Errors (FUSE): a modular framework to diagnose differences between hydrological models. Water Resour. Res. 44, W00B02, doi:10.1029/2007WR006735.

De Marsily, G., 1994. Tribune libre : quelques réflexions sur l'utilisation des modèles en hydrologie. Rev. Sci. Eau 7, 219–234.

De Vito, S., 1997. A gruesome problem for the curve-fitting solution. Brit. J. Phil. Sci. 48, 391–396.

Dean, S., Freer, J.E., Beven, K.J., Wade, A.J., Butterfield, D., 2009. Uncertainty assessment of a process-based integrated catchment model of phosphorus (INCA-P). Stochastic Environ. Res. Risk Assess. 23, 991–1010, doi:10.1007/s00477-008-0273–z.

Dowe, D.L., Gardner, S., Oppy, G., 2007. Bayes not bust! Why simplicity is no problem for Bayesians Brit. J. Phil. Sci. 58, 709–754.

Engeland, K., Xu, C.Y., Gottschalk, L., 2005. Assessing uncertainties in a conceptual water balance model using Bayesian methodology. J. Hydrol. 50, 45–63.

Fenicia, F., Savenije, H.H.G., Matgen, P., Pfister, L., 2008. Understanding catchment behavior through model concept improvement. Water Resour. Res. 44, W01402, doi:10.1029/2006WR005563.

Forster, M., 1995. Bayes and Bust: simplicity as a problem for a probabilist's approach to confirmation. Brit. J. Phil. Sci. 46, 399–424.

Forster, M., Sober, E., 1994. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. Brit. J. Phil. Sci. 45, 1–35.

Freer, J.E., McMillan, H., McDonnell, J.J., Beven, K.J., 2004. Constraining TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. J. Hydrol. 291, 254–277.

Freeze, R.A., Harlan, R.L., 1969. Blueprint for a physically-based, digitally simulated, hydrologic response model. J. Hydrol. 9, 237–258.

Ganoulis, J., 1996. Tribune libre : sur la modélisation des phénomènes hydrologiques. Rev. Sci. Eau 9, 421–434.

Goldstein, M., Rougier, J., 2006. Bayes linear calibrated prediction for complex systems. J. Amer. Stat. Assoc. 101, 1132–1143.

Groff, R., 2004. Critical realism, post-positivism and the possibility of knowledge. Routledge, London.

Hitchcock, C., Sober, E., 2004. Prediction vs. accommodation and the risk of overfitting. Brit. J. Phil. Sci. 55, 1–34.

Howson, C., 2003. Hume's problem: induction and the justification of belief. Clarendon Press, Oxford.

Howson, C., Urbach, P., 1993. Scientific reasoning: the Bayesian approach, 2nd edition. Open Court, Chicago.

Iorgulescu, I., Beven, K.J., Musy, A., 2007. Flow, mixing, and displacement in using a data-based hydrochemical model to predict conservative tracer data. Water Resour. Res. 43, W03401, doi:10.1029/2005WR004019.

Inter-governmental Panel on Climate Change (IPCC), Climate Change 2007: The Physical Science Basis. Summary for Policy Makers, WMO: Geneva, 2007.

Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of mathematical models. J. Roy. Statist. Soc. D63, 425–450.

Klemeš, V., 1986. Operational testing of hydrologic simulation models. J. Hydrol. 31, 13–24.

Konikow, L.F., Bredehoeft, J.D., 1992. Groundwater models cannot be validated? Adv. Water Resour. 15, 75–83.

Koutsoyiannis, D., 2011. Hurst-Kolmogorov dynamics and uncertainty. J. Amer. Water Resour. Assoc. 47, 481–495, doi:10.1111/j.1752-1688.2011.00543.x.

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

Lamb, R., Beven, K.J., Myrabø, S., 1998. Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model. Adv. Water Resour. 22, 305–317.

Lawry, J., 2006. Modelling and reasoning with vague concepts. Springer, Berlin.

Lawry, J., 2008. Appropriateness measures: an uncertainty model for vague concepts. Synthese 161, 255–269, doi:10.1007/s11229-007-9158-9.

Le Moine, N., Andréassian, V., Perrin, C., Michel, C., 2007. How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study over 1040 French catchments. Water Resour. Res. 43, W06428, doi:10.1029/2006WR005608.

Liu, Y., Freer, J.E., Beven, K.J., Matgen, P., 2009. Towards a limit 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.

Manning, R., 1891. On the flow of water in open channel and pipes. Trans. Instn. Civ. Engrs. Ireland 20, 161–207.

Mathevet, T., Garçon, R., 2010. Tall tales from the hydrological crypt: are models monsters? Hydrol. Sci. J. 55, 857–871.

Mikkelson, G.M., 2006. Realism versus instrumentalism in a new statistical framework. Philos. Sci. 73, 440–447.

Morton, A., 1993. Mathematical models: questions of trustworthiness. Brit. J. Phil. Sci. 44, 659–674.

Murphy, J.M., Booth, B.B.B., Collins, M., Harris, G.R., Sexton, D.M.H.R., Webb, M.J., 2007. A methodology for probabilistic prediction of regional climate change from perturbed physiscs ensembles. Phil. Trans. Roy. Soc. Lond. A365, 1993–2008.

Niiniluoto, I., 1987. Truthlikeness. D. Reidel, Dordrecht.

Niiniluoto, I., 1998. Verisimilitude: the third period. Brit. J. Phil. Sci 49, 1–29.

O'Hagan, A., Oakley, J.E., 2004. Probability is perfect but we can't elicit it perfectly. Reliability Eng. Syst. Safety 85, 239–248.

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

Page, T., Beven, K.J., Freer, J., 2007. Modelling the chloride signal at the Plynlimon catchments, Wales using a modified dynamic TOPMODEL. Hydrol. Process. 21, 292–307.

Pappenberger, F., Frodsham, K., Beven, K.J., Romanovicz, R., Matgen, P., 2007. Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. Hydrol. Earth Syst. Sci. 11, 739–752.

Philip, J.R, 1980. Field heterogeneity: some basic issues. Water Resour. Res. 16, 443–448.

Press, S.J., Tanur, J.M., 2001. The subjectivity of scientists and the Bayesian approach. Wiley, Chichester.

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

Refsgaard, J.C., Henriksen, H.J., 2004. Modelling guidelines–terminology and guiding principles. Adv. Water Resour. 27, 71–82.

Reggiani, P., Sivapalan, M., Hassanizadeh, S.M., 1998. A unifying framework of watershed thermodynamics: balance equations for mass, momentum, energy and entropy and the second law of thermodynamics. Adv. Water Resour. 22, 367–398.

Reggiani, P., Sivapalan, M., Hassanizadeh, S.M., Gray, W.G., 1999. A unifying framework of watershed thermodynamics: constitutive relationships. Adv. Water Resour. 23, 15–39.

Reggiani, P., Sivapalan, M., Hassanizadeh, S.M., 2000. Conservation equations governing hillslope responses: physical basis of water balance. Water Resour. Res. 38, 1845–1863.

Richards, L.A., 1931. Capillary conduction of liquids through porous mediums. Physics 1, 313–333.

Rougier, J.C., Sexton, D.M.H.R., 2007. Inference in ensemble experiments. Phil. Trans. Roy. Soc. Lond. A365, 2133–2144.

Santer, B.D., Wigley, T.M.L., Taylor, K.E., 2011. The reproducibility of observational estimates of surface and atmospheric temperature change. Science 334 , doi:10.1126/science.1216273.

Schertzer, D., Tchiguirinskaia, I., Lovejoy, S., Hubert, P., 2010. No monsters, no miracles: in nonlinear sciences hydrology is not an outlier! Hydrol. Sci. J. 55, 965–979.

Sober, E., 1999. Instrumentalism revisited. Critica 31, 3–38.

Sober, E., 2002. Instrumentalism, parsimony, and the Akaike framework. Philos. Sci. 69, S112–S123.

Stephenson, G.R., Freeze, R.A., 1974. Mathematical simulation of subsurface flow contributions to snowmelt runoff, Reynolds Creek, Idaho. Water Resour. Res. 10, 284–298.

Taleb, N.N., 2010. The Black Swan, Second Edition. Penguin, London.

Tarantola, A., 2005. Inverse problem theory and model parameter estimation. SIAM, Philadelphia, PA.

Tarantola, A., 2006. Popper, Bayes and the inverse problem. Nature Physics 2, 492–494.

Taylor, B., Models, 2006. Truth and realism. OUP, Oxford, UK.

Von Bertalanffy, L., 1968. General Systems Theory. Braziller, New York.

Wagener, T., Wheater, H., Gupta, H.V., 2004. Rainfall–runoff modelling in gauged and ungauged catchments. Imperial College Press, London.

Weinberg, A., 1972. Trans-science. Minerva 10, 209–222.

Westerberg, I., Guerrero, J.L., Seibert, J., Beven, K.J., 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.

Young, P.C., 1998. Data-based mechanistic modelling of environmental, ecological, economic and engineering systems. Environ. Model. Softw. 13, 105–122.

Young, P.C., 2001. Data-based mechanistic modelling and validation of rainfall-flow models. In: Anderson, M.G., Bates, P.D. (Eds.), Model validation: perspectives in hydrological science. John Wiley & Sons, Chichester, UK, pp. 117–161.

Young, P.C., 2003. Top-down and data-based mechanistic modeling of rainfall-flow dynamics at the catchment scale. Hydrol. Process. 17, 2195–2217.

Young, P.C., 2011. Recursive estimation and time-series analysis. Springer-Verlag, Berlin.

Young, P.C., Beven, K.J., 1994. Data-based mechanistic modelling and the rainfall-flow nonlinearity. Environmetrics 5 (3), 335–363.

Young, P.C., Ratto, M., 2009. A unified approach to environmental systems modeling. Stochast. Environ. Res. Risk Assess. 23, 1037–1057.