Model evaluation and performance

B. Beck

Volume 3, pp 1275–1279

in

Encyclopedia of Environmetrics
(ISBN 0471 899976)

Edited by

Abdel H. El-Shaarawi and Walter W. Piegorsch

Model evaluation and performance

Whether the behavior of a model matches the behavior of the (real) system−sufficiently well−has always been a matter of great interest, marked by many papers over many years, but especially distinctively and originally by Caswell [7]. The contemporary phrase for what one seeks to achieve in resolving this issue is 'evaluation' [14]. While it might seem strange for a label to be of significance, earlier terms used for describing this process of evaluating model performance have provoked rather vigorous debate, within which the word 'validation' was first to be replaced by 'history matching' [12], to which was then added the phrase 'quality assurance' [2, 3], but from which debate 'evaluation' may eventually emerge as the most appropriate descriptor. The difficulty of finding a label for the process is as follows. Validation and assurance prejudice the expectations of the outcome of the procedure towards only the positive−the model *is* valid or its quality *is* assured−whereas evaluation is neutral in what might be expected of the outcome. Because models of environmental systems have become so widespread in serving purposes affecting a substantially more aware and engaged audience of (scientifically) lay stakeholders, words used within the scientific enterprise can have meanings that are misleading in contexts outside the confines of the laboratory world. The public knows well that supposedly authoritative scientists can have diametrically opposed views on the benefits of proposed measures to protect the environment. Proclaiming that a model underpinning development and assessment of the consequences of these measures is *valid* harks back to an arguably outdated view of science having a unique access to the singular truth of the matter. It does not sit comfortably in the contemporary scene of what some have called post-normal science [11]; nor is it even what the builders of mathematical models of environmental systems themselves believe to be the case [6].

Essence of the Procedure

In lay terms, these are the essential questions one would like to have answered in seeking to evaluate a model:

1. Has the model been constructed of approved materials, i.e. approved constituent hypotheses (in scientific terms)?
2. Does its behavior approximate well that observed in respect of the real thing?
3. Does it work, i.e. does it fulfill its designated task, or serve its intended purpose?

Peer Review

Conventionally, the first of these questions has been answered through the process of peer review, for which extensive guidelines are available, for example, from the US **Environmental Protection Agency** [19]. Many definitions of model evaluation, such as the several entries containing the word 'validation' in the *Concise Encyclopedia of Environmental Systems* [20], treat only the quantitative, statistical aspects of the procedure (and its philosophical basis). Some things, however, are extremely difficult to deal with in quantitative terms, most notably the quality of the constituent hypotheses and the 'pedigree' of the process mechanisms incorporated into the structure of the model [10]. Whether these many components of the model have been endorsed by overwhelming consensus in their choice, or are strongly disputed, or considered highly speculative, are factors material to model evaluation, and ones ideally suited to the process of peer review. Indeed, they are especially important the more difficult it becomes to evaluate the model's overall performance against field data, and field data (in turn) will be increasingly hard to acquire for the ever more ambitious models proposed for addressing the ever more comprehensive analysis of environmental systems, including the adaptive response of communities and societies to environmental change (as in integrated assessment; see, for example, [17]).

While peer review of a model will encompass more than merely an assessment of the quality of the materials of the model's construction−for example, the adequacy of the technical qualifications of any persons wishing to use the developed model−it is important to indicate this feature as essentially an *internal* index of evaluation, cast in terms of the model's parametric space. Judgment about the model is being made by reference to the model's intrinsic mechanisms, identified in effect by the model's parameters, which determine how the input (causative) stimuli are transcribed into output responses. In principle, any such judgment ought to reflect the generic properties of the model, irrespective of the current task to which it has been assigned. In practice, however, it must inevitably reflect the accumulating experience with the model, and its earlier successful/unsuccessful performance, up to–but not including–the present task (whatever this may be).

Matching History

The second of the above questions is where the vast majority of attention has been focused. It is simply so self-evidently vital to our accepting the model as trustworthy in some sense, as epitomized in the 'matching of history', wherein the curves of the model's output responses can be seen–and quantified in numerical terms–to pass amongst the dots of the observed responses of the environmental system under study.

In summary, what we seek as the outcome of a matching of history–in fact, ideally the outcome of the attempt to match at least a second, independent history (the first having been used for calibration)–are the following: (a) a sequence, or set, of residual errors of mismatch of small magnitude, with a mean value approaching zero, and statistical properties closely approximating those of a white-noise sequence; (b) errors (uncertainty) attaching to the estimates of the model's parameters, when derived from calibration, whose variances are of small magnitude; and (c) no significant correlations among these errors, i.e. small error covariances between the various model parameters. We need two independent histories because, as Bohlin [5] has so aptly put it: 'one good fit makes a data description; two good fits makes a system description'. In this is embraced the following agreed procedure, around which a stable consensus has been gathered over decades. A first set of past data is used to calibrate the model, i.e. to derive values to be used for the complete set of the model's parameters, where these values are associated with a good, in principle, 'best', match of the model with the data; the calibrated model, with no further adjustments whatsoever, is then tested against the second set of data, which it should match acceptably; and 'acceptability' as such has generally been judged solely in terms of outcome (a) above (*see* **Cross-validation**). When restricted in this way, it is possible to see how the mere matching of history can be described as essentially an *external* index of evaluation, cast in the output space of the model. Assessment strictly of the residual errors of (output) mismatch calls for some comparison of data derived from the model with data deduced from sources of knowledge or experience utterly independent of the specific model under scrutiny. Typically, appropriate data of this sort will be those derived from empirical observation, but they might be obtained from alternative candidate models. Since most complex models of environmental systems will share similar pedigrees, however, this necessarily undermines the struggle to achieve maximal independence in the two sets of data juxtaposed at the heart of the test.

The anatomy of the test itself has four parts:

1. *the raw data*, for example, the sequences of model outputs, the observed (system) output responses, and the differences between these two sets of data;
2. *summarizing properties of these data*, such as (statistical) **distribution functions**, the moments of these distributions (e.g. their means), and the sets of coefficients appearing in correlation functions and regression relationships (relating estimated to observed output data);
3. *the decision*, quintessentially of whether to accept or reject the model as having matched history, into the making of which may be funneled both the raw data and their summarizing properties; and
4. *the decision statistics*, used to guide consistent application of the rules for accepting or rejecting the model, for example, the χ^2, Student's t, Kolmogorov–Smirnov, and Mann–Whitney–Wilcoxon test statistics, among others.

One is not obliged to employ these formal methods of statistical hypothesis testing [13, 15]. The person charged with evaluating the model may simply make up his/her own mind on the basis of the data presented, without reaching for a formalized decision rule. However, it is undoubtedly the use of these methods that has dominated conventional understanding of what constitutes testing of whether history has been matched or not. In short, when most of us think of model evaluation, a statistical test of the deviations between observed and computed output responses comes immediately to mind.

Such evaluation according to the residual errors of mismatch alone–according to outcome (a) above, that is–says nothing directly of the quality of the materials of which the model has been constructed. Indeed, such tests imply one can accept or reject the model as a whole, without the subtlety of being able to accept some and reject others of its component parts. Outcomes (b) and (c), having to do with the uncertainty attaching to the set of model parameters, provide evidence of this, with the quality of the constituent materials being roughly comparable with the inverse of estimated parametric uncertainty. In the ideal, over successive evaluations against independent histories, one would want to see successive reductions in the uncertainty attaching to the model's parameters. Thus, changes in the external index of evaluation (matching of a variety of histories) would be accompanied by accumulating evidence regarding the internal index, which concerns itself with statements about the quality of the model's construction (approval of its constituent hypotheses). Awareness of these secondary kinds of outcomes, though once not widespread, can be crucial.

Almost all models suffer from a lack of identifiability; put simply, many combinations of values for the model's parameters may permit the model to fit the observed data more or less equally well [1, 4]. In other words, history may have been matched, but at the cost of substantial ambiguity in the parameterization of the model, as reflected in large error covariances among the estimates of the parameters [signalled through outcome (c)]. If just a singular, supposedly 'uniquely best', set of values for the model's parameters has been obtained from calibration, then these can be retained for the test of the model against the second set of data, and the latent ambiguity may never come to light. Yet it is almost certainly there, and would manifest itself in the form of the following conundrum: if several (perhaps many) candidate parameterizations yield a good calibration against the first history, which of these should be called upon for the purposes of matching the second history? If the presumption throughout evaluation of the model is that only a single candidate parameterization of the model should exist, this conundrum will not arise, and that has been precisely the problem – for there should *always* be enquiry into the presence of ambiguity. Unfortunately, illuminating such ambiguity has been impeded by the absence of good systematic procedures for computing error variance–covariance properties for the parameters of models ranging beyond the simpler regression-type relationships [1].

To summarize, the behavior of the model may approximate well that observed in respect of the real thing, but there may be countless variations on the theme of the way in which the model might be constructed, some of which may not be 'approved' (when subjected to peer review).

Fulfilling a Designated Task

A model may be constructed for a variety of purposes, for instance, to provide:

1. a succinctly encoded archive of contemporary knowledge;
2. an instrument of prediction (in support of making a decision or formulating a policy);
3. a device for communicating scientific notions to a scientifically lay audience;
4. an exploratory vehicle for discovery of our ignorance.

No one would wish to tolerate extensive ambiguity in a model intended to serve the first of these purposes. Striving to overcome the seemingly unpalatable lack of model identifiability could in this sense be a worthy, if barely winnable, struggle. Yet a lack of identifiability does not preclude a positive outcome of the evaluation, especially from the pragmatic perspective of the model fulfilling the second of the above purposes. The supporting argument runs as follows (it implies the use of ancillary analyses of sensitivity and uncertainty [1, 16]). We presume a priori no uniquely best parameterization of the model. The model is reconciled with history, yielding possibly a multiplicity of 'acceptably good' such parameterizations, hence the intrinsic uncertainty. This multiplicity of parameterizations is accounted for in some way in generating a variety of predicted consequences for each of the policy options. Providing a clear preference for adopting one of the options can then be identified, and this preference survives, say, all manner of analyses of sensitivity of the predictions, their uncertainties, and the rankings of the options to the evident uncertainty and ambiguity in the calibrated model, one can move forward with the preferred option, relatively secure in the knowledge that this decision is robust against uncertainty in the model [8]. The model will have served its purpose and will thereby have been evaluated positively, in that sense.

Models, as Caswell observed long ago [7], are objects *designed* to fulfill clearly expressed tasks, just as hammers, screwdrivers, and other tools have been designed to serve identified or stated purposes. Some of the tasks listed above are more purely what would be called scientific (1 and 4), some obviously more pragmatic (2 and 3). Conceiving therefore of a model as a tool, and imagining how we might evaluate the appropriateness (or otherwise) of that tool, specifically for serving pragmatic purposes, has begun to broaden the concept of evaluation itself, beyond the customary domains of peer review and matching of history [2, 3]. Our perspective can be placed outside the traditional view of models as computerized articulations of theory whose purpose, at bottom, is to make predictions of a future state of nature, ultimately falsifiable by subsequent observation when the time comes. How would we, from the changed perspective, evaluate whether a model was well- or ill-designed as a communications device, for example? For the time being we have only started to ask such questions.

Open Questions

Like the subject of pollutant dispersion, wherein unresolved intricacies are continually being revealed, discussion of the subject of evaluating models (formerly referred to as validation) seems destined to continue (e.g. [18]). For one thing, the task is becoming no easier, since it must follow our apparently boundless ambition for the development and application of models of the behavior of environmental systems. Ever larger models will be constructed. They will be ever more dependent upon multidisciplinary knowledge bases, extremely difficult to scrutinize, and doubtless strongly immune to empirical refutation. We can already sense a shift away from the rigour of evaluating very high-order models (VHOMs), for instance in **oceanography**, where application of the algorithms of data assimilation is rising to the fore [9]. In other words, the outlook is coalescing around the view that the relatively sparse data can but be assimilated into the current theory, not employed to root out ruthlessly its inadequacies. We do indeed have a problem with evaluation in the sense of matching history, and the situation may be little better in respect of peer review, simply because there will be few peers for such VHOMs having no conflict of interest. It will not suffice to be cast into a state of mental paralysis on the issue of their evaluation, but neither will it be sufficient to argue that their quality has been assured simply by virtue of every conceivable constituent hypothesis having been incorporated a priori. There is scope for much primary thought to be invested in the topic of evaluating VHOMs in particular.

Ten years ago, had one been asked, validation−as it was referred to then−would have been defined as the assessment of a model's predictive performance against a second set of (independent) data given parameter values identified from a first set of data. Today we are armed with a wider palette of metaphors and analogs (the legal process, quality assurance in the design of tools, and quality assurance in controlling procedures in an analytical laboratory) with which to fashion a broader protocol for the conduct of evaluating a model.

References

[1] Beck, M.B. (1987). Water quality modeling: a review of the analysis of uncertainty, *Water Resources Research* **23**, 1393−1442.

[2] Beck, M.B. & Chen, J. (2000). Assuring the quality of models designed for predictive purposes, in *Sensitivity Analysis*, A. Saltelli, K. Chan, & E.M. Scott, eds, Wiley, Chichester, pp. 401−420.

[3] Beck, M.B., Ravetz, J.R., Mulkey, L.A. & Barnwell, T.O. (1997). On the problem of model validation for predictive exposure assessments, *Stochastic Hydrology and Hydraulics* **11**, 229−254.

[4] Beven, K.J. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling, *Advances in Water Resources* **16**, 41−51.

[5] Bohlin, T. (1993). Validation of identified models, in *Concise Encyclopedia of Environmental Systems*, P.C. Young, ed., Pergamon Press, Oxford, pp. 645−650.

[6] Bredehoeft, J.D. & Konikow, L.F. (1993). Groundwater models: validate or invalidate, *Ground Water* **31**, 178−179.

[7] Caswell, H. (1976). The validation problem, in *Systems Analysis and Simulation in Ecology*, Vol. IV, B.C. Patten, ed., Academic Press, New York, pp. 313−325.

[8] Duchesne, S., Beck, M.B. & Reda, A.L.L. (2001). Ranking stormwater control strategies under uncertainty: the River Cam case study, *Water Science and Technology* **43**, 311−320.

[9] Evensen, G. (1994). Inverse methods and data assimilation in nonlinear ocean models, *Physica D* **77**, 108−129.

[10] Funtowicz, S.O. & Ravetz, J. (1990). *Uncertainty and Quality in Science for Policy*, Kluwer, Dordrecht.

[11] Funtowicz, S.O. & Ravetz, J. (1993). Science for the post-normal age, *Futures* **25**, 735−755.

[12] Konikow, L.F. & Bredehoeft, J.D. (1992). Groundwater models cannot be validated, *Advances in Water Resources* **15**, 75–83.

[13] Luis, S.J. & McLaughlin, D.B. (1992). A stochastic approach to model validation, *Advances in Water Resources* **15**, 15–32.

[14] Oreskes, N. (1998). Evaluation (not validation) of quantitative models for assessing the effects of environmental lead exposure, *Environmental Health Perspectives* **106**, 1453–1460.

[15] Reckhow, K.H., Clements, J.T. & Dodd, R.C. (1990). Statistical evaluation of mechanistic water-quality models, *Journal of Environmental Engineering* **116**, 250–268.

[16] Saltelli, A., Chan, K. & Scott, E.M., eds (2000). *Sensitivity Analysis*, Wiley, Chichester.

[17] Schneider, S. (1997). Integrated assessment modeling of climate change: transparent rational tool for policy-making or opaque screen hiding value-laden assumptions?, *Environmental Modeling and Assessment* **2**, 229–250.

[18] Shelly, A., Ford, E.D. & Beck, M.B. (2000). *Quality Assurance of Environmental Models*, NRCSE-TRS 042, National Center for Research on Statistics and the Environment, University of Washington, Seattle.

[19] US EPA (1998). *Science Policy Council Handbook: Peer Review*, Report EPA 100-B-98-001, Office of Science and Policy, Office of Research and Development, US Environmental Protection Agency, Washington.

[20] Young, P.C., ed. (1993). *Concise Encyclopedia of Environmental Systems*, Pergamon Press, Oxford.

(*See also* **Modeling, environmental**)

B. Beck