Modelling guidelines—terminology and guiding principles

Jens Christian Refsgaard *, Hans Jørgen Henriksen

Department of Hydrology, Geological Survey of Denmark and Greenland (GEUS), Øster Voldgade 10, Copenhagen DK-1350, Denmark

Received 29 October 2002; received in revised form 7 August 2003; accepted 18 August 2003

Abstract

Some scientists argue, with reference to Popper’s scientific philosophical school, that models cannot be verified or validated. Other scientists and many practitioners nevertheless use these terms, but with very different meanings. As a result of an increasing number of examples of model malpractice and mistrust to the credibility of models, several modelling guidelines are being elaborated in recent years with the aim of improving the quality of modelling studies. This gap between the views and the lack of consensus experienced in the scientific community and the strongly perceived need for commonly agreed modelling guidelines is constraining the optimal use and benefits of models. This paper proposes a framework for quality assurance guidelines, including a consistent terminology and a foundation for a methodology bridging the gap between scientific philosophy and pragmatic modelling. A distinction is made between the conceptual model, the model code and the site-specific model. A conceptual model is subject to confirmation or falsification like scientific theories. A model code may be verified within given ranges of applicability and ranges of accuracy, but it can never be universally verified. Similarly, a model may be validated, but only with reference to site-specific applications and to pre-specified performance (accuracy) criteria. Thus, a model’s validity will always be limited in terms of space, time, boundary conditions and types of application. This implies a continuous interaction between manager and modeller in order to establish suitable accuracy criteria and predictions associated with uncertainty analysis.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Model guidelines; Scientific philosophy; Validation; Verification; Confirmation; Domain of applicability; Uncertainty

1. Introduction

Models describing water flows, water quality and ecology are being developed and applied in increasing number and variety. With the requirements imposed by the EU Water Framework Directive the trend in recent years to base water management decisions to a larger extent on model studies and to use more sophisticated models is likely to be reinforced. At the same time insufficient attention is generally given to documenting the predictive capability of the models. Therefore, contradictions emerge regarding the various claims of model applicability on the one hand and the lack of documentation of these claims on the other hand. Hence, the credibility of the models is often questioned, and sometimes with good reason.

As emphasised by e.g. Forkel [12] modelling studies involve several partners with different responsibilities.

The ‘key players’ are code developers, model users and water resources managers. However, due to the complexity of the modelling process and the different backgrounds of these groups, gaps in terms of lack of mutual understanding easily develop. For example, the strengths and limitations of modelling applications are most often difficult, if not impossible, to assess by the water resources managers. Similarly, the transformation of water managers’ objectives to specific performance criteria can be very difficult to assess for the model users. Due to lack of documentation and transparency, modelling projects can be difficult to audit, and without a considerable effort it is hardly possible to reconstruct, repeat and reproduce the modelling process and its results.

In the water resources management community a number of different guidelines on good modelling practice have been prepared. One of the most, if not the most, comprehensive examples of modelling guidelines has been developed in The Netherlands [37] as a result of a process involving all the main players in the Dutch water management field. The background for this process was a perceived need for improving the quality in modelling.
by addressing malpractice such as careless handling of input data, insufficient calibration and validation and model use outside its scope [34]. Similarly, the background for modelling guidelines for the Murray–Darling Basin in Australia was a perception among the end-users that model capabilities may have been ‘over-sold’, and that there is a lack of consistency in approaches, communication and understanding among and between modellers and water resources managers, often resulting in considerable uncertainty for decision making [25].

A key problem in relation to establishment of generally acceptable modelling guidelines is confusion on terminology. For example the terms validation and verifications are used with different, and some times interchangeable, meaning by different authors. The confusion arises from both semantic and philosophical considerations [32]. Another important problem is the lack of consensus related to the so far non-conclusive debate on the fundamental question concerning whether a water resources model can be validated or verified, and whether it as such can be claimed to be suitable or valid for particular applications [3,11,16,20,26].

Finally, modelling guidelines have to reflect and be in line with the underlying philosophy of environmental modelling which have changed significantly during the past decades from what in retrospect may be called rather naive enthusiasms (see for example Freeze and Harlan [13]; Abbott [1]—many of us focussed on the huge potentials of sophisticated models outlined in these early days without reflecting too much on the associated limitations) to what now appears to be a much more balanced and mature view (e.g. Beven [7,9]).

Thus, there is a gap between the theory and practice, i.e. between the various, contradictory views and the lack of a common terminology and methodology in the scientific community on the one side, and the need of having quality assurance guidelines for practical model applications on the other side. The objective of the present paper is to establish guiding principles for quality assurance guidelines, including establishing a consistent terminology and a foundation for a methodology bridging the gap between scientific philosophy and pragmatic modelling.

2. Key opinions in the scientific community

The present paper does not attempt to provide a full review of all relevant papers on this subject. Rather, it provides a review of a few selected characteristic examples.

2.1. Terminology

No unique and generally accepted terminology and methodology exist at present in the scientific community with respect to modelling protocol and guidelines for good modelling practise. Examples of general methodologies exist [4,32,33], but they use different terminology and have significant differences with respect to the underlying scientific philosophy.

A rigorous and comprehensive terminology for model credibility was presented by Schlesinger et al. [33]. This terminology was developed by a committee composed of members from diverse disciplines and background with the intent that it could be employed in all types of simulation applications. In regard to terminology, distinctions are made between model qualification (adequacy of conceptual model), model verification (adequacy of computer programme) and model validation (adequacy of site-specific model). With the exception of a few important terms, such as generic model code and model calibration, which are not considered by Schlesinger et al. [33], their proposed terminology includes all the important elements of the modelling process.

Konikow and Bredehoefdt [20], in their thought provoking paper, express the view that “the terms validation and verification have little or no place in groundwater science; these terms lead to a false impression of model capability”. Their main argument relates to the anti-positivistic view that a theory (in this case a model) can never be proved to be generally valid, but may in contrary be falsified by just one example. They argue and recommend that the term history matching, which does not indicate a claim of predictive capability, should be used instead.

Oreskes et al. [26], in their classic and philosophically based paper, distinguish between verification, validation and confirmation:

- **Verify** is “an assertion or establishment of truth”. To verify a model therefore means to demonstrate its truth. According to the authors “verification is only possible in closed systems in which all the components of the system is established independently and are known to be correct. In its application to models of natural systems, the term verification is highly misleading. It suggests a demonstration of proof that is simply not accessible”. They argue that mathematical components are subject to verification, because they are part of closed systems, but numerical models in application cannot be verified because of uncertainty of input parameters, scaling problems and uncertainty in observations.
- **The term validation is weaker than the term verification.** Thus validation does not necessarily denote an establishment of truth, but rather “the establishment of legitimacy, typically given in terms of contracts, arguments and methods”. They argue that “the term valid may be useful for assertions about a generic model code but is clearly misleading if used to refer to actual model results in any particular realisation”.

• The term *confirmation* is weaker than the terms verification and validation. It is used with regard to a theory, when it is found that the theory is in agreement with empirical observations. As discussed below such agreement does not prove that the theory is true, it only confirms it.

Oreskes et al. [26] do not define how the terms verification and validation should be used, but rather define their meaning and set limitations to the contexts in which they meaningfully can be used.

An important distinction is made between open and closed systems. A system is a closed system if its true conditions can be predicted or computed exactly. This applies to mathematics and mostly to physics and chemistry. Systems where the true behaviour cannot be computed due to uncertainties and lack of knowledge on e.g. input data and parameter values are called open systems. The systems we are dealing with in water resources management, based on geosciences, biology and socio-economy, are open systems.

It may be argued that e.g. the behaviour of a groundwater flow system can be predicted correctly if all the details of the subsurface (soil system and geological system) media were known, because the fundamental physical laws governing the flow are known. However, in practice it will never be possible to know all the details of the media down to molecular scale, and hence uncertainties will always exist. For instance, several alternative representations of the subsurface system at microscopic scale will be able to provide the same flow field at a macroscopic scale. Therefore, the results from a groundwater flow model are said to be non-unique. In addition, as the system is a so-called open system, the boundary conditions generate further uncertainty.

Matalas et al. [24] draw a distinction between the terms ‘model’ and ‘theory’. They state that “a theory represents a synthesis of understanding, which provides not only a description of what constitutes the states of the system and their connectedness (i.e. postulated concepts), but also deducted consequences from these postulates. A model is an analogy or an abstraction, which … may be derived intuitively and without formal deductive capability”.

Rykiel [32] argues that models can be validated as acceptable for pragmatic purposes, whereas theoretical validity is always provisional. In this respect he, like Matalas et al. [24], distinguishes between scientific models and predictive (engineering) models. Scientific models can be corroborated (confirmed) or refuted (falsified) in the sense of hypothesis testing, while predictive models can be validated or invalidated in the sense of engineering performance testing. Thus according to Rykiel [32], validation is not a procedure for testing scientific theory or for certifying the ‘truth’ of current scientific understanding, but rather a testing of whether a model is acceptable for its intended use.

Within the hydraulic engineering community attempts have been made to establish a common quality assurance methodology IAHR [18]. The IAHR methodology comprises guidelines for standard validation documents, where validation of a software package is considered in four steps [10,23]: conceptual validation, algorithmic validation, software validation and functional validation. It is noted that the term validation in the IAHR methodology corresponds to what other authors call code verification, while schemes for validation of site-specific models are not included.

### 2.2. Scientific philosophical aspects of verification and validation

Different principal schools of philosophical thought exist on the issue of verification and validation. During the second half of the 19th century and the first half of the 20th century positivism was the dominant philosophical school. Matalas et al. [24] characterises the positivistic school in the following way: “… theories are proposed through inductive logic, and the proposed theories are confirmed or refuted on the basis of critical experiments designed to verify the consequences of the theories. And through theory reduction or adoption of new or modified theories, science is able to approach truth”. The logic rationale behind positivism is the inductive method, i.e. the inference from singular statements, such as accounts of results of observations or experiments, to universal statements, such as hypothesis or theories.

Popper [29] opposed the positivistic school arguing that science is deductive rather than inductive, and that theories cannot be verified, only falsified. The deductive method implies inferences from a universal statement to a singular statement, where conclusions are logically derived from given premises. Science is considered as a hypothetico-deductive activity, implying that empirical observations must be framed as deductive consequences of a general theory or scientific law. If the observations can be shown to be true then the theory or law is said to be corroborated. Popper used the term corroborate instead of confirmation, because he “wanted a neutral term to describe the degree to which a theory has stood up to severe tests and proved its mettle”.

The greater the number and diversity of confirming observations the more credible the theory or law becomes. But no matter how much data and how many confirmations we have, there will always be the possibility that more than one theory can explain the observations. Over time the false theories are likely to be confronted with observations that falsify them. Thus, scientific theories are never certain or proved but only hypotheses subject to corroboration or falsification.
Popper [29] distinguished between two kinds of universal statements: the ‘strictly universal’ and the ‘numerical universal’. The strictly universal statements are those usually dealt with when speaking about theories or natural laws. They are a kind of ‘all-statement’ claiming to be true for any place and any time. In contrary numerical universal statements refers only to a finite class of specific elements within a finite individual spatio-temporal region. A numerical universal statement is thus in fact equivalent to conjunctions of singular statements.

Kuhn [21] also strongly criticised positivism, and in a discussion of selection of correct scientific theories (paradigms) states “… few philosophers of science still seek absolute criteria for the verification of scientific theories. Noting that no theory can ever be exposed to all possible relevant tests, they ask not whether a theory has been verified but rather about its probability in the light of the evidence that actually exists. And to answer that question one important school is driven to compare the ability of different theories to explain the evidence at hand.”

According to the deductive approach a given system is reduced into elements or sub-systems that are closed, i.e. without uncertainties from the boundary or initial conditions, and a given hypothesis is then confirmed by use of causal relationships and rigorous logic. The deductive method is the traditional scientific philosophy and methodology for ‘exact sciences’ such as physics and chemistry. Hansen [15] and Baker [5] argue that this deductive or ‘theory-directed’ scientific method is not suitable to earth sciences, such as geology and biology, which are characterised by open systems, and where many of the signs in the historical development process are not preserved. Instead, they argue for another scientific method, which they, respectively, denote ‘holistic’ or ‘earth-directed’. The earth-directed scientific method does not focus on idealised theories verified in experimental laboratories. Instead, it is oriented towards observations in nature, uncontrolled by artificial constraints. The earth-directed method, being more ‘soft’ and accepting conclusions on the complex state of nature from an integration of many observations, but without the logical rigorous proof required by the deductive method, can be argued to be well in line with Popper’s philosophy where the scientific knowledge comprises a variety of falsifiable theories that are subject to tests against observations [15].

2.3. Philosophy of environmental modelling

Following several papers (ranging from Beven [6] to [7]) with comprehensive critique against the predominant philosophy underlying most environmental modelling, Beven [9] outlines a new philosophy for modelling of environmental systems. The basic aim of this new approach is to extend the most common, past approach with a more realistic account of uncertainty rejecting the idea of being able to identify only one optimal model as being the most reliable for a given case. His basic idea is in line with Oreskes et al. [26] that verification and validation of environmental models is impossible, because natural systems are open. Instead environmental models may be non-unique subject to only a conditional confirmation, due to e.g. errors in model structure, calibration of parameters and period of data used for evaluation. Due to this there will always be the possibility of equifinality in that many different model structures and parameter sets may give simulations that cannot be falsified from the available observational data. Beven therefore argues that the range of behaviour models (structures and parameter sets) is best represented in terms of mapping of the ‘landscape space’ into the ‘model space’, and that uncertainty predictions should consider all the behavioural models.

3. Proposed terminology and methodological framework

The following terminology is inspired by the generalised terminology for model credibility proposed by Schlesinger et al. [33], but modified and extended to accommodate some of the scientific philosophical issues raised above. The simulation environment is divided into four basic elements as shown in Fig. 1. The inner arrows describe the processes that relate the elements to each other, and the outer circle refers to the procedures that evaluate the credibility of these processes.

In general terms a model is understood as a simplified representation of the natural system it attempts to describe. However, in the terminology proposed below a distinction is made between three different meanings of the general term model, namely the conceptual model, the model code and the model that here is defined as a site-specific model. The most important elements in the terminology and their interrelationships are defined as follows:

**Reality**: The natural system, understood here as the study area.

**Conceptual model**: A description of reality in terms of verbal descriptions, equations, governing relationships or ‘natural laws’ that purport to describe reality. This is the user’s perception of the key hydrological and ecological processes in the study area (perceptual model) and the corresponding simplifications and numerical accuracy limits that are assumed acceptable in order to achieve the purpose of the modelling. A conceptual model thus includes both a mathematical description (equations) and a descriptions of flow processes, river system elements, ecological structures, geological features, etc. that are required for the particular purpose of modelling. By drawing an analogy to the scientific
philosophical discussion above the conceptual model in other words constitutes the scientific hypothesis or theory that we assume for our particular modelling study.

**Model code**: A mathematical formulation in the form of a computer program that is so generic that it, without program changes, can be used to establish a model with the same basic type of equations (but allowing different input variables and parameter values) for different study areas.

**Model**: A site-specific model established for a particular study area, including input data and parameter values.

**Model confirmation**: Determination of adequacy of the conceptual model to provide an acceptable level of agreement for the domain of intended application. This is in other words the scientific confirmation of the theories/hypotheses included in the conceptual model.

**Code verification**: Substantiation that a model code is in some sense a true representation of a conceptual model within certain specified limits or ranges of application and corresponding ranges of accuracy.

**Model calibration**: The procedure of adjustment of parameter values of a model to reproduce the response of reality within the range of accuracy specified in the performance criteria.

**Model validation**: Substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model.

**Model set-up**: Establishment of a site-specific model using a model code. This requires, among other things, the definition of boundary and initial conditions and parameter assessment from field and laboratory data.

**Simulation**: Use of a validated model to gain insight into reality and obtain predictions that can be used by water managers. This includes insight into how reality can be expected to respond to human interventions. In this connection uncertainty assessments of the model predictions are very important.

**Performance criteria**: Level of acceptable agreement between model and reality. The performance criteria apply both for model calibration and model validation.

**Domain of applicability (of conceptual model)**: Prescribed conditions for which the conceptual model has been tested, i.e. compared with reality to the extent possible and judged suitable for use (by model confirmation).

**Domain of applicability (of model code)**: Prescribed conditions for which the model code has been tested, i.e. compared with analytical solutions, other model codes or similar to the extent possible and judged suitable for use (by code verification).

**Domain of applicability (of model)**: Prescribed conditions for which the site-specific model has been tested, i.e. compared with reality to the extent possible and judged suitable for use (by model validation).

The credibility of the descriptions or the agreements between reality, conceptual model, model code and model are evaluated through the terms confirmation, verification, calibration and validation. Thus, the relation between reality and the scientific description of reality which is constituted by the conceptual model with its theories and equations on flow and transport processes, its interpretation of the geological system and ecosystem at hand, etc., is evaluated through the confirmation of the conceptual model. As a logical consequence of our
position on scientific methodology, we use the term confirmation in connection with conceptual model. This implies that we agree that it is never possible to prove the truth of a theory/hypothesis and as such of a conceptual model. And even if a site-specific model is eventually accepted as valid for specific conditions, this is not a proof that the conceptual model is true, because, due to non-uniqueness, the site-specific model may turn out to perform right for the wrong reasons.

Methods for conceptual model confirmation should follow the standard procedures for confirmation of scientific theories. This implies that conceptual models should be confronted with actual field data and be subject to critical peer reviews. Furthermore, the feedback from the calibration and validation process may also serve as a means by which one or a number of alternative conceptual model(s) may be either confirmed or falsified.

The ability of a given model code to adequately describe the theory and equations defined in the conceptual model by use of numerical algorithms is evaluated through the verification of the model code. Use of the term verification in this respect is in accordance with Oreskes et al. [26], because mathematical equations are closed systems. The methodologies used for code verification include comparing a numerical solution with an analytical solution or with a numerical solution from other verified codes. However, some programme errors only appear under circumstances that do not routinely occur, and may not have been anticipated. Furthermore, for complex codes it is virtually impossible to verify that the code is universally accurate and error-free. Therefore, the term code verification must be qualified in terms of specified ranges of application and corresponding ranges of accuracy. A code may be applied outside its documented ranges of application, but in such cases it must not carry the label ‘verified’ and caution should be expressed with respect to its results.

The application of a model code to be used for setting up a site-specific model is usually associated with model calibration. The model performance during calibration depends on the quantity and quality of the available input and observation data as well as on the conceptual model. If sufficient accuracy cannot be achieved either the conceptual model and/or the data have to be re-evaluated. A discussion of the problems and methodologies in model calibration is provided by Gupta et al. [14].

Often the model performance during calibration is used as a measure of the predictive capability of a model. This is a fundamental error. Many studies (e.g. Refsgaard and Knudsen [31]; Lidén [22]) have demonstrated that the model performance against independent data not used for calibration is generally poorer than the performance achieved in the calibration situation. Therefore, the credibility of a site-specific model’s capability to make predictions about reality must be evaluated against independent data. This process is denoted model validation. In designing suitable model validation tests a guiding principle should be that a model should be tested to show how well it can perform the kind of task for which it is specifically intended [19]. This implies for instance that for the case where a model is intended to be used for conditions similar to conditions where test data exist, such as extension of streamflow records, a standard split-sample test may be applied. However, models are often intended to be used as management tools to help answer questions such as: What happens to the water resources if land use is changed? In such case no site-specific test data exist and the question of defining a validation test scheme becomes non-trivial.

4. Discussion

4.1. Scientific philosophical aspects

The fundamental view expressed by scientific philosophers is that verification and validation of numerical models of natural systems is impossible, because natural systems are never closed and because the mapping of model results are always non-unique [26]. Thus, seen from a theoretical point it is tempting to conclude that the establishment of modelling guidelines comprising these terms simply is not possible.

On the other hand, there is a large and increasing need to establish guidelines to improve the quality of modelling, and such guidelines need to address the issues of verification and validation in order to be operational in practise. Irrespective of what the scientific community decides regarding terminology and validation methodology, including the associated philosophical aspects, models are being used more and more to support water resources management in practise. As long as the present situation continues, characterised by a large degree of confusion on terminology and methodology, the potential benefits of using models are severely constrained. They are often subject to either ‘overselling’ or ‘mis-trust’, and misunderstandings between model users and water resources managers may easily occur in the absence of a commonly accepted and understood ‘language’. Thus, establishment of a terminology and methodology that bridge the gap between scientific philosophy and pragmatic modelling is a key challenge and an important one.

This gap between a scientific philosophical and a pragmatic modelling position is also clearly reflected in the dialogue between Konikow and Bredehoeft [20] and De Marsily et al. [11]. Following the Popperian school, Konikow and Bredehoeft [20] express the view that “the terms validation and verification have little or no place
in ground-water science; these terms lead to a false impression of model capability”. De Marsily et al. [11], in a response, argue for a more pragmatic view: “... using the model in a predictive mode and comparing it with new data is not a futile exercise; it makes a lot of sense to us. It does not prove that the model will be correct for all circumstances, it only increases our confidence in its value. We do not want certainty; we will be satisfied with engineering confidence.”

With regard to scientific methodology we fundamentally agree with the views of Popper [29] and the earth-directed theoretical method described by Baker [5]. Consequently, we agree with the view of Oreskes et al. [26], Konikow and Bredehoeft [20] and many others that it is not possible to carry out model verification or model validation, if these terms are used without restriction to domains of applicability and levels of accuracy.

The restrictions in use of the terms confirmation, verification and validation imposed by the respective domains of applicability imply, according to Popper’s views, that the conceptual model, model code and site-specific models can only be classified as numerical universal statements as opposed to strictly universal statements. This distinction is fundamental for our proposed methodology and its link to scientific philosophical theories.

4.2. Model confirmation, verification and validation

An important aspect of our proposed methodology lies in the separation between the three different ‘versions’ of the word model, namely the conceptual model, the model code and the site-specific model. This separation is in line with Matalas et al. [24] and Rykiel [32], who distinguish between the theory (conceptual model) and the engineering model (the site-specific model). Similarly, Schlesinger et al. [33] distinguish between conceptual model and computerised model. Schlesinger et al. [33], Matalas et al. [24] and Rykiel [32] do not separate the model code from the site-specific model.

Due to this distinction it is possible, at a general level, to talk about confirmation of a theory or a hypothesis about how nature can be described using the relevant scientific method for that purpose, and, at a site-specific level, to talk about validity of a given model within certain domains of applicability and associated with specified accuracy limits.

As Beven [9] argues we need to distinguish between our qualitative understanding (perceptual model) and the practical implementation of that understanding in our conceptual model. As we have defined a conceptual model as combination of a perceptual model and the simplifications acceptable for a particular model study a conceptual model becomes site-specific and even case specific. For example a conceptual model of a ground-water aquifer may be described as two-dimensional for a study focussing on regional groundwater heads, while it may need to include more complex three-dimensional geological structures for detailed simulation of solute transport studies.

Confirmation of a conceptual model is a non-trivial issue. It is hardly possible to prescribe general test procedures, in particular not exact tests. Conceptual models are more difficult in some domains than in others. For example, the process descriptions/equations and the actual system is relatively easily identifiable in a hydrodynamic river flow system as compared to a groundwater system or an ecosystem, because the geology will never be completely known in a groundwater system and the biological processes may not be well known in an ecosystem. The more complex and difficult the conceptual model becomes the more ‘soft’ the confirmation tests may turn out to be. Thus, expert knowledge in terms of peer reviews may be an important element of such tests.

In cases where considerable uncertainty exists in the conceptual model, the possibility of testing alternative conceptual models should be promoted. An example of this is given by Troldborg [35], who reports a study where three scientists developed alternative geological interpretations for the same area, and three numerical groundwater models were set-up and calibrated on this basis. During this process, or in the subsequent validation phase, one or more of these models may turn out to perform so poorly that the underlying conceptual model has to be rejected. This approach of building the uncertainty of our knowledge of reality into alternative conceptual models, which are subsequently subject to a confirmation test, is fully in line with Popper’s scientific philosophical school. Unfortunately, this is very seldom pursued in practise.

Code verification is not an activity that is carried out from scratch in every modelling study. In a particular study it has to be ascertained that the domain of applicability for which the selected model code has been verified covers the conditions specified in the actual conceptual model. If that is not the case, additional verification tests have to be conducted. Otherwise, the code explicitly must be classified as not verified for this particular study, and the subsequent simulation results therefore have to be considered with extra caution.

Establishment of validation test schemes for the situations, where the split-sample test is not sufficient, is an area, where limited work has been carried out so far. The only rigorous and comprehensive methodology reported in literature is that of Klemes [19]. He proposed a systematic scheme of validation tests, where a distinction is made between simulations conducted for the same catchment as was used for calibration (split-sample test) and simulations conducted for ungauged catchments (proxy-basin tests). He also distinguished between
cases where catchment conditions such as climate, land use and ground water abstraction are stationary (split-sample test) and cases where they are not (differential split-sample test). A further discussion, including examples, of Klemes’s test scheme is given in Refsgaard [30]. The two key principles are: (a) the validation tests must be carried out against independent data, i.e. data that have not been used during calibration, and (b) the model should be tested to show how good it can perform the kind of task for which it is specifically intended to be applied subsequently. This implies e.g. that multi-site validation is needed if predictions of spatial patterns are required, and multi-variable checks are required if predictions of the behaviour of individual subsystems within a catchment is needed. Thus, a model should only be assumed valid with respect to outputs that have been explicitly validated. This means for instance that a model which is validated against catchment runoff cannot automatically be assumed valid also for simulation of erosion on a hillslope within the catchment, because smaller scale processes may dominate here; it will need validation against hillslope soil erosion data.

From a theoretical point of view the procedures outlined by Klemes [19] for the proxy-basin and the differential split-sample tests, where tests have to be carried out using data from similar catchments, are weaker than the usual split-sample test, where data from the specific catchment are available. However, no obviously better testing schemes exist. Therefore, this will have to be reflected in the performance criteria in terms of larger expected uncertainties in the predictions. It must be realised that the validation test schemes proposed above are so demanding that many applications today would fail to meet them. Thus, for many cases where either proxy-basin and differential split-sample tests are required, suitable test data simply do not exist. This is for example the case for prediction of regional scale transport of potential contamination from underground radionuclide deposits over the next thousands of years. In such case model validation is not possible. This does not imply that these modelling studies are not useful, only that their output should be recognised to be somewhat more uncertain than is often stated and that the term 'validated model' should not be used. Thus, a model’s validity will always be confined in terms of space, time, boundary conditions, types of application, etc.

According to the methodology, model validation implies substantiating that a site-specific model can produce simulation results within the range of accuracy specified in the performance criteria for the particular study. Hence, before carrying out the model calibration and the subsequent validation tests quantitative performance criteria must be established. In determining the acceptable level of accuracy a trade-off will, either explicitly or implicitly, have to be made between costs, in terms of data collection and modelling work, and associated benefits that can be obtained due to more accurate model results. Consequently, the acceptable level of accuracy will vary from case to case and must be seen in a socio-economic context. It should therefore usually not be defined by the modeller, but in a dialogue between the modeller and the manager.

4.3. Need for interaction between manager, code developer and modeller

As discussed above, the validation methodologies presently used, even in research projects, are generally not rigorous and far from satisfactory. At the same time models are being used in practise and daily claims are being made on validity of models and on the basis of, at the best, not very strict and rigorous test schemes. An important question then, is how can the situation be improved in the future? As emphasised by Forkel [12] improvements cannot be achieved by the research community alone, but requires an interaction between the three main ‘players’, namely water resources managers, code developers and model users (modellers).

The key responsibilities of the water resources manager are to specify the objectives and define the acceptance limits of accuracy performance criteria for the model application. Furthermore, it is the manager’s responsibility to define requirements for code verification and model validation. In many consultancy jobs accuracy criteria and validation requirements are not specified at all, with the result being that the model user implicitly defines them in accordance with the achieved model results. In this respect it is important in the terms of references for a given model application to ensure consistency between the objectives, the specified accuracy criteria and validation requirements are not associated benefits that can be obtained due to more accurate model results. Consequently, the acceptable level of accuracy will vary from case to case and must be seen in a socio-economic context. It should therefore usually not be defined by the modeller, but in a dialogue between the modeller and the manager.

4.3. Need for interaction between manager, code developer and modeller

As discussed above, the validation methodologies presently used, even in research projects, are generally not rigorous and far from satisfactory. At the same time models are being used in practise and daily claims are being made on validity of models and on the basis of, at the best, not very strict and rigorous test schemes. An important question then, is how can the situation be improved in the future? As emphasised by Forkel [12] improvements cannot be achieved by the research community alone, but requires an interaction between the three main ‘players’, namely water resources managers, code developers and model users (modellers).

The key responsibilities of the water resources manager are to specify the objectives and define the acceptance limits of accuracy performance criteria for the model application. Furthermore, it is the manager’s responsibility to define requirements for code verification and model validation. In many consultancy jobs accuracy criteria and validation requirements are not specified at all, with the result being that the model user implicitly defines them in accordance with the achieved model results. In this respect it is important in the terms of references for a given model application to ensure consistency between the objectives, the specified accuracy criteria, the data availability and the financial resources. In order for the manager to make such evaluations, some knowledge on the modelling process is required.

The model user has the responsibility for selection of a suitable code as well as for construction, calibration and validation of the site-specific model. In particular, the model user is responsible for preparing validation documents in such a way that the domain of applicability and the range of accuracy of the model are explicitly specified. Furthermore, the documentation of the modelling process should ideally be done in enough detail that it can be repeated several years later, if required. The model user has to interact with the water resources manager on assessments of realistic model accuracies. Furthermore, the model user must be aware of the capabilities and limitations of the selected code and interact with the code developer with regard to reporting of user experience such as shortcomings in documentation, errors in code, market demands for extensions, etc.
The key responsibilities of the developer of the model code are to develop and verify a model code. In this connection it is important that the capabilities and limitations of the code appear in the documentation. As code development is a continuous process, code maintenance and regular updating with new versions improved as a response to user reactions become important. Although a model code should be comprehensively documented, there will in practise always occur doubts once in a while on its functioning, even for experienced users. Hence, active support to and dialogue with model users are crucial for ensuring operational model applications at a high professional level.

4.4. Performance criteria—when is a model good enough?

A critical issue in relation to the methodological framework is how to define the performance criteria. We agree with Beven [9] that any conceptual model is known to be wrong and hence any model will be falsified if we investigate it in sufficient detail and specify very high performance criteria.

Clearly, if one attempts to establish a model that should simulate the truth it would always be falsified. However, this is not a very useful information. Therefore, we are using the conditional validation, or the validation restricted to domain of applicability (or numerical universal as opposed to strictly universal in Popperian terms). The good question is then what is good enough? Or in other words what are the criteria? How do we select them?

A good reference for model performance is to compare it with uncertainties of the available field observations. If the model performance is within this uncertainty range we often characterise the model as good enough. However, usually it is not so simple. How wide confidence bands do we accept on observational uncertainties—ranges corresponding to 65%, 95% or 99%? Do we always then reject a model if it cannot perform within the observational uncertainty range? In many cases even results from less accurate models may be very useful.

Therefore, our answer is that the decision on what is good enough generally must be taken in a socio-economic context. For instance, the accuracy requirements to a model to be used for an initial screening of alternative options for location of a new small well field for a small water supply will be much smaller than the requirements to a model that is intended to be used for the final design of a large well field for a major water supply in an area with potential damaging effects on precious nature and other significant conflicts of interests. Thus, we believe that the accuracy criteria cannot be decided universally by modellers or researchers, but must be different from case to case depending on how much is at stake in the decision to depend on the support from model predictions. This implies that the performance criteria must be discussed and agreed between the manager and the modeller beforehand. However, as the modelling process and the underlying study progresses with improved knowledge on the data and model uncertainties as well as on the risk perception of the concerned stakeholders it may well be required to adjust the performance criteria in a sort of adaptive project management context [27].

4.5. The role of uncertainty assessments

Should we then trust a model if it happens to pass a validation test? Are we sure that this model is the best one and that the underlying conceptual basis and input data are basically correct?

Yes on the one hand, in such case we may trust a model as a suitable tool to make predictions through model simulations. But on the other hand, we can never be sure that a model that passes a validation test will have a sound conceptual basis. It could be right for the wrong reasons, e.g. by compensating error in conceptual model (model structure) with errors in parameter values.

And we know that it would be possible to find many other models that can pass the validation test, and that it would not be possible beforehand to identify one of these models as the best one in all respects. Having realised this equifinality problem the relevant question is what we should do to address it in practical cases. In this respect our framework prescribes that model predictions (see definition of ‘simulation’ in Section 3) made subsequent to passing a validation test should include uncertainty assessments. Hence, we basically agree with Beven [9] that uncertainty assessments are necessary, and that such uncertainty analyses should include uncertainty on model structure, parameter values etc. Different methodologies exist for conducting uncertainty assessments, e.g. Beven [8] and Van Asselt and Rotmans [36].

5. Guiding principles and future perspectives for modelling guidelines

5.1. Guiding principles

In our opinion the two key factors causing the poor quality of the modelling work in practise are: (a) too poor quality of the modelling work done by practitioners (inadequate use of guidelines and quality assurance procedures and inadequate role play between manager (client) and modeller (consultant)) and (b) lack of data and methodology in the hydrological science. Modelling guidelines like [25,37] almost exclusively address the former issue while scientific literature like [7,9] focus on the latter issue. In our opinion it is crucial that the two lines of action are combined. This implies that we need to define modelling guidelines that are both operational
in practise and scientifically founded. The framework we have described here attempts to establish one such a bridge between the two fields, i.e. pragmatic modelling and natural science. An important aspect of this framework is in a scientifically consistent way to enable the manager and the modeller to make the compromises that are required in practise.

On this background the following five key principles for pragmatic modelling have emerged:

- **A terminology** that is internally consistent. We acknowledge that many authors in the scientific literature use different terminology and that, in particular, some authors do not use the terms verification and validation. However, these terms are also widely used, and we need in practise to have understandable terms for these operations. Thus, with the clear distinction between conceptual model, model code and site-specific model and the restrictions to domains of applicability (numerical universal in Popperian sense) we believe that our terminology is in accordance with the main stream of scientific philosophy.

- **We never talk about universal code verification or universal model validation**, but always restrict these terms to clearly defined domains of applicability. This is a necessary assumption for the consistency of the terminology and methodology and must be emphasised explicitly in any guidelines.

- **Validation tests** against independent data that have not also been used for calibration are necessary in order to be able to document the predictive capability of a model.

- **Model predictions achieved through simulation** should be associated with **uncertainty assessments** where amongst others the uncertainty in model structure and parameter values should be accounted for.

- **A continuous interaction between manager and modeller** is crucial for the success of the modelling process. One of the key aspects in this regard is to establish suitable performance criteria for the model calibration and validation tests. This dialogue is also very important in connection with uncertainty assessments.

5.2. **Future challenges**

Some of the issues dealt with in the present manuscript are still not fully explored. The four most important future challenges are:

- **Establishment of accuracy criteria** for a modelling study is a very important issue and one where we maybe differ from most scientific literature. Modellers often establish numerical accuracy criteria in order to classify the goodness of a given model [2,17,28]. These attempts are very useful in making the performance more transparent and quantitative, but do not provide an objective means to decide what the optimal accuracy criteria really should be in a given case. According to our framework no universal accuracy criteria can be established, i.e. it is generally not possible from a natural scientific point of view to tell when a model performance is good enough. Such acceptance criteria will vary from case to case depending on the socio-economic context, i.e. what is at stake in the decisions to be supported by the model predictions. The good question now is: how do we translate the ‘soft’ socio-economic objectives to ‘hard-core’ model performance criteria? This is obviously a challenge that cannot be solved by natural science alone, but need to be addressed in a much broader context including aspects of economy, stakeholder interests and risk perception. Until we become better to overcome this challenge we will, however, not be able to arrive at the optimal balance between the costs of modelling and the derived societal benefits. Although this work has hardly begun yet, and we know that it is a very difficult road, we see no real alternative.

- **Although all experience shows that models generally perform poorer in validation tests against independent data than they do in calibration tests, model validation** is in our opinion a much neglected issue, both in many modelling guidelines and in the scientific literature. Maybe many scientists have not wanted to use the term validation due to the scientific philosophically related controversies, but in any case many scientists are not advocating the need for model validation. One of the unfortunate consequences of this ‘lack of interest’ is that not much work has been devoted to developing suitable validation test schemes since Klemes [19]. In our opinion further development of suitable testing schemes and imposing them to all modelling projects is a major future challenge.

- A third issue that requires considerable attention is how do we decide among alternative model structures and parameter sets (the equifinality problem). If we use multiple criteria one model may be better on one criteria and another on another criteria. In our opinion we need not necessarily chose. We know that all conceptual models are wrong and we know that wrong conceptual models are compensated by biased model parameter values through calibration. But, unless we can falsify a conceptual model directly, which is very difficult, or unless the resulting model is falsified through the validation test, this model is a possible candidate for predictions. And if several models pass the validation tests we may not be able to tell which one is the best. In such case they should all be considered suitable, and the fact that they provide different predictive results should be used as part of the uncertainty assessments. Work on this relatively
A new paradigm has just begun [9] and a lot of work is still required to further develop and operationalise it.

Finally, there are many more challenges related to uncertainty in water resources management. Quality assurance and uncertainty assessments are two aspects that are very closely linked. Initially, the manager has to define accuracy criteria from a perception of which uncertainty level he believes is suitable in a particular case (see above). Subsequently, as the modelling study proceeds, the dialogue between modeller and manager has to continue with the necessary trade-off between modelling accuracy and cost of modelling study. In the uncertainty assessments it is very important to go beyond the traditional statistical uncertainty analysis. Thus, e.g. aspects of scenario uncertainty and ignorance should generally be included and in addition the uncertainties originating from data and models often needs to be integrated with socio-economic aspects in order to form a suitable basis for the further decision process [36]. Thus, like with the accuracy criteria (above) the use of uncertainty assessments in water resources management goes beyond natural science.

Acknowledgements

The present work was carried out within the Project ‘Harmonising Quality Assurance in model based catchments and river basin management (HarmoniQuA)’, which is partly funded by the EC Energy, Environment and Sustainable Development programme (Contract EVK2-CT2001-00097). The constructive comments and suggestions to the manuscript by the HarmoniQuA project team and by our colleague William (Bill) G. Harrar are acknowledged. Finally, the constructive criticisms by Keith Beven, University of Lancaster; Rodger Grayson, University of Melbourne and a third, anonymous referee helped to improve the manuscript significantly.

References


