Available online at www.sciencedirect.com
ScienceDirect
Advances in Water Resources 29 (2006) 1586–1597
Advances in Water Resources
www.elsevier.com/locate/advwatres

A framework for dealing with uncertainty due to model structure error

Jens Christian Refsgaard [a,*], Jeroen P. van der Sluijs [b], James Brown [c], Peter van der Keur [a]

[a] Department of Hydrology, Geological Survey of Denmark and Greenland (GEUS), Oster Voldgade 10, 1350 Copenhagen, Denmark
[b] Copernicus Institute for Sustainable Development and Innovation, Department of Science Technology and Society, Utrecht University, Utrecht, The Netherlands
[c] University of Amsterdam (UVA), Amsterdam, The Netherlands

Received 29 July 2004; received in revised form 6 September 2005; accepted 21 November 2005
Available online 5 January 2006

Abstract

Although uncertainty about structures of environmental models (conceptual uncertainty) is often acknowledged to be the main source of uncertainty in model predictions, it is rarely considered in environmental modelling. Rather, formal uncertainty analyses have traditionally focused on model parameters and input data as the principal source of uncertainty in model predictions. The traditional approach to model uncertainty analysis, which considers only a single conceptual model, may fail to adequately sample the relevant space of plausible conceptual models. As such, it is prone to modelling bias and underestimation of predictive uncertainty.

In this paper we review a range of strategies for assessing structural uncertainties in models. The existing strategies fall into two categories depending on whether field data are available for the predicted variable of interest. To date, most research has focussed on situations where inferences on the accuracy of a model structure can be made directly on the basis of field data. This corresponds to a situation of 'interpolation'. However, in many cases environmental models are used for 'extrapolation'; that is, beyond the situation and the field data available for calibration. In the present paper, a framework is presented for assessing the predictive uncertainties of environmental models used for extrapolation. It involves the use of multiple conceptual models, assessment of their pedigree and reflection on the extent to which the sampled models adequately represent the space of plausible models.

Keywords: Environmental modelling; Model error; Model structure; Conceptual uncertainty; Scenario analysis; Pedigree

1. Introduction

1.1. Background

Assessing the uncertainty of model simulations is important when such models are used to support decisions about water resources [6,33,23,39]. The key sources of uncertainty in model predictions are (i) input data; (ii) model parameter values; and (iii) model structure (=conceptual model). Other authors further distinguish uncertainty in model context, model assumptions, expert judgement and indicator choice [46,54,48] but these are beyond the scope of this paper. Uncertainties due to input data and due to parameter values have been dealt with in many studies, and methodologies to deal with these are well developed. However, no generic methodology exists for assessing the effects of model structure uncertainty, and this source of uncertainty is frequently neglected.

Any model is an abstraction, simplification and interpretation of reality. The incompleteness of a model

* Corresponding author. Tel.: +45 38 14 27 76; fax: +45 38 14 20 50.
E-mail address: jcr@geus.dk (J.C. Refsgaard).

0309-1708/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2005.11.013

structure and the mismatch between the real causal structure of a system and the assumed causal structure as represented in a model always result in uncertainty about model predictions. The importance of the model structure for predictions is well recognised, even for situations where predictions are made on output variables, such as discharge, for which field data are available [16,8]. The considerable challenge faced in many applications of environmental models is that predictions are required beyond the range of available observations, either in time or in space, e.g. to make extrapolations towards unobservable futures [2] or to make predictions for natural systems, such as ecosystems, that are likely to undergo structural changes [4]. In such cases, uncertainty in model structure is recognised by many authors to be the main source of uncertainty in model predictions [44,13,31,28].

1.2. An example – five alternative conceptual models

The problem is illustrated for a study conducted by the County of Copenhagen in 2000 involving a real water management decision [11,37]. The County of Copenhagen is the authority responsible for water resources management in the county where the city of Copenhagen abstracts groundwater for most of its water supply. According to a new Water Supply Act the county had to prepare an action plan for protection of groundwater against pollution. As a first step, the county asked five groups of Danish consulting firms to conduct studies of the aquifer's vulnerability towards pollution in a 175 km^2 area west of Copenhagen, where the groundwater abstraction amounts to about 12 million m^3/year. The key question to be answered was: which parts of this particular area are most vulnerable to pollution and need to be protected? The five consultants were among the most well reputed consulting firms in Denmark, and they were known to have different views and preferences on which methodologies are most suitable for assessing vulnerability. As the task was one of the first consultancy studies on a new major market for preparation of groundwater protection plans it was considered a prestigious job to which the consultants generally allocated some of their most qualified professionals.

The five consultants used significantly different approaches. One consultant based his approach on annual fluctuations of piezometric heads assuming that larger fluctuations represent greater interaction between aquifer and surface water systems and hence a larger vulnerability. Several consultants used the DRASTIC multi-criteria method [1], but modified it in different ways by changing weights and adding new, mainly geochemically oriented, criteria. One consultant based his approach on advanced hydrological modelling of both groundwater and surface water systems using the MIKE SHE code [40], while two other consultants used simpler groundwater modelling approaches. Thus, the five consultants had different perceptions of what causes groundwater pollution and used models with different processes and causal relationships to describe the possibility of groundwater pollution in the area. In addition, their different interpretations and interpolations made from common field data resulted in significantly different figures for e.g. areal means of precipitation and evapotranspiration and the thickness of various geological layers [37].

The conclusions of the five consultants regarding vulnerability to nitrate pollution are shown in Fig. 1. It is apparent that the five estimates differ substantially from each other. In the present case, no data exist to validate the model predictions, because the five models were used to make extrapolations. Thus, it is not possible, from existing field data, to tell which of the five model estimates are more reliable. The differences in prediction originate from two main sources: (i) data and parameter uncertainty and (ii) conceptual uncertainty. Although the data and parameter uncertainties were not explicitly assessed by any of the consultants (as is common in such studies), the substantial differences in model structures and the fact that the consultants all used the same raw data point to structural uncertainty as the main cause of difference between the five model results and as a major source of uncertainty in model predictions.

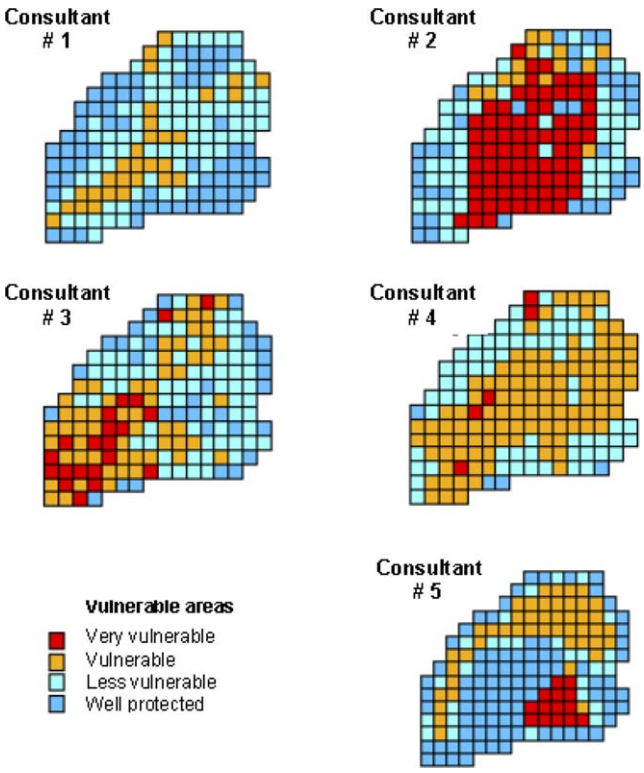

Fig. 1. Model predictions on aquifer vulnerability towards nitrate pollution for a 175 km^2 area west of Copenhagen [11].

Usually a water manager bases their decisions on the conclusions from only one study. The uniqueness of the present study was that five consultants were asked to answer the same question on the basis of the same data. In this respect the differences between the five estimates are striking and clearly do not provide a sound basis for deciding anything about which areas should be protected. A worrying question, which is left unanswered, is whether the basis for decisions is similarly poor in the many other cases where only a single conceptual model has been adopted and where millions of DKK have subsequently been used to prepare and implement action plans.

1.3. Objective and outline of paper

The objective of this paper is to review possible strategies for dealing with model structure errors and to outline a framework for handling the effects of model structure errors on predictive uncertainty, with particular emphasis on situations where model predictions represent extrapolations to situations not covered by calibration data and are often outside the domain on which our knowledge on the dynamics of the system and our understanding of its causal relationships is based.

The paper is organised so that reviews of existing strategies and the discussion of their potentials and limitations are given in Section 2. A new framework is presented in Section 3 for analysing the uncertainties due to model structure errors when models are used for making extrapolations beyond their calibration base. Finally, the problems and perspectives of the new framework are discussed in Section 4. The terminology used is defined in Appendix.

2. Review of possible strategies

2.1. Classification

The existing strategies for assessing uncertainty due to incomplete or inadequate model structure may be grouped into the categories shown in Fig. 2. The most important distinction is whether data exist that makes it possible to make inferences on the model structure uncertainty directly. This requires that data are available for the output variable of predictive interest and for conditions similar to those in the predictive situation. In other words it is a distinction between whether the model predictions can be considered as interpolations or extrapolations relative to the calibration situation.

The two main categories are thus equivalent to different situations with respect to model validation tests. According to Klemes' classical hierarchical test scheme [26,38], the interpolation case corresponds to situations where the traditional split-sample test is suitable, while the extrapolation case corresponds to situations where no data exist for the concerned output variable (proxy-basin test) or where the basin characteristics are considered non-stationary, e.g. for predictions of effects of climate change or effects of land use change (differential split-sample test).

In the review of existing strategies given below examples of studies have been selected to illustrate the classification and the common approaches. It is not an

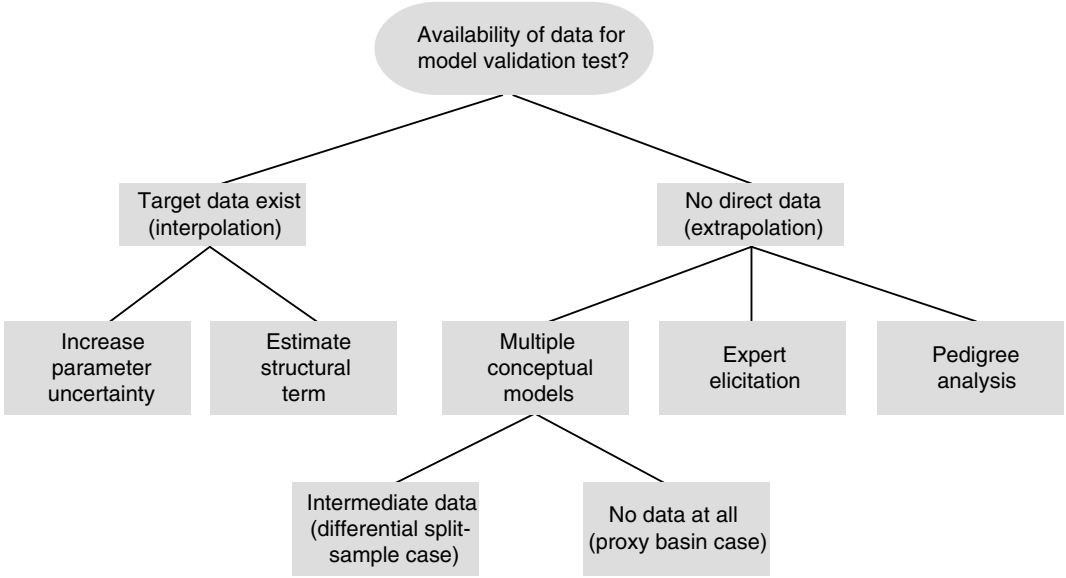

Fig. 2. Classification of existing strategies for assessing conceptual model uncertainty.

exhaustive review, but illustrates the range of approaches available to diagnose structural uncertainty in models.

2.2. Data exist – interpolation

In this situation, calibration is usually carried out against a sample of the existing field data to ensure some kind of optimal parameter values, and then the model predictions are compared with the remaining ('independent') field data. The deviations between model predictions and independent field observations can be used to infer the model's conceptual error. Different methodologies can be used in this respect.

2.2.1. Increasing parameter uncertainty to account for structural uncertainty

One strategy is to increase the parameter uncertainty to a level where it is assumed to compensate for omitting model structure error from the analysis. Van Griensven and Meixner [45] provide an example of this. They assess the total predictive uncertainty without identifying or quantifying the underlying sources of uncertainty. They use the split-sample approach assessing ranges of predictive uncertainty from analyses of predictions and data for a period different from the calibration period. Their total predictive uncertainty is assessed by increasing the model parameter uncertainty beyond the magnitudes estimated during calibration to a level where the resulting predictive uncertainty intervals bracket the observations. This technique does not introduce a separate stochastic term for the structural uncertainty, but represents the structural term in the parameter term. The model structure error is likely to influence the model simulations in non-random and temporally varying ways. By compensating the model structure error by increasing the variance of a temporally constant random variable the results from this approach can be questioned, particularly if used for predictions in situations where split-sample tests are not made.

2.2.2. Estimation of the structural uncertainty term

Other strategies attempt to estimate the structural contribution to uncertainty in the model predictions. An example of such an approach is given by Radwan et al. [35], who estimate the total predictive uncertainty from a statistical analysis of the residuals between model predictions and observations. Further, they analyse the propagated uncertainties from model input and parameter values. By subtracting these two uncertainties from the total predictive uncertainty they assign the remaining predictive uncertainty to be an effect of model structure uncertainty. It is then possible to add the model structure uncertainty when making other predictions. This approach assumes that the uncertainties from different sources are additive. This assumption is questionable, because the combination of uncertainties is often non-linear due to interactions, correlations and dependencies between variables in a model. It also assumes that the differences in predictions and observations are caused by structural error and not by the poor specification of input and parameter uncertainty, nor by errors in the observations.

Vrugt et al. [53] present another stochastic approach based on a simultaneous parameter optimisation and data assimilation with an ensemble Kalman filter. By specifying values for measurement error and a so-called 'stochastic forcing term', representing structural uncertainty, they are able to estimate the dynamic behaviour of the model structure uncertainty. Both techniques assume a smooth contribution from structural uncertainty, but an important advantage of the latter is that parameter innovations (an output from the Kalman filter) may be used to diagnose non-stationarity in system structure.

2.3. No direct data – extrapolation

In cases where model structure errors cannot be assessed directly due to a lack of relevant data, the main strategy is to do the extrapolation with multiple conceptual models. Two supporting methods can be used here for the generation and qualification of each of the alternative models: expert elicitation and pedigree analysis (Fig. 2).

2.3.1. Multiple conceptual models

In the scenario approach a number of alternative conceptual models are considered. For each of these, the model input and parameter uncertainties may be analysed and the differences between model predictions are then seen as a measure of the model structure uncertainty. The idea of using alternative or competing candidate model structures was introduced in water quality modelling some time ago [5]. The issue typically dealt with here is whether models developed for current conditions can yield correct predictions when used under changed control. Van Straten and Keesman [50] note in this respect that good performance at the calibration stage does not guarantee correctly predicted behaviour, due to non-stationarity of the underlying processes in space or time.

The multiple modelling approach has also been used in flood forecasting. For example, Butts et al. [8] use 10 different model structures to evaluate structural uncertainty in flood predictions. They conclude that exploring an ensemble of model structures provides a useful approach in assessing simulation uncertainty.

In groundwater modelling different conceptual models are typically based on different geological interpretations [18,43,42,30,34]. Højberg and Refsgaard [21] present an example using three different conceptual

models, based on three alternative geological interpretations for a multi-aquifer system in Denmark. Each of the models was calibrated against piezometric head data using inverse technique. The three models provided equally good and very similar predictions of groundwater heads, including well field capture zones. However, when using the models to extrapolate beyond the calibration data to predictions of flow pathways and travel times the three models differed dramatically. When assessing the uncertainty contributed by the model parameter values, the overlap of uncertainty ranges between the three models significantly decreased when moving from groundwater heads to capture zones and travel times. They conclude that the larger the degree of extrapolation, the more the underlying conceptual model dominates over the parameter uncertainty and the effect of calibration.

The strategy of applying several alternative models based on codes with different model structures is also common in climate change modelling. In its description of uncertainty related to model predictions of both present and future climates the Intergovernmental Panel on Climate Change (IPCC) [22] bases its evaluation on scenarios of many (up to 35) different models. The same strategy is followed in the dialogue model [52]. Dialogue is a so-called integrated assessment model (IAM) of climate change. It has been developed as an interactive decision-support tool for energy supply policy making. Dialogue simulates the cause effect chain of climate change, using mono-disciplinary sub-models for each step in the chain. The chain starts with scenarios for economic growth, energy demand, fuel mix etc., leading to emissions of greenhouse gasses, leading to changes in atmospheric composition, leading to radiative forcing of the climate, leading to climate change, leading to impacts of climate change on societies and ecosystems. Rather than selecting one mono-disciplinary sub-model for each step, as most other climate IAMs do, dialogue uses multiple models for each step (for instance, three different carbon cycle models, simplified versions of five different global climate model – outcomes, etc.), representing the major part of the spectrum of expert opinion in each discipline.

2.3.2. Expert elicitation

Expert elicitation can be used as a supporting method in uncertainty analysis. It is a structured process to elicit subjective judgements and ideas from experts. It is widely used in uncertainty assessment to quantify uncertainties in cases where there is no or too few direct empirical data available to infer uncertainty. Usually the subjective judgement is represented as a probability density function reflecting the experts' degree of belief. Expert elicitation aims to specify uncertainties in a structured and documented way, ensuring the account is both credible and traceable to its assumptions. Typically it is applied in situations where there is scarce or insufficient empirical material for a direct quantification of uncertainty [20]. An example with use of expert elicitation to estimate probabilities of alternative conceptual models is given by Meyer et al. [29]. They assessed probabilities as subjective values, from expert elicitation, reflecting a belief about the relative plausibility of each model based on its apparent consistency with available knowledge and data.

Expert elicitation can also be used to generate ideas about alternative causal structures (conceptual models) that govern the behaviour of a system. Techniques used in decision analysis include group model building [51] and the hexagon method [19] but these techniques usually aim to achieve consensus. From the point of view of model structure uncertainty, these elicitation techniques could perhaps be used to generate alternative conceptual models.

2.3.3. Pedigree analysis

Another supporting method is pedigree analysis. The idea comes from Funtowicz and Ravetz [17], who note that statistical uncertainty in terms of inexactness does not cover all relevant dimensions of uncertainty, including the methodological and epistemological dimensions. To promote a more differentiated insight into uncertainty they propose to extend good scientific practice with five qualifiers for quantitative scientific information: numeral unit, spread, assessment, and pedigree (NUSAP). By adding expert judgement of reliability (assessment) and systematic multi-criteria evaluation of the processes by which numbers have been produced (pedigree), NUSAP has extended the statistical approach to uncertainty (inexactness) with the methodological (unreliability) and epistemological ignorance dimensions. By providing a separate qualification for each dimension of uncertainty, it enables flexibility in their expression.

Each special sort of information has its own aspects that are key to its pedigree, so different pedigree matrices using different pedigree criteria can be used to qualify different sorts of information. Early applications of pedigree analysis of environmental models have focussed on parameter pedigree, using proxy representation, empirical basis, methodological rigor, theoretical understanding and validation as pedigree criteria. Later on, pedigree analysis has been extended to assessment of model assumptions and problem framing [49,12].

2.4. Discussion of strengths/weaknesses and potentials/limitations

The strategies used in 'interpolation', i.e. for situations that are similar to the calibration situation with respect to variables of interest and conditions of the natural system, have the advantage that they can be based directly on field data. A fundamental weakness is that

field data are themselves uncertain. Nevertheless, in many cases, they can be expected to provide relatively accurate estimates of, at least, the total predictive uncertainty for the specific measured variable and for the same conditions as those in the calibration and validation situation. Some of the methods cannot differentiate how the total predictive uncertainty originates from model input, model parameter and model structure uncertainty. Other methods attempt to do so. However, this distinction is, as recognised by many authors, e.g. Vrugt et al. [53], problematic. In the case of uncalibrated models, the parameter uncertainty is very difficult to assess quantitatively, and wrong estimates of model parameter uncertainty will influence the estimates of model structure uncertainty. In the case of calibrated models, estimates of model parameter uncertainty can often be derived from autocalibration routines. An inadequate model structure will, however, be compensated by biased parameter values to optimise the model fit with field data during calibration. Hence, the uncertainty due to model structure will be underestimated in this case.

A more serious limitation of the strategies depending on observed data is that they are only applicable for situations where the output variables of interest are measured (e.g. [35,45,53]). While relevant field data are often available for variables such as water levels and water flows, this is usually not the case for concentrations, or when predictions are desired for scenarios involving catchment change, such as land use change or climate change. Another serious limitation stems from an assumption that the underlying system does not undergo structural changes, such as changes in ecosystem processes due to climate change.

The strategy that uses multiple conceptual models benefits from an explicit analysis of the effects of alternative model structures. Furthermore, it makes it possible to include expert knowledge on plausible model structures. This strategy is strongly advocated by Neuman and Wierenga [31] and Poeter and Anderson [34]. They characterise the traditional approach of relying on a single conceptual model as one in which plausible conceptual models are rejected (in this case by omission). They conclude that the bias and uncertainty that results from reliance on an inadequate conceptual model are typically much larger than those introduced through an inadequate choice of model parameter values.

This view is consistent with Beven [7] who outlines a new philosophy for modelling of environmental systems. The basic aim of his approach is to extend traditional schemes with a more realistic account of uncertainty, rejecting the idea that a single optimal model exists for any given case. Instead, environmental models may be non-unique in their accuracy of both reproduction of observations and prediction (i.e. unidentifiable or equifinal), and subject to only a conditional confirmation, due to e.g. errors in model structure, calibration of parameters and period of data used for evaluation. A weakness of the multiple modelling strategy, is the absence of quantitative information about the extent to which each model is plausible. Furthermore, it may be difficult to sample from the full range of plausible conceptual models. In this respect, expert knowledge on which the formulations of multiple conceptual models are based, is an important and unavoidable subjective element. The level of subjectivity can be reduced if the scenarios are generated in a formalised and reproducible manner. For example, this is possible with the TPROGS procedure [9,10], by which alternative geological models can be generated stochastically. The subjectivity does not disappear with this approach. Rather, it is transferred from formulation of the geological model itself to assumptions on probability functions and correlation structures of the various geological units that are more easily constrained in practice.

The strategy of expert elicitation has the advantage that subjective expert knowledge can be included in the evaluation. It has the potential to make use of all available knowledge including knowledge that cannot be easily formalised otherwise. It can include views of sceptics, and reveals the level of expert disagreement on certain estimates. Expert elicitation also has several limitations. The fraction of experts holding a given view is not proportional to the probability of that view being correct. One may safely average estimates of model parameters, but if the expert's models were incommensurate, one cannot average models [25]. If differences in expert opinion are irresolvable, weighing and combining the individual estimates of distributions is impossible. In practice, the opinions are often weighted equally, although sometimes self-rating is used to obtain a weight-factor for the experts competence. Finally, the results of expert elicitation tend to be sensitive to the selection of the experts whose estimates are gathered.

In a review of four different case studies in which pedigree analysis was applied, Van der Sluijs et al. [49] show that pedigree analysis broadens the scope of uncertainty assessment and stimulates scrutiny of underlying methods and assumptions. Craye et al. [12] reported similar experiences. It facilitates structured, creative thinking on conceivable sources of error and fosters an enhanced appreciation of the issue of quality in information. It thereby enables a more effective criticism of quantitative information by providers, clients, and also users of all sorts, expert and lay. It provides differentiated insight in what the weakest parts of a given knowledge base are. It is flexible in its use and can be used on different levels of comprehensiveness: from a 'back of the envelope' sketch based on self-elicitation to a comprehensive and sophisticated procedure involving structured informed in-depth group discussions, covering each pedigree criterion. The scoring of pedigree criteria is to a certain

degree subjective. Subjectivity can partly be remedied by the design of unambiguous pedigree matrices and by involving multiple experts in the scoring. The choice of experts to do the scoring is also a potential source of bias. The method is relatively new, with a limited (but growing) number of practitioners. There is as yet no settled guideline for good practice. We must keep in mind that it is not a panacea for the problem of unquantifiable uncertainty.

3. New framework

We propose that conceptual uncertainty can be assessed by adopting a protocol based on the six elements shown in Fig. 3. The central aim is to establish a number of plausible conceptual models, with a range that adequately samples the space of possible conceptual models, to evaluate the tenability of each conceptual model and the overall range of models selected in relation to the perceived uncertainty on model structure and to propagate the uncertainties in each case.

STEP 1: Formulate a conceptual model. A conceptual model is established. Since we have defined a conceptual model as a combination of our qualitative process understanding and the simplifications acceptable for a particular modelling study, a conceptual model becomes highly site-specific and even case-specific. For example a conceptual model of an aquifer may be described as two-dimensional for a study focussing on regional groundwater heads, while it may need to include three-dimensional geological structures for detailed simulation of contaminant transport. Formulating a new conceptual model may involve changing or refining the model structure, e.g. by modifying the hydrogeological interpretations (in the case of groundwater models), dimensionality, temporal and spatial resolution, initial and boundary conditions and process descriptions (governing equations).

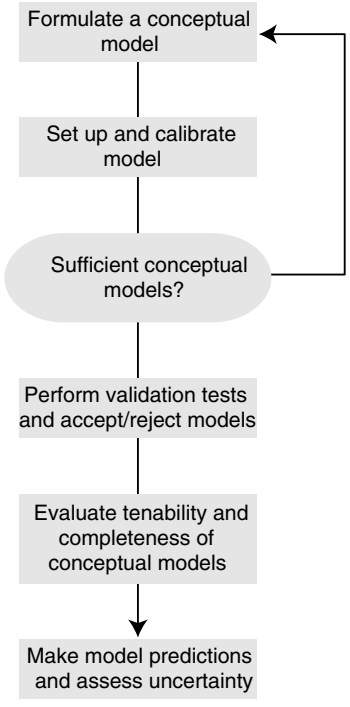

Fig. 3. Protocol for assessing conceptual model uncertainty.

STEP 2: Set up and calibrate model. On the basis of the formulated conceptual model a site- and case-specific model is set up. Subsequently the model is calibrated and the model parameter uncertainty assessed. For the purposes of 'interpolation' (i.e. relevant observations are available), the parameter uncertainty can reasonably be constrained through calibration. However, for the case of 'extrapolation', the risk of calibrating model parameters for prediction of unobserved variables is that the model becomes biased for the unobserved variable.

STEP 3: Sufficient conceptual models? The first two steps are repeated until sufficient conceptual models are included. This judgement will be influenced by the practical constraints on including additional models and the desire to include additional conceptual models that are substantially different from those already included.

STEP 4: Perform validation tests (to the extent data availability allows). In order to evaluate how well the models describe the system in question, the performances of each of the models are tested by comparing model predictions with independent field data, i.e. data not used for calibration. This may be achieved by splitting the sample data into a calibration and validation set, or, alternatively, by cross-validation (e.g. bootstrapping: [15]) against 'independent data'. The models whose predictive capability is deemed low are discarded and the reasons for these predictive failures are explored, where possible, for insight into the origins of structural uncertainty. In 'extrapolation' cases, data will usually not be available for validation tests and STEP 4 must be skipped. However, in some cases, it is possible to test 'intermediate' model results. For example a groundwater model aimed at prediction of concentration values can often be tested against groundwater head and discharge data, or sparse concentration data may be available for parts of the study area.

STEP 5: Evaluate tenability and completeness of conceptual models. The aim of this step is to analyse the retained models with respect to their predictive bias and uncertainty. This has two elements: (i) to evaluate the tenability of each conceptual model; and (ii) as far as possible, to evaluate the extent to which the retained models represent the space of plausible conceptual models. The tenability of the conceptual models is evaluated

Table 1
Pedigree matrix for evaluating the tenability of a conceptual model

Score	Supporting empirical evidence		Theoretical understanding	Representation of understood underlying mechanisms	Plausibility	Colleague consensus
	Proxy	Quality and quantity				
4	Exact measures of the modelled quantities	Controlled experiments and large sample direct measurements	Well-established theory	Model equations reflect high mechanistic process detail	Highly plausible	All but cranks
3	Good fits or measures of the modelled quantities	Historical/field data uncontrolled experiments small sample direct measurements	Accepted theory with partial nature (in view of the phenomenon it describes)	Model equations reflect acceptable mechanistic process detail	Reasonably plausible	All but rebels
2	Well correlated but not measuring the same thing	Modelled/derived data indirect measurements	Accepted theory with partial nature and limited consensus on reliability	Aggregated parameterised meta model	Somewhat plausible	Competing schools
1	Weak correlation but commonalties in measure	Educated guesses indirect approx. rule of thumb estimate	Preliminary theory	Grey box model	Not very plausible	Embryonic field
0	Not correlated and not clearly related	Crude speculation	Crude speculation	Black box model	Not at all plausible	No opinion

through expert reviews. First, the strength of the tenability of each conceptual model is evaluated by using the pedigree matrix in Table 1. A structured procedure for the elicitation of pedigree scores is given by Van der Sluijs et al. [47]. Note that there is no need to arrive at a consensus pedigree score for each criterion: if experts disagree on the pedigree scores for a given model, this reflects further epistemological uncertainty surrounding that model. Next, the adequacy of the retained conceptual models to represent the range of plausible models is evaluated. This is an assessment of whether the space of the retained conceptual models is sufficient to encapsulate the relevant range of plausible conceptual models without becoming impractical. This has strong similarities to Dunn's concept of context validation [14]. Context validity refers to the validity of inferences that we have estimated the proximal range of rival hypotheses. Context validation can be performed by a bottom-up process to elicit from experts rival hypotheses on causal relations governing the dynamics of a system. One could argue that an infinite number of conceivable models might exist. However, it has been shown in projects where such elicitation processes were used, that the cumulative distribution of unique rival models flattens out after consultation of a limited number of experts, usually somewhere between 20 and 25 when chosen with diverse enough backgrounds [27].

STEP 6: Make model predictions and assess uncertainty. Together with model predictions of the desired variables, uncertainty assessments are carried out. This will typically include uncertainty in input data and parameter values in addition to the conceptual uncertainty. Furthermore, on the basis of the goodness of the conceptual models, evaluated in STEP 5 the goodness of the assessed predictive uncertainty associated with the model structure should be evaluated.

4. Discussion and conclusions

4.1. Methodologies to assess conceptual uncertainty

As discussed above, the existing strategies fall into two main categories, each with limitations. The strategies where model structure errors are assessed from observed data are confined to interpolation cases, understood as cases where the model can be calibrated and validated against field data for the variables of predictive interest and where the natural system does not undergo structural change. The strategies used for situations involving extrapolation depend either on multiple conceptual models (preferred) or on expert elicitation or pedigree analysis for a single conceptual model (usually less preferred).

The novelty of our proposed framework is the combination of multiple conceptual models and the pedigree

approach for assessing the overall tenability of these models in one formalised protocol. Some of our proposed steps are similar to other approaches for dealing with equifinality, multiple possible models and the rejection of non-behavioural model [6,31]. Other steps are based on qualitative approaches, including expert knowledge in a structured manner [20,49]. The aim of our new framework is not to identify the ''true'' model structure or the cause of the errors in the existing model structure. Instead, we propose an approach that integrates different types of knowledge, not previously combined, such as quantitative and qualitative uncertainty, to estimate the impact of model structure uncertainty on model predictions.

The GLUE approach (generalised likelihood uncertainty estimation, [6,7]) also operates with a range of alternative models. Although almost all applications of GLUE reported so far operate with only one model structure and many alternative model parameter sets, it is possible to use GLUE with alternative model structures [24]. In addition to prescribing multiple conceptual models, an important difference between our proposed approach and GLUE is that we recommend parameter optimisation is conducted as part of the calibration in order to take full advantage of the information in field data. There are different opinions about whether calibration by parameter optimisation is advisable or not. The main advantage of calibration is that it improves the ability of the model to reproduce hydrological behaviour of a system within the limits of observed behaviour [31]. An important by-product is that it provides useful information about the uncertainty of model parameters. The disadvantage is that parameter optimisation may result in biased parameter values to compensate for errors in model structure and that many parameter sets (i.e. many models) perform more or less equally well but provide different results. In implementing our framework, model calibration might be skipped and many models with different parameter sets retained, as in the GLUE approach. The reason we are not advocating such an approach is partly for pragmatic reasons (very large computational requirements) and partly that we aim to focus on model structure uncertainty rather than parameter uncertainty.

Although intended for use in a very different context, the central aim behind our proposed protocol is similar to the approach of IPCC [22], who assign a level of confidence to their assessment of climate change by evaluating predictions from multiple models. The level of confidence placed in a particular finding reflects both the degree of consensus amongst modellers and the quantity of evidence that is available to support the finding. IPCC [22] classifies the confidence qualitatively in three levels: (i) 'well established', (ii) 'evolving' and (iii) 'speculative'.

4.2. Critical issues for implementing the new protocol

4.2.1. Performance criteria – threshold for accepting/rejecting models

A critical issue in relation to acceptance/rejection of models (STEP 4 above) is how to define performance criteria. We agree with Beven [7] that any conceptual model is (known to be) wrong in an absolute sense, and hence that any model will be rejected if we investigate it in sufficient detail and specify very high performance criteria. On the other hand, the whole point in modelling is to simplify.

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 may characterise the model as good enough. However, usually it is less straightforward. For example, how wide should the confidence bands be before we reject models or accept them within observational uncertainties – ranges corresponding to 65%, 95% or 99%? Indeed, the differences between 95% and 99% may be significant in practical terms. Do we always then reject a model if it cannot perform within the observational uncertainty range? How reasonable are our estimates of uncertainty in observations? In many cases, even the results from less accurate models may be very useful.

Another reference for what is acceptable accuracy is the use of a benchmark model as discussed by e.g. Seibert [41]. The difficulty is then transferred to selecting an appropriate benchmark.

Our answer is that the decision on performance criteria must, in general, be taken in a socio-economic context, for which predictive uncertainties must be clearly explained and open to interpretation beyond small groups of scientists. 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 the nature of a decision and the risks involved.

4.2.2. Qualitative assessment of tenability of conceptual models

Pedigree analysis structures the critical appraisal of alternative model structures and provides insight in the state of knowledge on which each of the conceivable model structures is based. However, it does not give an indication of the relative quality of the various model structures. With reference to Table 1, the pedigree analysis for a simple statistical model (A) and a complex mechanistic model (B) could, for example, result in statements like:

- Model A is weakly correlated to the predicted variable (Proxy, score 1), based on a large sample of direct measurements (Quality and quantity, score

4), built on a preliminary theory and a black box model (Theoretical understanding, score 1; Representation of mechanisms, score 1), somewhat plausible (Plausibility, score 2) and controversial among colleagues (Colleague consensus, score 2);
- Model B exactly addresses the desired predictive variable (Proxy, score 4), is based on data with rule of thumb estimates (Quality and quantity, score 1), built on a well-established theory with model equations reflecting high process details (Theoretical understanding, score 4; Representation of mechanisms, score 4), reasonably plausible and accepted by all colleagues except rebels (Plausibility and Colleague consensus, score 3).

Such statements cannot be integrated in a quantitative uncertainty analysis in terms of probabilities, but they should be available as the best possible scientifically based characterisation of uncertainties and as such be made available to those involved in the decision making process.

Furthermore, as the selected conceptual models can never cover all possibilities, but instead cover limited range, it is important to emphasise that the overall uncertainty of model predictions cannot be assessed in an absolute sense, only in a conditional or relative sense [7,31]. Our suggested method does not alter this fundamentally. However, we believe that the outcome of the proposed formalised review is a qualitative assessment that is more useful in a decision making context than unstructured information, or verbose information from scientific outlets that is not always available to the decision maker. The challenge is to design environmental management strategies that are robust against the uncertainties identified. Inclusion of a wider range of conceivable model structures may help to anticipate surprises that would have been overlooked otherwise.

4.2.3. Different degrees of extrapolation

Our proposed framework deals with situations where predictions involve extrapolations beyond available field data. However, there are different degrees of extrapolation (Fig. 2). If we look at the situation where a three-dimensional groundwater model is calibrated against groundwater head and discharge data, model predictions of groundwater recharge to a given layer is a smaller extrapolation than model predictions of groundwater age or contaminant concentration. In both situations, model predictions are carried out for variables that have not been used as calibration targets and for which no traditional split-sample validation tests are possible. The type of validation test recommended for such situation is a proxy-basin test, which according to the principles in Klemes [26] and Refsgaard [38], for instance, could imply that validation tests have to be conducted in two similar catchments where relevant data (e.g. concentrations) exist, and where such data are not used for calibration. The residuals in the other catchments can then be seen as a measure of the uncertainty to be expected in the catchment of interest.

If model predictions are made for groundwater heads in cases involving groundwater abstraction, and the existing data available for calibration and validation tests do not include such abstraction, we also have an extrapolation case, although of a different nature. In this case we have data for the variable of predictive interest, but the catchment characteristics are non-stationary. This corresponds to the situation of model validation denoted by a differential split-sample test [26,38]. The differential split-sample test scheme recommended by Klemes also operates by tests on similar catchments where data for the type of non-stationary situation exist. Differential split-sample tests are often less demanding than proxy-basin tests [36]. A similar type of differential split-sample situation arises when predictions are required for a system in which structural change is expected (e.g. [50,4].

In cases where the conceptual models can be transferred to other catchments in a reliable and reproducible way, such proxy-basin and differential split-sample tests could be conducted and the results used to evaluate the goodness of the underlying conceptual models. It is worth noting that Klemes' test schemes, which also apply for cases of extrapolation, operate with tests for two alternative catchments. This has clear similarities with our strategy of recommending the use of multiple conceptual models.

4.3. Perspectives

In many cases where environmental models are used to make predictions that are extrapolations beyond the calibration base, no suitable framework exists for assessing the effects of model structure error. The proposed framework is composed of elements originating from different scientific disciplines. The elements are well tested individually, but not previously applied in such an integrated manner for water resources or environmental modelling applications. The full framework still needs to be tested in real-life cases.

Acknowledgement

For the three authors from GEUS and UVA the present work was supported by the Project 'Harmonised Techniques and Representative River Basin Data for Assessment and Use of Uncertainty Information in Integrated Water Management' (www.harmonirib.com), which is partly funded by the EC Energy, Environment and Sustainable Development programme (Contract EVK1-2002-00109). The constructive comments of

Hoshin V. Gupta and two anonymous reviewers are acknowledged.

Appendix. Terminology

The terminology used is mainly based on Refsgaard and Henriksen [39]:

Reality: The system that we aim to represent with the model, understood here as the study area.

Conceptual model: A representation 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 therefore includes a mathematical description (equations) of assumed processes and a description of the objects they interact with, including river system elements, ecological structures, geological features, etc. that are required for the particular purpose of modelling.

Model code: A generic mathematical description of a conceptual model, implemented in a computer program. It is generic in the sense that, without program changes, it can be used to establish a model with the same basic type of equations (but allowing different input variables and parameter values) for a different study area.

Model: A case-specific tailored version of a model code established for a particular study area and set of modelling objectives (output variables) including specific input data and parameter values.

Model confirmation: Determination of the adequacy of the conceptual model to provide an acceptable performance for the domain of intended application.

Code verification: Substantiation that a model code adequately represents a conceptual model within certain specified limits or ranges of application and corresponding ranges of accuracy.

Model calibration: The procedure of adjusting the parameter values of a model in such a way that the model reproduces an observed response of the system represented in the model 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. Note that various authors have criticised the use of the word validation for predictive models because universal validation of a model is in principle impossible and therefore prefer to use the term model evaluation [32,3]. In our definition [39] the term validation is not used in a universal sense, but is always restricted to clearly defined domains of applicability and performance accuracy ('numerical universal' in Popperian sense).

Pedigree: Pedigree conveys an evaluative account of the production process of information, and indicates different aspects of the underpinning and scientific status of the knowledge used. Pedigree is expressed by means of a set of pedigree criteria to assess these different aspects. Criteria for model parameter pedigree are for instance proxy representation, empirical basis, methodological rigor, theoretical understanding and validation. Assessment of pedigree involves qualitative expert judgement. To minimise arbitrariness and subjectivity in measuring strength, a pedigree matrix is used to code qualitative expert judgements for each criterion into a discrete numeral scale from 0 (weak) to 4 (strong) with linguistic descriptions (modes) of each level on the scale [49].

References

[1] Aller LT, Bennet T, Lehr JH, Petty RJ. DRASTIC: a standardized system for evaluating ground water pollution potential using hydrogeologic setting, US EPA Robert S. Kerr Environmental Research Laboratory, EPA/600/287/035, Ada, OK, 1987.

[2] Babendreier JE. National-scale multimedia risk assessment for hazardous waste disposal. In: International workshop on uncertainty, sensitivity and parameter estimation for multimedia environmental modelling held at US Nuclear Regulatory Commission, Rockville (MD), August 19–21, 2003. Proceedings, pp. 103–9.

[3] Beck MB. Model evaluation and performance. In: El-Shaarawi AH, Piegorsch WW, editors. Encyclopedia of environmetrics, vol. 3. Chichester: John Wiley & Sons, Ltd; 2002. p. 1275–9.

[4] Beck MB. Environmental foresight and structural change. Environ Modell Software 2005;20:651–70.

[5] Beck MB, van Straten G, editorsUncertainty and forecasting of water quality. Springer-Verlag; 1983.

[6] Beven K, Binley AM. The future of distributed models, model calibration and uncertainty predictions. Hydrol Process 1992;6:279–98.

[7] Beven K. Towards a coherent philosophy for modelling the environment. Proc Roy Soc London, A 2002;458(2026): 2465–84.

[8] Butts MB, Payne JT, Kristensen M, Madsen H. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow prediction. J Hydrol 2004;298:242–66.

[9] Carle SF, Fog GE. Transition probability based on indicator geostatistics. Math Geol 1996;28(4):453–77.

[10] Carle SF, Fog GE. Modeling spatial variability with one and multidimensional contineous-lag Markov chains. Math Geol 1997;29(7):891–917.

[11] Copenhagen County. Pilot project on establishment of methodology for zonation of groundwater vulnerability. In: Proceedings from seminar on groundwater zonation, November 7, 2000, County of Copenhagen [in Danish].

[12] Craye M, van der Sluijs JP, Funtowicz S. A reflexive approach to dealing with uncertainties in environmental health risk science and policy. Int J Risk Assess Manage 2005;5(2):216–36.

[13] Dubus IG, Brown CD, Beulke S. Sources of uncertainty in pesticide fate modelling. Sci Total Environ 2003;317:53–72.

[14] Dunn W. Using the method of context validation to mitigate type III errors in environmental policy analysis. In: Hisschemöller M,

Hoppe HV, Dunn W, Ravetz J, editors. Knowledge, power and participation in environmental policy. Policy studies review annual, vol. 12. New Jersey (USA): Transaction Publishers. p. 417–36.
[15] Efron B, Tibshirani RJ. An introduction to the bootstrap. Monographs on statistics and applied probability. New York: Chapman and Hall; 1993.
[16] Franchini M, Pacciani M. Comparative analysis of several conceptual rainfall-runoff models. J Hydrol 1992;122:161–219.
[17] Funtowicz SO, Ravetz JR. Uncertainty and quality in science for policy. Dordrecht: Kluwer; 1990. p. 229.
[18] Harrar WG, Sonnenborg TO, Henriksen HJ. Capture zone, travel time and solute transport predictions using inverse modelling and different geological models. Hydrogeol J 2003;11(5):536–48.
[19] Hodgson AM. Hexagons for systems thinking. Eur J Oper Res 1992;59:220–30.
[20] Hora SC. Acquisition of expert judgement: examples from risk assessment. J Energy Eng 1992;118:136–48.
[21] Højberg AL, Refsgaard JC. Model uncertainty – parameter uncertainty versus conceptual models. Water Sci Technol 2005;52(6):177–86.
[22] IPCC. Climate change 2001: the scientific basis. Contribution of working group I to the third assessment report of the intergovernmental panel of climate change [Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA, editors]. Cambridge University Press, Cambridge (UK) and New York (NY, USA). p. 881.
[23] Jakeman AJ, Letcher RA. Integrated assessment and modelling: features, principles and examples for catchment management. Environ Modell Software 2003;18:491–501.
[24] Jensen JB. Parameter and uncertainty estimation in groundwater modelling. PhD thesis, Department of Civil Engineering, Aalborg University, Series Paper no. 23, 2003.
[25] Keith DW. When is it appropriate to combine expert judgements? Clim Change 1996;33:139–43.
[26] Klemes V. Operational testing of hydrological simulation models. Hydrol Sci J 1986;31:13–24.
[27] Kloprogge P, van der Sluijs JP. The inclusion of stakeholder knowledge and perspectives in integrated assessment of climate change. Climatic Change, in press.
[28] Linkov I, Burmistrov D. Model uncertainty and choices made by modelers: lessons learned from the international atomic energy model intercomparisons. Risk Anal 2003;23(6):1297–308.
[29] Meyer PD, Ye M, Neuman SP, Cantrell KJ. Combined estimation of hydrogeologic conceptual model and parameter uncertainty. NUREG/CR-6843 Report, NRC, Washington, DC, 2004.
[30] National Research Council. Conceptual models of flow and transport in the vadose zone. Washington, DC: National Academy Press; 2001.
[31] Neuman SP, Wierenga PJ. A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites. University of Arizona, Report NUREG/CR-6805, 2003.
[32] Oreskes N, Shrader-Frechette K, Belitz K. Verification, validation, and confirmation of numerical models in the Earth Sciences. Science 1994;263:641–6.
[33] Pahl-Wostl C. Towards sustainability in the water sector – the importance of human actors and processes of social learning. Aquat Sci 2002;64:394–411.
[34] Poeter E, Anderson D. Multiple ranking and inference in ground water modeling. Ground Water 2005;43(4):597–605.
[35] Radwan M, Willems P, Berlamont J. Sensivity and uncertainty analysis for river quality modelling. J Hydroinform 2004:83–99.
[36] Refsgaard JC, Knudsen J. Operational validation and intercomparison of different types of hydrological models. Water Resources Res 1996;32(7):2189–202.
[37] Refsgaard JC, Hansen LK, Vahman M. Groundwater zonation in Copenhagen County – Intercomparision of thematic results from different consultants. In: Seminar on groundwater zonation, County of Copenhagen, November 7, 2000 [in Danish].
[38] Refsgaard JC. Towards a formal approach to calibration and validation of models using spatial data. In: Grayson R, Blöschl G, editors. Spatial patterns in catchment hydrology: observations and modelling. Cambridge University Press; 2001. p. 329–54.
[39] Refsgaard JC, Henriksen HJ. Modelling guidelines – terminology and guiding principles. Adv Water Resources 2004;27:71–82.
[40] Refsgaard JC, Storm B. MIKE SHE. In: Singh VP, editor. Computer models of watershed hydrology. Water Resources Publication; 1995. p. 809–46.
[41] Seibert J. On the need for benchmarks in hydrological modelling. Hydrol Process 2001;15(6):1063–4.
[42] Selroos JO, Walker DD, Strom A, Gylling B, Follin S. Comparison of alternative modelling approaches for groundwater flow in fractured rock. J Hydrol 2001;257:174–88.
[43] Troldborg L. The influence of conceptual geological models on the simulation of flow and transport in quaternary aquifer systems. PhD Thesis. Geological Survey of Denmark and Greenland, Report 2004/107.
[44] Usunoff E, Carrera J, Mousavi SF. An approach to the design of experiments for discriminating among alternative conceptual models. Adv Water Resources 1992;15:199–214.
[45] Van Griensven A, Meixner T. Dealing with unidentifiable sources of uncertainty within environmental models. In: Pahl C, Schmidt S, Jakeman T, editors. iEMSs 2004 international congress: ''Complexity and integrated resources management''. International Environmental Modelling and Software Society, Osnabrück, Germany, June 2004.
[46] Van der Sluijs JP. Anchoring amid uncertainty; On the management of uncertainties in risk assessment of anthropogenic climate change, Ph.D. thesis, Utrecht University, 1997. p. 260.
[47] Van der Sluijs JP, Potting J, Risbey JS, Van Vuuren D, de Vries B, Beusen A, et al. Uncertainty assessment of the IMAGE/TIMER B1 CO2 emissions scenario, using the NUSAP method. Report commissioned by the Netherlands National Research Program on global Air Pollution and Climate Change, RIVM, Bilthoven, The Netherlands, 2002. p. 225.
[48] Van der Sluijs JP, Risbey JS, Kloprogge P, Ravetz JR, Funtowicz SO, Corral Quintana S, et al. RIVM/MNP Guidance for uncertainty assessment and communication: detailed guidance, report commissioned by RIVM/MNP – Copernicus Institute, Department of Science, Technology and Society, Utrecht University, Utrecht, The Netherlands, 2003. p. 71.
[49] Van der Sluijs JP, Craye M, Funtowicz SO, Kloprogge P, Ravetz J, Risbey JS. Combining quantitative and qualitative measures of uncertainty in model based foresight studies: the NUSAP system. Risk Anal 2005;25(2):481–92.
[50] Van Straten G, Keesman KJ. Uncertainty propagation and speculation in projective forecasts of environmental change: a lake-eutrophication example. J Forecast 1991;10:163–90.
[51] Vennix JAM. Group model-building: tackling messy problems. Syst Dyn Rev 1999;15(4).
[52] Visser H, Folkert RJM, Hoekstra J, De Wolff JJ. Identifying key sources of uncertainty in climate change projections. Clim Change 2000;45:421–57.
[53] Vrugt JA, Diks CGH, Gupta HV. Improved treatment of uncertainty in hydrologic modelling: combining the strengths of global optimization and data assimilation. Water Resources Res 2005;41(1). Art No W01017.
[54] Walker WE, Harremoës P, Rotmans J, Van der Sluijs JP, Van Asselt MBA, Janssen P, et al. Defining uncertainty. A conceptual basis for uncertainty management in model-based decision support. Integr Assessment 2003;4(1):5–17.