On the concept of model structural error

K. Beven
Environmental Science, Lancaster University, Lancaster, LA1 4YQ, UK (E-mail: k.beven@lancaster.ac.uk)

Abstract A consideration of model structural error leads to some particularly interesting tensions in the model calibration/conditioning process. In applying models we can usually only assess the total error on some output variable for which we have observations. This total error may arise due to input and boundary condition errors, model structural errors and error on the output observation itself (not only measurement error but also as a result of differences in meaning between what is modelled and what is measured). Statistical approaches to model uncertainty generally assume that the errors can be treated as an additive term on the (possibly transformed) model output. This allows for compensation of all the sources of error, as if the model predictions are correct and the total error can be treated as “measurement error.” Model structural error is not easily evaluated within this framework. An alternative approach to put more emphasis on model evaluation and rejection is suggested. It is recognised that model success or failure within this framework will depend heavily on an assessment of both input data errors (the “perfect” model will not produce acceptable results if driven with poor input data) and effective observation error (including a consideration of the meaning of observed variables relative to those predicted by a model).

Keywords Error compensation; model calibration; model rejection; uncertainty

Introduction

Environmental models are wrong and are known to be wrong (Morton, 1993; Beven, 2002a,b). This means that in the application of environmental models it is necessary to deal with uncertainties associated with model structural error. This has long been recognised in principle. However, it is seldom considered in any practical application. It is still relatively rare that model applications are associated with an evaluation of the associated uncertainties. It is more often the case in applications that the model is applied as if it were correct and free of structural error.

There are good reasons for this, of course. The model is not the only source of error in the modelling process. There are, in addition, errors associated with inputs and boundary conditions, errors associated with the parameter estimates used in the model, errors associated with the observations with which the model will be compared (including commensurability errors for some variables, see below). There are also different forms of model structural error: errors due to the use of incorrect representations of processes in heterogeneous and non-stationary systems; conceptual errors arising from the representation of the characteristics of the processes; processes that are not represented because we are not sure how they should be represented; processes that are not represented because we do not yet know that they might be important; and implementation errors (numerical algorithms, choice of space and time discretisations) that might make effective parameter values scale and numerical algorithm dependent in ways that are not clearly understood in nonlinear systems.

There are, in consequence, numerous potential sources of error in the modelling process which will operate to produce error in the model predictions in nonlinear ways. The different sources may be more or less important in different applications. However, to assess the nature of the errors we have generally only comparisons with direct observations on the system that are limited in either space or time. Deconstructing the potential sources of the errors revealed by this comparison is an ill-posed problem, i.e. there are an
infinite number of ways in which input, structural and observation errors might interact to
give the same series of residuals in space or time.

Solving ill-posed problems of this sort requires simplifications. Traditionally, the main
methodology for making these simplifications to produce a tractable solution has been
within the framework of statistical inference, and most recently of Bayesian statistical
inference (see, for example, Kennedy and O’Hagen, 2001). This is not, however, the only
framework that might be considered and in what follows it will be argued that some
other ways of simplifying the problem might be worth considering (Beven, 2005).

The mathematics of model structural error
Consider a general representation of the modelling problem with error terms. At any
point in space or time \((x,t)\) we expect that

\[
O(x,t) + e_O(x,t) = M(\theta_e, I, e_I, x, t) + e_M(\theta_e, I, e_I, x, t) + e_r
\]  

(1)

where \(O\) is a vector of observations, \(M\) is a vector of modelled values, \(\theta\) is a vector of
model parameter values, \(I\) is a vector of inputs and boundary conditions, \(e_O\) is an obser-
vation error, \(e_I\) is an input error, \(e_M(\theta_e, I, e_I, x, t)\) is a model error and \(e_r\) is a random
error.

What is actually observable, on the other hand, is:

\[
e(x,t) = O(x,t) - M(\theta_e, I, e_I, x, t) = e_M(\theta_e, I, e_I, x, t) - e_O(x,t) + e_r
\]  

(2)

It is therefore evident that the various terms on the right hand side cannot be disaggre-
gated from the single observable \(e\) unless some very strong assumptions are made. The
usual assumption in statistical inference is that all the terms on the right hand side can be
treated as a lumped “measurement error” and represented by an appropriate statistical
model (e.g. additive Gaussian, autocorrelated noise).

However, while writing the model error \(e_M(\theta, I, e_I, x, t)\) as an additive term in (1) is
formally correct at any \((x,t)\), that does not mean that the function it represents is in any
way simple. The model estimate \(M()\) depends on parameter errors and input and bound-
ary condition errors in complex and nonlinear ways. This would be the case even if, for
some hypothetical case, the model could be assumed formally correct. In real appli-
cations, however, this is not the case. The model is wrong and is known to be wrong and
its deficiencies will interact nonlinearly with errors in parameters and input and boundary
conditions to induce nonstationarity and heteroscedasticity in this term.

There are other sources of error that are not easily distributed between the terms on
the right hand side of (2). In some cases, for example, the meaning of an observation of a
variable is not necessarily the same as the meaning of a model prediction of that same
variable at the same time and space. This is the commensurability problem (Beven, 2000,
2002a,b). It is a particular problem in distributed models where measurement technique
limitations mean that an observation taken at one spatial scale (i.e. soil moisture, water
table level) is not easily compared with a prediction at the model grid/element scale. This
could be considered either as part of the observation error or as part of the model error.
In that the terms of the right hand side of (2) add, it does not matter. However, any
attempt to identify model structural error does require that this be defined. Since a fair
evaluation of the model itself (and of model structural error) requires commensurability
at the prediction scale, this should perhaps be separated for clarity as an additional term
\[ e_C \text{ depending on the model discretisation, such that:} \]
\[ e(x,t) = e_M(\theta, e_0, I, e_I, x, t) - e_C(\Delta x, \Delta t, x) - e_O(x, t) + e_r \]  
\[ (3) \]

**The role of model calibration in the identification of model structural error**

Given the observed total model error on the left hand side of (3), model calibration usually involves some attempt to minimise the total model error in some way by changing \( M(\theta, e_0, I, e_I, x, t) \), since \( O(x, t) \) must normally be considered fixed. Since the four terms on the right hand side of (3) cannot be differentiated \textit{a priori}, model calibration normally requires a simplification of the problem.

The traditional approach to model calibration in hydrological modelling has been to simplify (1) to the form:

\[ O(x, t) = M(\theta, I, x, t) + e(x, t) \]  
\[ (4) \]

with the aim of minimising the total error in some way. This assumes that the effect of all sources of error can be subsumed into the total error series as if the model were correct and that the input and boundary condition data and observations were known precisely.

Furthermore, if the total error \( e(x,t) \) can be assumed to have a relatively simple form (or can be suitably transformed to a simple form) then a formal statistical likelihood function can be defined, dependent on the assumed error structure. Thus, an evaluation made for observations at a single site for total model errors can be assumed to have zero mean, constant variance, independent in time and Gaussian distribution. Here, the likelihood function takes the form:

\[ L(e|M(\theta, I, x, t)) = (2\pi\sigma^2)^{-T/2}\exp \left[-\frac{1}{2\sigma^2} \sum_{t=1}^{T} e_t^2 \right] \]  
\[ (5) \]

where \( e_t = O(x,t) - M(\theta, I, x, t) \) at time \( t \), \( T \) is the total number of time steps and \( \sigma^2 \) is the residual error variance. For total model errors that can be assumed have a constant bias, constant variance, autorelated in time and Gaussian distribution, the likelihood function takes the form:

\[ L(e|M(\theta, I, x, t)) = (2\pi\sigma^2)^{-T/2}(1 - \alpha^2)^{\mu/2} \]

\[ \times \exp \left[-\frac{1}{2\sigma^2} \left( \frac{1}{\alpha^2} - \frac{1}{\mu} - \frac{1}{\sigma^2} \right) \left( \frac{1}{\sigma^2} \right) \sum_{t=2}^{T} e_t^2 - \frac{1}{\alpha^2} \sum_{t=2}^{T} \left( e_t - \mu - \alpha(e_{t-1} - \mu) \right)^2 \right] \]  
\[ (6) \]

where \( \mu \) is the mean residual error (bias) and \( \alpha \) is the lag 1 correlation coefficient of the total model residuals in time.

A significant advantage of this formal statistical approach is that when the assumptions are satisfied, the theory allows the estimation of the \textit{probability} with which an observation will be predicted, conditional on the model and parameter values, and the probability density functions of the parameter estimates (which under these assumptions will be multivariate normal). As more data are made available, the use of these likelihood functions will also lead to reduced uncertainty in the estimated parameter values (even if the total error variance is not reduced). O’Hagan (2004) has suggested that this is the \textit{only} satisfactory way of addressing the issue of model uncertainty; without proper probability estimates statements about modelling uncertainty will have no meaning.

There is the issue, however, about when probability estimates based on additive (transformed) error structures are meaningful. From a purely empirical point of view, a test of the actual model residuals \( e(x,t) \) for validity relative to the assumptions made in
formulating the likelihood function might be considered sufficient to justify probability statements of uncertainty. From a theoretical point of view, however, there has to be some concern about treating the full sources of error in (3) in this type of aggregated form. Model structural errors will, in the general case, be nonlinear, nonstationary and non-additive. Input and boundary condition errors, and any parameter errors, will also be processed through the model structure in nonlinear and nonstationary and non-additive ways.

Kennedy and O’Hagen (2001) have tried to address this problem by showing that all sources of error might be represented within a hierarchical Bayesian framework. In particular, where the term \( e_M(\theta, e_0, I, e_I, x, t) \) is simple, it might be possible to estimate this as a “model inadequacy function”, or more recently “model discrepancy function” (O’Hagan, 2004). In principle, this could take any (nonlinear) form (although the most complex in the cases they considered was a constant bias, which can, in any case, be included as a parameter in (6)). The aim is to extract as much structural information from the total error series as possible, hopefully leaving a Gaussian id residual error term, \( e_r \). The model discrepancy function can then also be used in prediction, under the assumption that the nature of the structural errors in calibration will be “similar” in prediction.

It should be noted, however, that the model discrepancy function is not a direct representation of model structural error. It is a compensatory term for all the unknown sources of error in (1). These sources of error could, in principle, be considered explicitly in the Bayesian hierarchy if good information were available as to their nature. This will be rarely the case in hydrological modelling applications, where, for example, rainfall inputs to the system may be poorly known for all events, even in small experimental catchments and even the most fundamental equation – the water balance – cannot be closed by measurement (Beven, 2001, 2002b). Thus, disaggregation of the different error components will be necessarily poorly posed and ignoring potential sources of error, including model structural error, may result in an overestimation of the information content of additional data, leading to an unjustified overconfidence in estimated parameter values (Thiemann et al., 2001; Beven and Young, 2003). In representing the modelling process by a simplified form (4), the error model is required to compensate for all sources of deficiency.

Model structural error and error compensation

The question of using an error model to compensate for deficiencies in the modelling process is a particularly interesting one. Given that our models are wrong and known to be wrong, then it would seem to be a sensible strategy. If models (or parameter sets) are being compared, then the degree of compensation and the residual error variances can be compared and only those that show satisfactory levels of performance in terms of prediction uncertainties can be retained.

There are, however, many examples in the literature of the misapplication of likelihood principles to hydrological modelling problems arising from making standard assumptions (as represented by equations (5) or (6) above) to situations where the modelling errors are more complex (Beven and Young, 2003).

An interesting example in this context is that provided by Kavetski et al. (2003) in their attempt to provide a “Bayesian Total Error Analysis.” They have taken a hierarchical Bayesian approach to the rainfall-runoff modelling problem, including the effects of rainfall input errors which they have represented by a statistical distribution of multipliers on each rainfall event. The approach is implemented recursively, such that the best multiplier for the first rainfall event (and its uncertainty) is identified from the error series associated with that event; then the multiplier from the second event (and its uncertainty)
is identified; and so on. Prior distributions of the multipliers are assumed to constrain the potential range of multipliers considered.

This example highlights nicely the difference between a statistical consideration of error in the modelling process and the way that error might be processed by the mechanistic nature of the system. A hydrological model that underpredicts using the measured rainfalls for an event (for whatever reason, possibly only because of error in assumed initial conditions) can be corrected by a multiplier for that event of greater than one. This will result in a smaller residual variance, even though the measured rainfalls might have been a good estimate of the inputs to the catchment for that event. In a sequence of events, underprediction in one event means that antecedent wetness prior to the next event might be overestimated resulting in an overprediction of runoff for a following event. This could be compensated by a multiplier for that event of less than one. However, this will then have an effect on the following event, and so on.

Thus, the possibilities for compensation of different types of error allowed in the statistical approach are not resolved by taking different sources of error into account more explicitly. We can also note that, while it may be possible to reduce residual variances by allowing more degrees of freedom (such as rainfall multipliers), in calibration it may be difficult to use the information gained in this way in prediction unless some structure to the compensating factors can be discerned.

Finally, it should be noted that error compensation within this statistical framework might have the effect of saving a model from rejection. When a model and associated error model are calibrated as if the model were correct, then the error variance will adjust (if the error structure assumptions are valid) to cover the observations in the calibration period, regardless of whether the model structure is adequate or not. It is clear in some past published studies in rainfall-runoff modelling that some inadequate models have been accepted in this way. Inadequacies, due to nonlinearities or other errors not revealed in calibration, may then only be seen in prediction, strongly suggesting that the minimum test of a split-record evaluation of model predictions should always be made.

**An alternative approach: allowing model rejection in a context of model equifinality**

In a series of papers (Beven, 1993, 2000, 2002a,b, 2005), I have argued that the potential for model equifinality should be considered in applications of environmental models. The equifinality thesis suggests that, given the expected limitations of the available models and data, there may be many model structures and parameter sets within model structures that may provide acceptable simulations and that, without additional information with which to differentiate between these behavioural models, it is sensible to include them all (or a set that spans the range of functional behaviours that results in acceptable predictions) in the prediction process. The equifinality thesis is the central concept of the Generalised Likelihood Uncertainty Estimation (GLUE) methodology (Beven and Binley, 1992; Beven and Freer, 2001), which can include formal Bayes methods as special cases when the assumptions of likelihood functions such as (5) and (6) can be assumed valid, but which also allows more subjective model evaluation measures.

There is, however, a fundamental philosophical difference with parameter inference based on statistical likelihood (Beven, 2002b, 2005). The GLUE methodology does not purport to estimate the probability of predicting an observation given the model, but rather attempts to evaluate the predicted distribution of a variable that is always conditional on the models considered, the ranges of parameter values considered, the evaluation measures used and the input and output data available to the application for model evaluation. The prediction distributions do not consider the error associated with a particular model.
run explicitly. There is instead an assumption that the error series associated with a model run in calibration will have “similar” characteristics in prediction (note the similar assumption about model structural error in the formal likelihood approach above). Thus, in weighing the predictions of multiple models to form the predictive distribution for a variable, there is an implicit weighting of the error series associated with those models, without the need to consider different sources of error explicitly (explicit error models can be handled in this framework by treating them as additional model components) (Romano-wicz et al., 1996).

One of the most interesting features of the GLUE methodology is the complementarity of model equifinality and model rejection. Equifinality accepts that there may be multiple models that may be useful in prediction and that any attempt to identify an optimal model might be illusory. However, if multiple models are to be accepted as acceptable or behavioural, it is evident that models can also be rejected (given a likelihood of zero) where they can be shown to be non-behavioural (Freer et al., 2002). Thus, there is always a possibility that all the models tried will be rejected (unlike the statistical approach where it is possible to compensate for model deficiencies by some error structure).

However, at this point the limitations of implicit handling of error series in the GLUE methodology become apparent since it is possible that some hypothetical “perfect” model could be rejected if driven by poor input and boundary condition data or if compared with poor observation data. Thus, there is a need for a more explicit consideration of sources of error in this framework, as expressed in (3), while retaining the possibility of model rejection.

A potential methodology has been proposed by Beven (2005). Combining (1) and (3) we have:

\[ O(x, t) + e_O(x, t) + e_C(\Delta x, \Delta t, x, t) = M(\theta, e_0, I, e_I, x, t) + e_M(\theta, e_0, I, e_I, x, t) + e_r \]  

(7)

The error terms on the left hand side of (7) represent the measurement error and the commensurability error between observed and predicted variables. The model term, \( M(\theta, e_0, I, e_I, x, t) \), will reflect error in input and boundary conditions, model parameters, and model structure. The error term, \( e_M(\theta, e_0, I, e_I, x, t) \), can now be interpreted as a compensatory error term for model deficiencies, but which must also reflect error in input and boundary conditions, model parameters, and model structure. Finally there may be a random error term, \( e_r \).

Equation (7) has been written in this form both to highlight the importance of measurement errors and the commensurability error issue and to reflect the real difficulty of separating input and boundary condition errors, parameter errors, and model structural error in nonlinear cases. There is no general theory available for doing this in nonlinear dynamic cases. The one simplification that can be made in (7) is that, if it is applied on a model by model basis, model parameter error has no real meaning. It is the model structure and set of effective parameter values together that process the (non-error free) input data and determine total model error in space and time. Thus (7) could be rewritten, for any model structure, as:

\[ O(x, t) + e_O(x, t) + e_C(\Delta x, \Delta t, x, t) = M(\theta, I, e_I, x, t) + e_M(\theta, I, e_I, x, t) + e_r \]  

(8)

and \( e_M(\theta, I, e_I, x, t) \) is a model specific error term.

The question that then arises within this framework is whether \( e_M(\theta, I, e_I, x, t) \) is acceptable in relation to the terms \( e_O(x, t) + e_C(\Delta x, \Delta t, x, t) \). This is equivalent to asking if the following inequality holds:

\[ O_{\min}(x, t) < M(\theta, I, e_I, x, t) < O_{\max}(x, t) \quad \text{for all} \quad O(x, t) \]  

(9)
where $O_{\text{min}}(x,t)$ and $O_{\text{max}}(x,t)$ are acceptable limits for the prediction of the output variables given $\varepsilon_0(x,t)$ and $\varepsilon_C(\Delta x, \Delta t, x, t)$ which together might be termed an “effective observation error”. The effective observation error takes account both of real measurement errors and of commensurability errors between observed and predicted variables. When defined in this way, the effective observation error need not have zero mean or constant variance, nor need it be Gaussian in nature, particularly where there may be physical constraints on the nature of that error. Note that the commensurability error may be model implementation dependent in that the difference between observed and predicted variables may depend on model time and space discretisations and measurement scales in relation to expected time and space heterogeneities of the observable quantities. However, it should really be possible to develop a methodology for making prior estimates of both measurement and commensurability errors since they should be independent of individual model runs. An objective evaluation of each model run using (8) should then be possible. If a model does not provide predictions within the specified range, for any $O(x,t)$, then it should be rejected as non-behavioural.

Within the range, for all $O(x,t)$, a positive weight could be assigned to the model predictions, $M(\theta, I, \varepsilon_1, x, t)$, according to its level of apparent performance. The simplest possible weighting scheme that need not be symmetric around the observed value, given an observation $O(x,t)$ and the acceptable range $[O_{\text{min}}(x,t), O_{\text{max}}(x,t)]$, is the triangular relative weighting scheme, but other bounded weighting schemes could be used (including truncated Gaussian forms). A core range of observational ambiguity (equal weighting) could be added if required, in both timing and magnitude if necessary (Beven, 2005; Pappenberger and Beven, 2005).

This methodology gives rise to some interesting possibilities. Within this framework there is no possibility of a representation of model error being allowed to compensate for poor model performance, even for the “optimal” model. If there is no model that proves to be behavioural then it is an indication that there are conceptual, structural or data errors (though it may still be difficult to decide which is the most important). There is, perhaps, more possibility of learning from the modelling process on occasions when it proves necessary to reject all the models tried.

However, this type of evaluation requires that consideration also has to be given to input and boundary condition errors, since, as noted before, even the “perfect” model might not provide behavioural predictions if it is driven with poor input data error. Thus, it should be the combination of input/boundary data realisation (within reasonable bounds) and model structure and parameter set in producing $M(\theta, I, \varepsilon_1, x, t)$ that should be evaluated against the effective observational error. The result will (hopefully) still be a set of behavioural models, each associated with some likelihood weight. Any compensation effect between an input realisation (and initial and boundary conditions) and model parameter set in achieving success in the calibration period will then be implicitly included in the set of behavioural models.

There is also the possibility that the behavioural models defined in this way do not provide predictions that span the range of the acceptable error around an observation. The behavioural models might, for example, provide simulations of an observed variable $O(x,t)$ that all lie in the range $O(x,t)$ to $O_{\text{max}}(x,t)$, or even just a small part of it. They are all still acceptable, but are apparently biased. This provides real information about the performance of the model (and/or other sources of error) that can be investigated and allowed for specifically at that site in prediction.
Conclusions

There is uncertainty about uncertainty estimation in environmental modelling (Beven, 2004). There remain many difficult issues to resolve, in particular the role of model structural error in limiting the value of information in constraining potential model representations of a system. This will undoubtedly be difficult to resolve because of the essential paradox that results from the fact that attempts to increase the process understanding in environmental models inevitably increase the number of parameters that must be defined in any application (Beven, 2002a). This would not be a problem if such parameters could be defined a priori as “universal constants.” However, environmental models must be applied to a particular place with all its unique characteristics (Beven, 2000, 2002a), necessitating a process of calibration or conditioning that will inevitably involve multiple sources of error as set out here.

The formalism outlined above to recognise and address these sources of error seems to provide a very natural approach to model calibration and evaluation, that avoids making difficult assumptions about the nature of the modelling errors other than the effective observational error. It also focuses attention on the difference between a model predicted variable (as subject to input and boundary condition uncertainty) and what can actually be observed in the assessment of the effective observational error where this is appropriate; potential compensation between input and structural error; and the possibility of real model failure.

It does not, however, allow the separation of model structural error since it is not clear that this can be done unless very strong assumptions are made about the nature of the input and boundary condition errors. Where these can be assumed to be negligible, then the term $\varepsilon_M(\theta, I(\varepsilon_I = 0), x, t)$ will indeed represent the structural error associated with any particular model and the best model found (by whatever performance measure) could be said to provide an estimate of minimum model structural error in time and space.

This will not, however, be the general case. There will be input and boundary condition error and it may indeed be necessary to take account of these errors for a “good” model structure to satisfy the rejection criterion (9). Correspondingly, however, it will be possible for a particular input error series to compensate for a “poor” model structure in producing simulations that satisfy (9) (as also noted above in the Bayesian methodology of Kavetski et al., 2003). The acceptance of incorrect models should be expected to be reduced, however, as more data are collected. Experience to date with hydrological models suggests that this is not the main problem encountered in model rejection based strategies. The main problem is that even using relatively relaxed rejection criteria, all models are rejected.

It is always, of course, possible to avoid rejection of all the models tried by extending the range of acceptable error (or adding a compensating statistical error model). However, the important point is that there would need to be an explicit recognition and argument for doing so. An approach based on rejection rather than optimisation also tends to focus attention on particular parts of the record that are not well simulated or particular “outliers” in the error series. In this way we might learn more about the model performance (or data) and, hopefully, more about hypotheses about processes. We will not learn much from models that are declared to be successful because their limitations are compensated by a statistical error component. We will learn more from the analysis of justifiable model rejections when no models appear to be behavioural.
References