# Identification and Evaluation of Watershed Models

Thorsten Wagener[1] and Howard S. Wheater

*Department of Civil and Environmental Engineering, Imperial College of Science, Technology and Medicine, London, United Kingdom*

Hoshin V. Gupta

*SAHRA, NSF STC for Sustainability of semi-Arid Hydrology and Riparian Areas Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona*

Conceptual modeling requires the identification of a suitable model structure and, within a chosen structure, the estimation of parameter values (and, ideally, their uncertainty) through calibration against observed data. A lack of objective approaches to evaluate model structures and the inability of calibration procedures to distinguish between the suitability of different parameter sets are major sources of uncertainty in current modeling procedures. This is further complicated by the increasing awareness of model structural inadequacies. A framework for the identification and evaluation of conceptual rainfall-runoff models is presented, based on multi-objective performance and identifiability approaches, and a novel dynamic identifiability analysis (DYNIA) method which results in an improved use of available information. The multi-objective approach is mainly used to analyze the performance and identifiability of competing models and model structures, while the DYNIA allows periods of high information content for specific parameters to be identified and model structures to be evaluated with respect to failure of individual components. The framework is applied to a watershed located in the South of England.

## 1. INTRODUCTION

Many if not most rainfall-runoff model structures currently used can be classified as conceptual. This classification is based on two criteria: (1) the structure of these models is specified prior to any modelling being undertaken, and (2) (at least some of) the model parameters do not have a direct physical interpretation, in the sense of being independently measurable, and have to be estimated through calibration against observed data [*Wheater et al*., 1993]. Calibration is a process of parameter adjustment (automatic or manual), until observed and calculated output time-series show a sufficiently high degree of similarity.

Conceptual rainfall-runoff (CRR) model structures commonly aggregate, in space and time, the hydrological processes occurring in a watershed (also called catchment), into a number of key responses represented by storage components (state variables) and their interactions (fluxes). The model parameters describe aspects such as the size of those storage components, the location of outlets, the distribution of storage volumes etc. Conceptual parameters, therefore, usually refer to a collection of aggregated processes and they may cover a large number of sub-processes that cannot be represented separately or explicitly [*Van Straten and Keesman*, 1991]. The underlying assumption however is that these parameters are, even if not measurable properties,

[1] Now at SAHRA, NSF STC for Sustainability of semi-Arid Hydrology and Riparian Areas, Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona

Calibration of Watershed Models
Water Science and Application Volume 6

10/1029/006WS03

at least constants and representative of inherent properties of the natural system [*Bard*, 1974, p.11].

The modeller's task is the identification of an appropriate CRR model (or models) for a specific case, i.e. a given modelling objective, watershed characteristics and data set. A model is defined in this context as a specific parameter set within a selected model structure. Experience shows that this identification is a difficult task. Various parameter sets, often widely distributed within the feasible parameter space [e.g. *Duan et al*., 1992; *Freer et al*., 1996], and sometimes even different conceptualisations of the watershed system [e.g. *Piñol et al*., 1997; *Uhlenbrock et al*., 1999], may yield equally good results in terms of a predefined objective function. This ambiguity has serious impacts on parameter and predictive uncertainty [e.g. *Beven and Binley*, 1992], and therefore limits the applicability of CRR models, e.g. for the simulation of land-use or climate-change scenarios, or for regionalisation studies [*Wheater et al*., 1993].

Initially it was thought that this problem would disappear with improved automatic search algorithms, capable of locating the global optimum on the response surface [e.g. *Duan et al*., 1992]. However, even though powerful global optimisation algorithms are available today, single-objective calibration procedures still fail to completely replace manual calibration. One reason for this is that the resulting hydrographs are often perceived to be inferior to those produced through manual calibration from the hydrologist's point of view [e.g. *Gupta et al*., 1998; *Boyle et al*., 2000]. It has been suggested that this is due to the fundamental problem that single-objective automatic calibration is not sophisticated enough to replicate the several performance criteria implicitly or explicitly used by the hydrologist in manual calibration. This problem is increased by indications that, due to structural inadequacies, one parameter set might not be enough to adequately describe all response modes of a hydrological system. Therefore, there is a strong argument that the process of identification of dynamic, conceptual models has to be rethought [*Gupta et al*., 1998; *Gupta*, 2000].

Three reactions to this problem of ambiguity of system description can be found in the hydrological literature. The first is the increased use of parsimonious model structures [e.g. *Jakeman and Hornberger*, 1993; *Young et al*., 1996; *Wagener et al*., 2001b], i.e. structures only containing those parameters, and therefore model components, that can be identified from the observed system output. However, the increase in identifiability is bought at the price of a decrease in the number of processes described separately by the model. There is therefore a danger of building a model (structure) which is too simplistic for the anticipated purpose. Such a model (structure) can be unreliable outside the range of watershed conditions, i.e. climate and land-use, on which it was calibrated, due to the restriction to 'justifiable' components [*Kuczera and Mroczkowski*, 1998]. It is also particularly important that the data used has a high information content in order to ensure that the main response modes are excited during calibration [*Gupta and Sorooshian*, 1985, *Yapo et al*., 1996].

The second reaction is the search for calibration methods which make better use of the information contained in the available data time-series, e.g. streamflow and/or groundwater levels. Various research efforts have shown that the amount of information retrieved using a single objective function is sufficient to identify only between three and five parameters [e.g. *Beven*, 1989; *Jakeman and Hornberger*, 1993; *Gupta*, 2000]. Most CRR model structures contain a larger number. More information can become available through the definition of multiple objective functions to increase the discriminative power of the calibration procedure [e.g. *Gupta et al*., 1998; *Gupta*, 2000]. These measures can either retrieve different types of information from a single time-series, e.g. streamflow [e.g. *Wheater et al*., 1986; *Gupta et al*., 1998; *Dunne*, 1999; *Boyle et al*., 2000; *Wagener et al*., 2001a], or describe the performance of individual models with respect to different measured variables, e.g. groundwater levels [e.g. *Kuczera and Mroczkowski*, 1998; *Seibert*, 2000], saturated areas [*Franks et al*., 1998], or measurements of streamflow salinity [*Mroczkowski et al*., 1997; *Kuczera and Mroczkowski*, 1998]. However, the usefulness of additional data can depend on the adequacy of the model structure investigated. *Lamb et al.* [1998] found that the use of groundwater levels from one or only a few measurement points as additional output variable(s) helped to reduce the parameter uncertainty of Topmodel [*Beven et al*., 1995]. The use of many (>100) groundwater measurement points however, leads to an increase in prediction uncertainty indicating structural problems in the model. *Seibert and McDonnell* [this volume] show in a different approach how the parameter space can be constrained when soft data, i.e. qualitative knowledge of the watershed behaviour, is included in the calibration process. The soft data in their case included information, derived through experimental work, about the contribution of new water to runoff and the restriction of parameter ranges to a desirable range. The result is a more realistic model, which will however yield sub-optimal performances with respect to many specific objective functions, in their case the Nash-Sutcliffe efficiency measure [*Nash and Sutcliffe*, 1970]. *Chappell et al*. [1998] give another example of how expert knowledge of internal catchment dynamics (e.g. saturated areas) can be used to constrain the parameter space.

Thirdly, some researchers abandoned the idea of a uniquely identifiable model in favour of the identification of

a model population [e.g. *van Straten and Keesman*, 1991; *Beven and Binley*, 1992; *Gupta et al*., 1998]. This can for example be a population of models with varying degrees of (some sort of) likelihood to be representative of the watershed at hand, the idea behind the Generalized Likelihood Uncertainty Estimation (GLUE) approach [*Freer et al*., this volume]. Or an approach based on the recognition that the calibration of a rainfall-runoff model is inherently a multi-objective problem, resulting in a population of non-dominated parameter sets [*Goldberg*, 1989, p.201] in the presence of model structural inadequacies [*Gupta et al*., 1998].

Here, we seek to increase the amount of information made available from an output time-series and to guide the identification of parsimonious model structures, consistent with a given model application as explained below. We use multi-objective approaches to performance and identifiability analysis and a novel dynamic identifiability analysis (DYNIA) method for assumption testing. These can be integrated into a framework for model identification and evaluation. An application example at the end of this chapter shows the use of the framework for a specific case.

## 2. IDENTIFICATION OF CONCEPTUAL RAINFALL-RUNOFF MODELS

The purpose of identifiability analysis in CRR modelling is to find (the) model structure(s) and corresponding parameter set(s) which are representative of the watershed under investigation, while considering aspects such as modelling objectives and available data. This identifiability analysis can be split into two stages: model structure selection and parameter estimation, which can, however, not be treated as completely separate [*Sorooshian and Gupta*, 1985] (in order to evaluate model structures fully, one has to analyse their performance and behaviour which requires some form of parameter estimation).

Traditional modelling procedures commonly contain, amongst others, an additional third step [e.g *Anderson and Burt*, 1985]. This is a validation or verification step often used to show that the selected model really is the correct representation of the watershed under investigation. This results in the following three steps as part of a longer procedure:

(1) Selection or development of a model structure, and subsequently computer code, to represent the conceptualisation of the hydrologic system which the hydrologist has established in his or her mind for the watershed under study.
(2) Calibration of the selected model structure, i.e. estimation of the 'best' parameter set(s) with respect to one or more (often combined) criteria.
(3) Validation or verification of this model by (successfully) applying it to a data set not used in the calibration stage.

It is important to stress that the original meanings of the words validation and verification are different. Verification is the stronger statement, meaning to establish the truth, while validation means to establish legitimacy [*Oreskes et al*., 1994]. In the context of hydrological modelling, these terms are often used synonymously, describing a step to justify that the chosen model is an acceptable representation of the real system. An in-depth discussion on this topic can be found in *Oreskes et al*. [1994].

These three steps are similar to the logic of induction often used in science. This idea of induction is founded on the underlying assumption that a general statement can be inferred from the results of observations or experiments [*Popper*, 2000, p.27]. It includes the assumption that a hypothesis, e.g. a chosen model structure, can be shown to be correct, i.e. a hypothesis can be validated or verified, through supporting evidence. The steps taken in this traditional scientific method are [for example modified from *Magee*, 1977, p. 56]:

(1) Observation and experiment;
(2) inductive generalization, i.e. a new hypothesis;
(3) attempted verification of hypothesis, i.e. proof or disproof of hypothesis;
(4) knowledge.

However, the logical error in this approach is, (as *Magee* [1977, p. 20] derives from statements by the philosopher Hume), that no number of singular observation statements, however large, could logically entail an unrestrictedly general statement. In rainfall-runoff modelling this is equivalent to the statement that, however often a model is capable of reproducing the response of a particular watershed, it can never be concluded that the true model has been found. It could for example be that future measurements will capture more extreme events, exciting a response not captured by earlier data and therefore not included in the model. Similarly, Popper concluded that no theory or hypothesis could ever be taken as the final truth. It can only be said that it is corroborated by every observation so far, and yields better predictions than any known alternative. It will however, always remain replaceable by a better theory or turn out to be false at a later stage [*Popper*, 2000, p.33].

The idea that a model can be verified (verus, meaning true in Latin [*Oreskes et al*., 1994]) is therefore ill-founded and alternative modelling frameworks have to be found. One such alternative approach was suggested by *Popper* [2000].

He realised that, while no number of correctly predicted observations can lead to the conclusion that a hypothesis is correct, a single unexplained observation can lead to the falsification of the hypothesis. Hence he replaced the framework of verification with a framework of falsification, allowing the testing of a hypothesis.

This framework of falsification as suggested by Popper can be outlined as follows [modified from *Magee*, 1977, p.56]:

(1) The initial problem or question, often resulting from the fact that an existing hypothesis has failed;
(2) one (or more) proposed new hypothesis(es);
(3) deduction of testable propositions from the new hypothesis;
(4) attempted falsification of the new hypothesis by testing the propositions;
(5) preference established between competing hypotheses.

The procedure is repeated as soon as the new hypothesis fails. It is thus possible to search for the truth, but it is not possible to know when the truth has been found, a problem which has to be reflected in any scientific method.

Additionally, *Beven* [2000, p.304] pointed out that it is very likely, at least with the current generation of CRR models, that every model will fail to reproduce some of the behaviour of a watershed at some stage. However, even if one knows that the model is inadequate, one often has to use it due to the lack of alternatives. And for many cases, the use of this inadequate model will be sufficient for the selected purpose. Or as Wilfried Trotter put it more generally: In science the primary duty of ideas is to be useful and interesting even more than to be 'true' [*Beveridge*, 1957, p. 41].

How this general idea of hypothesis falsification can be put into a framework for CRR modelling is described below.

### *2.1. Identification of Model Structures*

A large number of CRR modelling structures is currently available. These differ, for example, in the degree of detail described, the manner in which processes are conceptualised, requirements for input and output data, and possible spatial and temporal resolution. Despite these differences, a number of model structures may appear equally possible for a specific study, and the selection process usually amounts to a subjective decision by the modeller, since objective decision criteria are often lacking [*Mroczkowski et al*., 1997]. It is therefore important to deduce testable propositions with respect to the assumptions underlying the model structure, i.e. about the hypothesis of how the watershed works, and to find measures of evaluation that give some objective guidance as to whether a selected structure is suitable or not. However, *Uhlenbrock et al*. [1999] have shown that it is difficult to achieve this using single-objective Monte-Carlo-based calibration approaches. They were able to derive good performances with respect to the prediction of streamflow, from sensible, as well as incorrect conceptualisations of a watershed. *Mroczkowski et al*. [1997] encountered similar problems when trying to falsify one of two possible model structures, including and excluding a groundwater discharge zone respectively, to represent two paired watersheds in Western Australia. This was impossible for both watersheds when only streamflow data was used. The additional use of stream chloride and groundwater level measurements allowed at least for the falsification of one of the model structures in case of the second watershed which had undergone considerable land-use changes.

Testable propositions about a specific model structure can be either related to the performance of the model or its components, or they can be related to its proper functioning.

A test of performance is the assessment whether or not the model structure is capable of sufficiently reproducing the observed behaviour of the natural system, considering the given quality of data. However, an overall measure of performance, aggregating the residuals over the calibration period, and therefore usually a number of response modes, hides information about how well different model components perform. It can be shown that the use of multiple-objectives for single-output models, measuring the model's performance during different response modes, can give more detailed information and allows the modeller to link model performance to individual model components [e.g. *Boyle et al*., 2001; *Wagener et al*., 2001a]. Additional information will also be available in cases where the model produces other measurable output variables, e.g. groundwater levels or hydro-chemical variables, as mentioned earlier.

Evaluation of the proper functioning of the model means questioning the assumptions underlying the model's structure, such as: Do the model components really represent the response modes they are intended to represent? And is the model structure capable of reproducing the different dominant modes of behaviour of the watershed with a single parameter set? A model structure is usually a combination of different hypotheses of the working of the natural system. If those hypotheses are to be individually testable, they should be related to individual model components and not just to the model structure as a whole [*Beck*, 1987; *Beck et al*., 1993].

One, already mentioned, underlying assumption of conceptual modelling is the consideration of model parameters as constant in time, at least as long as for example no changes in the watershed occur that would alter the hydrological

response, such as land-use changes. Different researchers [e.g. *Beck*, 1985; 1987; *Gupta et al*., 1998; *Boyle et al*., 2000; *Wagener et al*., 2001a] have shown that this assumption can be tested, and that the failure of a model structure to simulate different response modes with a single parameter set suggests inadequacies in the functioning of the model.

Beck used the Extended Kalman Filter (EKF) extensively to recursively estimate model parameters and to utilize the occurrence of parameter deviation as an indicator of model structural failure [e.g. *Beck*, 1985; 1987; *Stigter et al*., 1997]. For example, in the identification of a model of organic waste degradation in a river, changes in optimum parameter values in time from one location in the parameter space to another were identified [*Beck*, 1985]. Beck concluded from this observation that the model hypothesis had failed, i.e. the parameters were changing to compensate for one or more missing aspect(s) in the model structure. The subsequent step is to draw inference from the type of failure to develop an improved hypothesis of the model structure. However, there are limitations to the EKF approach. Beck concluded with respect to the use of the EKF for hypothesis testing that the performance of the EKF is not as robust as would be desirable and, inter alia, is heavily compromised by the need to make more or less arbitrary assumptions about the sources of uncertainty affecting the identification problem [*Beck*, 1987].

A trade-off in the capability to simulate different response modes can occur, as shown by *Boyle et al.* [2000] for the example for a popular complex rainfall-runoff model (Sacramento with 13 calibrated parameters [*Smith et al*., this volume]), thus it was not possible to reproduce (slow) recession periods and the remaining system response modes simultaneously. Their multi-objective analysis suggests that the cause for this problem is mainly an inadequate representation of the upper soil zone processes.

There are therefore ideas to address the problem of model structure identification in a more objective way. However, they are not without weaknesses, as the Beck statement about the use of EKF showed earlier in the text. These need to be addressed to derive more suitable approaches.

### *2.2. Identification of Parameters*

The second stage in the model identification process is the estimation of a suitable parameter set, usually referred to as calibration of the model structure. In this process, the parameters of a model structure are adjusted until the observed system output and the model output show acceptable levels of agreement. Manual calibration does this in a trial-and-error procedure, often using a number of different measures of performance and visual inspection of the hydrograph [e.g. *Gupta et al*., 1998; *Smith et al*., this volume]. It can yield good results and is often a good way to learn about the model, but it can be time consuming, requires extensive experience with a specific model structure and an objective analysis of parameter uncertainty is not possible. Traditional single-objective automatic calibration on the other hand is fast and objective, but will produce results which reflect the choice of objective function and may therefore not be acceptable to hydrologists concerned with a number of aspects of performance [*Boyle et al*., 2000]. In particular the aggregation of the model residuals into an objective function leads to the neglect and loss of information about individual response modes, and can result in a biased performance, fitting a specific aspect of the hydrograph at the expense of another. It also leads to problems with the identification of those parameters associated with response modes which do not significantly influence the selected objective function [*Wagener et al*., 2001a]. Selecting, for example, an objective function which puts more emphasis on fitting peak flows, e.g. the Nash-Sutcliffe efficiency value [*Nash and Sutcliffe*, 1970], due to its use of squared residual values [*Legates and McCabe*, 1999], will often not allow for the identification of parameters related to the slow response of a watershed [e.g. *Dunne*, 1999].

An example to demonstrate this problem is briefly presented. It uses a simple model structure consisting of a Penman two-layer soil moisture accounting component [*Penman*, 1949] to produce effective rainfall and a linear routing component using two conceptual reservoirs in parallel to transform it into streamflow. A comparison of hydrographs produced by different parameter sets within the selected structure, which yield similar objective function values, shows that these hydrographs can be visually different. Figure 1 shows a hundred days extract of six years of daily streamflow data, where the observed time-series (black line) is plotted with seven different realisations (grey lines), i.e. using the same model structure, but different parameter sets. The objective function used during calibration is the well known Root Mean Squared Error (RMSE). Each of the models presented yields a RMSE of 0.60mm/d when the complete calibration period (6 years) is considered. However, the hydrographs produced are clearly visually different. The added dotty plots of the two residence times of the (linear) routing component show that while the quick flow residence time, k(quick) is very well identified, the slow flow residence time, k(slow), is not. This is consistent with the observation that the main difference between the hydrographs can be observed during low flow periods. This effect is due to the use of squared residuals when calculating the RMSE.

This result demonstrates that traditional single-objective optimisation methods do not have the ability to distinguish

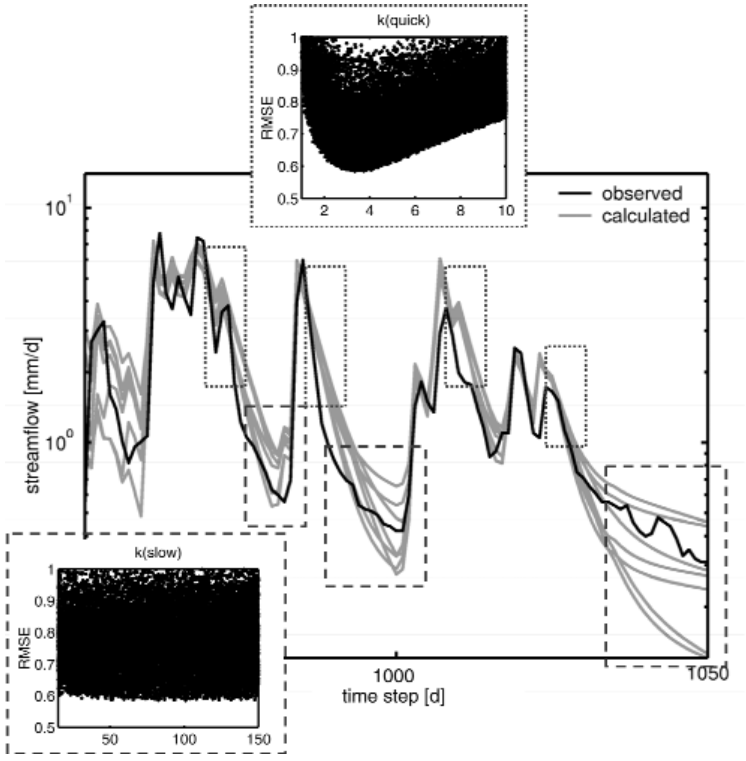


**Figure 1.** Hundred days extract of six years of daily streamflow data. Observed flow in black, seven different model realizations in gray. Inlets show dotty plots for the time constants k(quick) and k(slow) with respect to the Root Mean Squared Error (RMSE). The model structure used consists of a Penman soil moisture accounting and a parallel routing component of linear reservoirs with fixed flow distribution (see application example for details).

between visually different behaviour [*Gupta*, 2000]. The requirement for a parameter set to be uniquely locatable within the parameter space, i.e. to be globally identifiable, is that it yields a unique response vector [*Kleissen et al*., 1990; *Mous*, 1993]. The unique response vector, in this case a unique (calculated) hydrograph, might be achievable, but this uniqueness is often lost if the residuals are aggregated into a single objective function. Such problems cannot be solved through improved search algorithms. They are rather inherent in the philosophy of the calibration procedure itself.

Clearly, the complex thought processes which lead to successful manual calibration are very difficult to encapsulate in a single objective function. This is illustrated by the requirements defined by the US National Weather Service (NWS) for the manual calibration of the Sacramento model structure [*NWS*, 2001]:

(1) Proper calibration of a conceptual model should result in parameters that cause model components to mimic processes they are designed to represent. This requires the ability to isolate the effects of each parameter.
(2) Each parameter is designed to represent a specific portion of the hydrograph under certain moisture conditions.
(3) Calibration should concentrate on having each parameter serve its primary function rather than overall goodness of fit.

It can be seen from these requirements that manual calibration is more complex than the optimisation of a single objective function, and that traditional automatic calibration procedures will in general not achieve comparable results. It is for example often not possible to isolate the effects of individual parameters and treat them as independent entities as done in the manual approach described above. Another aspect is that the goal of single-objective optimisation is purely to optimise the model's performance with respect to a selected overall goodness of fit measure which is very different from requirement three. This is not to say that traditional 'single' objective functions are not important parts of any model evaluation. The point is rather that they are not sufficient and should be complemented by a variety of measures.

*Gupta et al*. [1998] review this problem in more detail and conclude that a multi-objective approach to automatic calibration can be successful. *Boyle et al*.[2000] show how such

a procedure can be applied to combine the requirements of manual calibration with the advantages of automatic calibration. A multi-objective algorithm is used to find the model population necessary to fit all aspects of the hydrograph. The user can then, if necessary, manually select a parameter set from this population to fit the hydrograph in the desired way. This will however, in the presence of model structural inadequacies, lead to a sub-optimal performance with respect to at least some of the other measures [*Boyle et al*., 2000; *Seibert and McDonnell*, this volume]. The resulting trade-off of the ability of different parameter sets to fit different aspects of the hydrograph usually leads to a compromise solution [*Ehrgott*, 2000] in cases where a single parameter set has to be specified. The procedure of *Boyle et al*. [2000] for example, analyses the local behaviour of the model additionally to its global behaviour [*Gupta*, 2000]. The global behaviour is described through objective functions such as overall bias or some measure of the overall variance, e.g. the Root Mean Squared Error (RMSE). The local behaviour is defined by aspects like the timing of the peaks, or the performance during quick and slow response periods [*Boyle et al*., 2000; 2001].

Recent research into parameter identification has thus moved away from simply trying to improve search algorithms, but has taken a closer look at the assumptions underlying (automatic) calibration approaches [e.g. *Gupta et al*., 1998]. This has lead to the use of multi-objective (MO) automatic approaches which so far have given promising results [*Boyle et al*., 2000; *Wagener et al*., 2001a]. Further investigations are required to make MO optimization a standard method for parameter estimation. For example questions such as the appropriate number and derivation of OFs within a MO approach must be resolved, and will probably depend on model structure and watershed characteristics [*Gupta*, 2000].

## 3. EVALUATION OF CONCEPTUAL RAINFALL-RUNOFF MODELS

It was established earlier that the idea of calibration and validation of CRR models is in principle ill-founded, i.e. to establish a model as the true representation of a hydrological system. The model identification problem is therefore seen here as a process of model evaluation. Within this process, models and model structures are evaluated with respect to different criteria and those that fail, in whatever way, are rejected as possible representations of the watershed under investigation. This will usually result in a population of feasible models or even model structures which can then be used for a (combined) prediction, which will result in a prediction range, rather than a single value for each time-step.

This evaluation should be at least with respect to three dimensions:

(1) Performance, with respect to reproducing the behaviour of the system.
(2) Uncertainty in the parameters, which is assumed to be inversely related to their identifiability.
(3) Assumptions, i.e. are any assumptions made during the development of the model (structure) violated.

The smaller the population of models (or even model structures) that survives this evaluation, i.e. those that are corroborated by it, the more identifiable is the representation of the natural system in mathematical form. Approaches to test models with respect to these three criteria are described below.

### *3.1. Evaluation of Competing Model Structures—Multi-objective Performance and Identifiability Analysis*

Multi-objective (MO) approaches can be applied to establish preferences between competing model structures or even model components, i.e. competing hypotheses, with respect to their performance and their identifiability. A MO approach is advantageous because the use of multiple objective criteria for parameter estimation permits more of the information contained in the data set to be used and distributes the importance of the parameter estimates among more components of the model. Additionally, the precision of some parameters may be greatly improved without an adverse impact on other parameters [*Yan and Haan*, 1991]. More detailed descriptions of MO model analysis can be found in the chapters by Gupta et al. and Boyle et al. [this volume].

*3.1.1. Measures of performance and identifiability.* It was already established earlier in the text that it is advantageous to evaluate the global and the local behaviour of models to increase the amount of information retrieved from the residuals in the context of single output rainfall-runoff models. Global behaviour is measured by traditional OFs, e.g. the RMSE or the bias for the whole calibration period, while different OFs have to be defined to measure the local behaviour. One way of implementing local measures is by partitioning the continuous output time series into different response periods. A separate OF can then be specified for each period, thus reducing the amount of information lost through aggregation of the residuals, e.g. by mixing high flow and recession periods.

Partitioning schemes proposed for hydrological time series include those based on: (a) Experience with a specific model structure (e.g. the Birkenes model structure in the

case of *Wheater et al*., 1986), i.e. different periods of the streamflow time series are selected based on the modeller's judgement. The intention of *Wheater et al*. [1986] was to improve the identifiability of insensitive parameters, so called minor parameters, with respect to an overall measure. Individual parameters, or pairs of parameters, are estimated using a simple grid search to find the best values for the individual objective functions. This is done in an iterative and sequential fashion, starting with the minor parameters and finishing with the dominant ones. (b) Hydrological understanding, i.e. the separation of different watershed response modes through a segmentation procedure based on the hydrologist's perception of the hydrological system (e.g. *Harlin*, 1991; *Dunne*, 1999; *Boyle et al*., 2000; *Wagener et al*., 2001a). For example, *Boyle et al*. [2000] propose hydrograph segmentation into periods 'driven' by rainfall, and periods of drainage. The drainage period is further subdivided into quick and slow drainage by a simple threshold value. (c) Parameter sensitivity [e.g. *Kleissen*, 1990; *Wagner and Harvey*, 1997; *Harvey and Wagner*, 2000], where it is assumed that informative periods are those time-steps during which the model outputs show a high sensitivity to changes in the model parameters [*Wagner and Harvey*, 1997]. *Kleissen* [1990] for example developed an optimisation procedure whereby only data segments during which the parameter shows a high degree of first order sensitivity are included in the calibration of that parameter (group) utilising a local optimisation algorithm. (d) Similar characteristics in the data derived from techniques like cluster analysis [e.g. *Boogard et al*., 1998] or wavelet analysis [*Gupta*, 2000] can be used to group data points or periods based on their information content. The different clusters could then be used to define separate objective functions.

While these methods help to retrieve more information, they also show some weaknesses. Approaches (a) and (b) are subjective and based on the hydrologist's experience, and so are not easily applicable to a wide variety of models and watersheds. Approach (c), while objective, does not recognise the effects of parameter dependencies, and may not highlight periods which are most informative about the parameters as independent entities, i.e. periods where the dependency with respect to other parameters is low. The sensitivity of the model performance to changes in the parameter is a necessary requirement, but it is not sufficient for the identifiability of the parameter. Furthermore, if the parameter sensitivity is measured locally [e.g. *Kleissen*, 1990], the result is not guaranteed over the feasible parameter space. However, *Wagner and Harvey* [1997] show that this problem can be reduced by implementing a Monte Carlo procedure where the sensitivity for a large number of different parameter combinations is assessed using parameter covariance matrices. Approach (d) is independent of any model structure and links between the results and the model parameters still need to be established.

There is therefore scope to improve the objectivity, applicability and robustness of approaches to hydrograph disaggregation, with the goal of improving model structure and parameter identifiability.

The evaluation of the model performance should, if possible, also include objective functions tailored to fit the specific purpose of the model. An example is the use of the model to investigate available quantities for abstraction purposes. Assuming that abstraction can only take place during periods when the water level is above a minimum environmentally acceptable flow and below a maximum water supply abstraction rate allows the definition of a specific objective function. This measure would only aggregate the residuals of the selected period and can give important information about how a model performs with respect to the anticipated task. However, it is important to mention that this should never be the only evaluation criterion.

However, how can one estimate the identifiability of the individual parameters with respect to the different OFs defined? A simple measure of parameter identifiability is defined by *Wagener et al*. [2001a]. It is based on the parameter population conditioned by the selected measure of performance (Figure 2). A uniform random sampling procedure is performed, and the resulting OF values are transformed so that the best performing parameter set is assigned the highest value and all measures sum to unity (these are termed support values in Figure 2). The best performing 10% of all parameter sets are selected and the cumulative marginal distributions for each parameter are plotted. A uniform distribution would plot as a straight line, while a population showing a clear peak will show a curved line. The stronger the conditioning, the stronger the curvature will be. The range of each parameter is subsequently split into M containers and the gradient of the cumulative distribution in each container is calculated. The highest gradient will occur where the conditioning of the distribution is strongest, i.e. at the location of a peak. The amplitude of the gradient is also indicated by the grey shading of the bar, with a darker colour indicating a higher gradient. Other measures of identifiability are possible [e.g. *Wagener et al*., 1999], but this one has been shown to be robust and easy to calculate.

*3.1.2. Multi-objective framework*. The above described multi-objective performance and identifiability approaches can be put into an analytical framework to estimate the appropriate level of model complexity for a specific case [Figure 3, adapted from *Wagener et al*., 2001a].

The hydrologist's perception of a given hydrological system strongly influences the level of conceptualisation that must be translated into the model structure. The importance of different system response modes, i.e. key processes that need to be simulated by the model, however, depends on the intended modelling purpose. Therefore, the level of model structural complexity required must be determined through careful consideration of the key processes included in the model structure and the level of prediction accuracy necessary for the intended modelling purpose.

On the other hand there is the level of structural complexity actually supported by the information contained within the observed data. It is defined here simply as the number of parameters, and therefore separate model components and processes, that can be identified. Other aspects of complexity [e.g. *Kleissen et al*., 1990] like the number of model states or interactions between the state variables, or the use of non-linear components instead of linear ones, are not considered here.

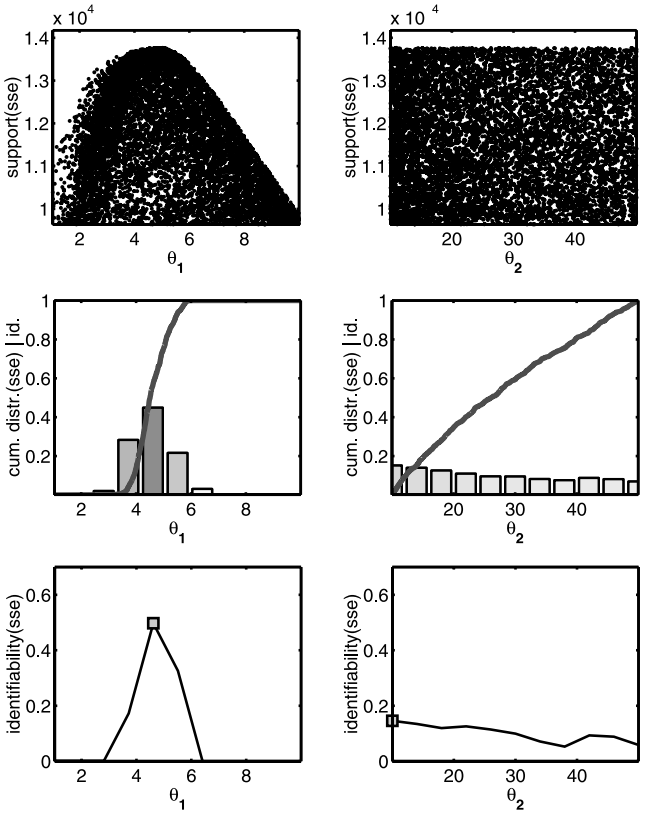


**Figure 2.** A measure of identifiability can be defined as follows: an initially uniform distribution is conditioned on some OF, the resulting dotty plot is shown in the top plots, selecting the top percentile (*e.g.* 10%) and plotting the cumulative distribution of the transformed measures leads to the middle plots, the gradient distribution of the cumulative distribution is a measure of identifiability, see bottom. The plots in the left column show an identifiable parameter, while the plots in the right column show a non-identifiable one.

An increase in complexity will often increase the performance. However, it will also often increase the uncertainty, for example due to reduction in parameter identifiability caused by increased parameter interaction. What trade-off between performance and identifiability is acceptable depends on the modelling purpose and the hydrologist's preference. In a regionalisation study, a more identifiable model with reduced performance might be adequate, while parameter identifiability might be of minor importance for extension of a single-site record.

It was already established earlier in the text that such a framework has to use a multi-objective approach to allow for an objective analysis. Using various objective functions to represent different system response modes is especially suitable for comparison studies since it allows us to attribute the model performance during different system response modes to different model components, for example either the moisture accounting or the routing components [*Wagener et al*., 2001a]. Using the segmentation approach by *Boyle et al*. [2000] as described earlier in the text, it is possible to establish that a certain model structure might perform better during "driven" periods because of a superior moisture accounting component, while another model structure containing a more appropriate slow flow routing component could result in higher performance during "non-driven slow" periods. A single-objective framework does not allow the comparison of model components and consequently important information relevant to identifying the most suitable model structure is lost. *Boyle et al*. [2001] use

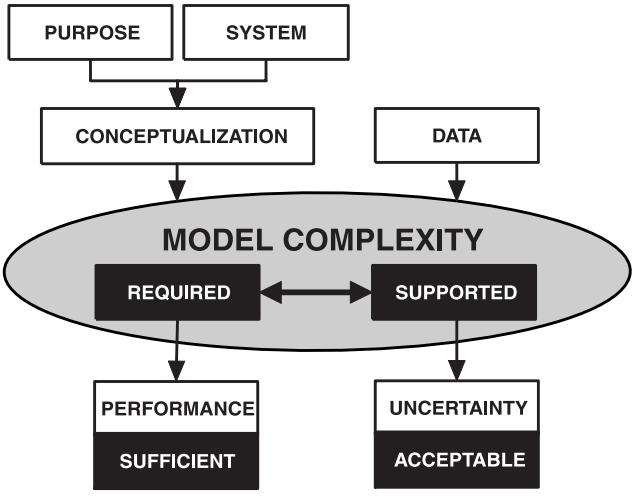


**Figure 3.** Framework for the evaluation of competing rainfall-runoff model structures.

this to evaluate the benefit of "spatial distribution" of model input (precipitation), structural components (soil moisture and streamflow routing computations) and surface characteristics (parameters) with respect to the reproduction of different response modes of the watershed system.

This framework will also necessarily be comparative, i.e. different models and usually different model structures will have to be compared to identify a suitable model or models. The reason is that the level of performance that can be reached is unknown, due to unknown influences of data error and of natural randomness. Those models and model structures that severely under-perform can be refuted and excluded from further consideration. In cases where all models fail, one has to go back and relax the criteria for under performance [*Beven*, 2000, p. 304].

Model structures producing more than a single output variable, e.g. groundwater levels or water quality parameters, can be tested with respect to all of those variables if measurements are available. One could say that the informative (or empirical) content of these structures is higher and they have therefore a higher degree of testability or falsifiability [*Popper*, 2000, p.113]. However, a hypothesis, or a model structure in our case, which has a higher informative content, is also logically less probable, because the more information a hypothesis contains, the more options there are for it to be false [*Popper*, 2000, p.119; *Magee*, 1977, p. 36]. Multi-output models are beyond the scope of this chapter though.

### *3.2. Evaluation of Individual Model Structures—Dynamic Identifiability Analysis*

There is an apparent lack of objective procedures to evaluate the suitability of a specific conceptual model structure to represent a specific hydrological system. It has been shown earlier how different and competing structures can be compared. However, it is also possible to analyse individual structures with respect to the third criterion mentioned in the beginning of section 3, namely the model assumptions.

*3.2.1. Failure, Inference and Improved Hypotheses*. Recently, *Gupta et al.* [1998; see also *Boyle et al.*, 2000 and *Wagener et al*., 2001a] showed how a multi-objective approach can be applied to give an indication of structural inadequacies. The assumption is that a model should be capable of representing all response modes of a hydrological system with a single parameter set. A failure to do so indicates that a specific model hypothesis is not suitable and should be rejected, or preferably, replaced by a new hypothesis which improves on the old one. This idea was already the basis of some of Beck's work [e.g. *Beck*, 1985], as described earlier in the text. *Wagener et al*. [2001c] developed a new approach based on this assumption. Their methodology analyses the identifiability of parameters within a selected model structure in a dynamic and objective manner, which can be used to analyze the consistency of locations of good performing parameter values in (parameter) space and in time.

In cases where the variation of parameter optima can be tracked in time it will sometimes be possible to directly relate changes in a particular parameter to variations in forcing or state variables [examples in *Beven*, 2000, p. 93ff.; and *Bashford and Beven*, 2000]. However, in many cases the development of improved hypotheses will be more complex and depend on the capability of the hydrologist. Unfortunately(?), there is no logical way to create new ideas; the hydrologist therefore has to apply his depth of insight and creative imagination to derive a new hypothesis, which can replace the old one, that has failed.

*3.2.2. Dynamic Identifiability Analysis*. The DYNamic Identifiability Analysis (DYNIA) is a new approach to locating periods of high identifiably for individual parameters and to detect failures of model structures in an objective manner. The proposed methodology draws from elements of the popular Regional Sensitivity Analysis [RSA; *Spear and Hornberger*, 1980; *Hornberger and Spear*, 1981] and includes aspects of the Generalized Likelihood Uncertainty Estimation [GLUE, *Freer et al*., this volume] approach, wavelet analysis [e.g. *Gershenfeld*, 1999] and the use of Kalman filtering for hypothesis testing as applied by *Beck* [1985].

In the original RSA approach, a model population is sampled from a uniform distribution. This population is divided into behavioural and non-behavioural models depending on whether a model resulted in a certain response or not [*Spear and Hornberger*, 1980]. *Beven and Binley* [1992] extended the approach by conditioning the model population on a likelihood measure, which in their case, can be a transformation of any measure of performance. These are the building blocks from which a new method of assessing the identifiability of parameters is created [*Wagener et al*., 2001c].

The steps taken in the procedure can be seen in the flow chart in Figure 4. Monte-Carlo sampling based on a uniform prior distribution is used to examine the feasible parameter space. The objective function associated with each parameter set, i.e. model, is transformed into a support measure, i.e. all support measures have the characteristic that they sum to unity and higher values indicate better performing parameter values. These are shown here in form of a dotty plot (Fig. 4(a)). The best performing parameter values (e.g. top 10 %) are selected and their cumulative distribution is calculated (Fig. 4(b)). A straight line will indicate a poorly identified parameter, i.e. the highest support values are widely distrib-

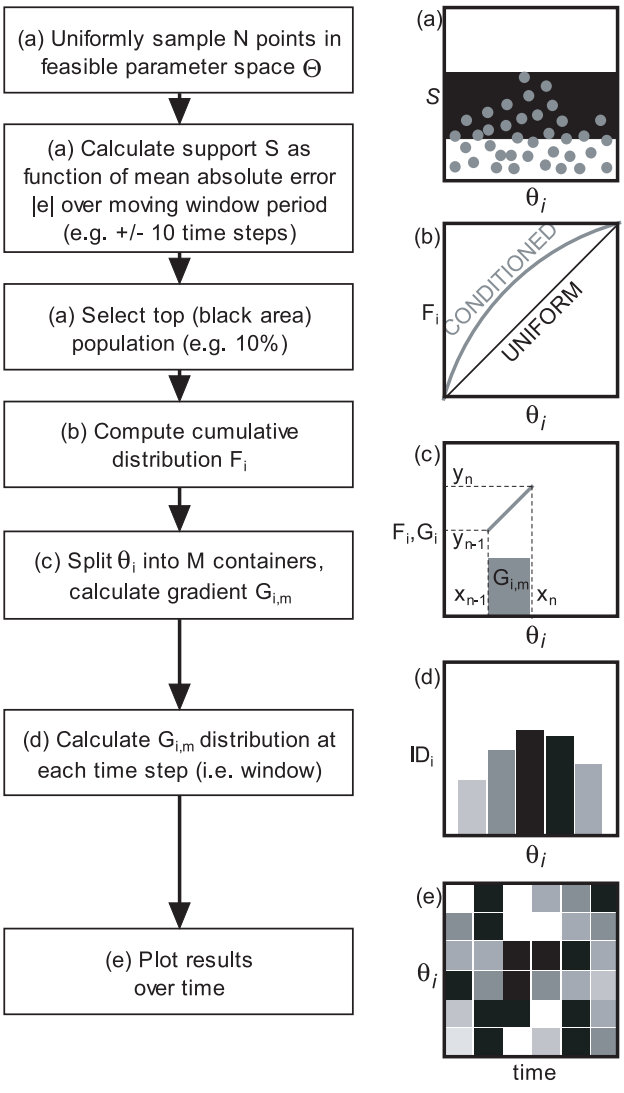


**Figure 4.** Schematic description of the DYNamic Identifiability Analysis (DYNIA) procedure.

uted over the feasible range. Deviations from this straight line indicate that the parameter is conditioned by the objective function used. The gradient of the cumulative support is the marginal probability distribution of the parameter, and therefore an indicator of the strength of the conditioning, and of the identifiability of the parameter. Segmenting the range of each parameter (e.g. into 20 containers) and calculating the gradient in each container leads to the (schematic) distribution shown in Fig. 4(d). The highest value, additionally indicated by the darkest colour, marks the location (within the chosen resolution) of greatest identifiability of the parameter. *Wagener et al*. [2001a] show how this measure of identifiability can be used to compare different model structures in terms of parameter uncertainty, which is assumed to be inversely related to identifiability. They calculate the identifiability as a function of measures of performance for the whole calibration period and for specific response modes, derived using the segmentation approach by *Boyle et al*. [2000] described earlier in the text. It can be shown that the identifiability of some parameters, and therefore individual model components, is greatly enhanced by this segmentation [*Wagener et al*., 2001a].

Calculating the parameter identifiability at every time step using only the residuals for a number of time steps n before and after the point considered, i.e. a moving window or running mean approach, allows the investigation of the identifiability as a function of time (Fig. 4(e)). The gradient distribution plotted at time step t therefore aggregates the residuals between t-n and t+n, with the window size being 2n+1. The number of time steps considered depends upon the length of the period over which the parameter is influential. For example, investigation of a slow response linear store residence time parameter requires a wider moving window than the analysis of a quick response residence time parameter. Different window sizes are commonly tested and the ones most appropriate are used to analyse individual parameters. A very small window size can lead to the result being largely influenced by errors in the data. However, this is not a problem where the data quality is very high, for example in the case of tracer experiments in rivers [*Wagener et al*., 2001d]. Conversely, if the window size is too big, periods of noise and periods of information will be mixed and the information will be blurred.

The results are plotted for each parameter versus time using a colour coding where a darker colour indicates areas, in parameter space and time, of higher identifiability. Care has to be taken when interpreting the DYNIA results of time steps at the beginning and the end of time-series. Here the full window size cannot be established and the result is distorted. This is an effect similar to the cone of influence in wavelet analysis [*Torrence and Compo*, 1998].

While this approach is not intended to evaluate parameter dependencies in detail, the significance of dependencies to the identifiability is implicit in the univariate marginal distribution which is structurally represented by Figure 4(d). A strong dependency during any period would tend to inhibit the information of a strong univariate peak, i.e. the effect of the involved parameters cannot be singled out. Parameter interdependence can be estimated in detail by the investigation of the response surface or the variance-covariance matrix [e.g. *Wheater et al*., 1986; *Hornberger et al*., 1985].

A limitation of the proposed measure of identifiability arises if any near-optimal parameter values are remote from the identified peak of the marginal distribution, as the rele-

vance of such values would be diminished. It is therefore important that a detailed investigation of the dotty plots is used to verify periods of high identifiability. The approach also requires that feasible parameter ranges are defined sensibly and the selected model population (usually the best 10%) represents only the top of the distributions.

DYNIA requires that sensible feasible ranges for each parameter can be defined and that the number of models (i.e. parameter sets) considered reflects the shape of the response surface. The procedure can then be applied to separate periods that do and those that do not contain information about specific parameters, and track parameter variations in time.

The subjective decision for a particular objective function in this procedure is usually not critical for the result and the mean absolute error criterion is usually adopted.

### *3.3. A Combined Framework of Corroboration and Rejection*

The earlier introduced multi-objective framework [*Wagener et al*., 2001a] can be extended to incorporate the DYNIA approach as an additional step in order to derive a framework of corroboration and rejection (Figure 5). Similar frameworks are for example proposed by *Beven* [2000, p.297ff.], and, more generally, by *Oreskes et al*. [1994].

The initial steps are similar to those in the multi-objective framework described earlier. The hydrologist selects (or develops) model structures that seem suitable for the given modeling purpose, watershed characteristics and data.

One can then apply a multi-objective procedure to establish preferences between the competing model structures, or preferably structural components. Under-performing structures (components) can be rejected at this stage, based on their performance and/or uncertainty.

During the next stage, the DYNIA approach can be used to further analyze the remaining model structures. Further rejections might be possible. The suitability of a model structure not failing is further corroborated. A model structure is (temporarily) accepted when no better performing structure can be found and no underlying assumption is violated.

In the last stage, the parameter space 'within' the remaining model structures can be analyzed to find all those models, i.e. parameter sets that are in line with the behavior of the natural system. It is very likely that such a procedure will result in a range of acceptable or 'behavioral' models or even model structures. The appropriate response is to combine the predictions of all models to derive an ensemble prediction of the systems behavior. A popular approach to do so is the GLUE approach [*Freer et al.*, this volume], however, other methods to combine the predictions of different models are possible [e.g. *Shamseldin et al*., 1997]. Within the GLUE approach, a likelihood value is derived for every model. The models are usually drawn from a uniform distribution. Basically any measure of performance which can be transformed so that higher values indicate better models and all measures add up to one, can be used as a likelihood measure in this approach. The likelihoods are then used to weight the prediction of every model at every time step. The cumulative distribution of the weighted streamflow values, even for different models, allows the extraction of percentiles, e.g. 5% and 95%, to derive the, in this case, 90% confidence limits for the predictions. The likelihoods of different models could be combined through simple addition.

## 4. APPLICATION EXAMPLE

### *4.1. Modelling Tools and Selected Model Structures*

The Rainfall-Runoff Modelling Toolbox (RRMT) and Monte-Carlo Analysis Toolbox (MCAT), developed at Imperial College, are used here for calculation and visualisation of results [*Wagener et al*., 1999; 2001b].

The RRMT has been developed in order to produce parsimonious, lumped model structures with a high level of

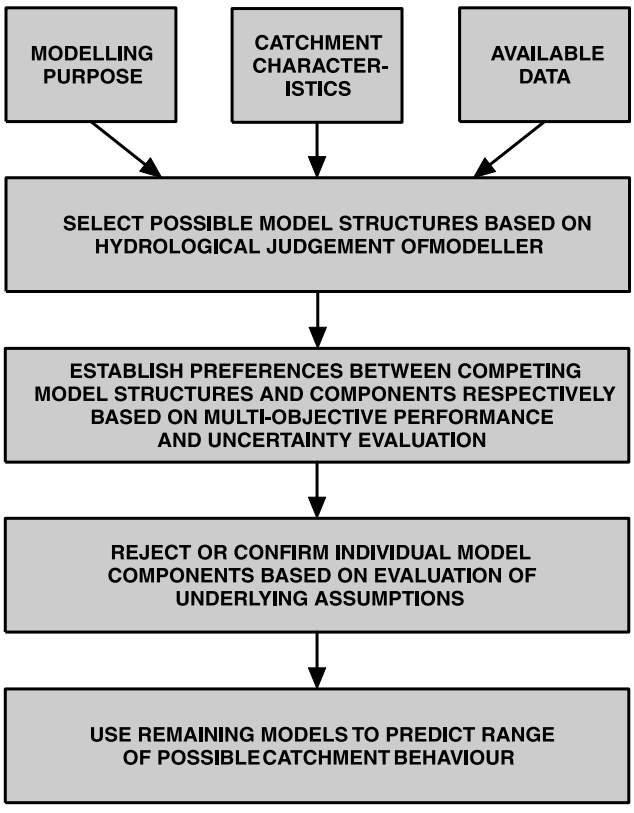


**Figure 5.** The proposed modeling procedure.

parameter identifiability. It is a generic modelling shell allowing its user to implement different model structures to find a suitable balance between model performance and parameter identifiability. Model structures that can be implemented are spatially lumped, relatively simple (in terms of number of parameters), and of conceptual or hybrid metric conceptual type. Hybrid metric-conceptual models utilise observations to test hypotheses about the model structure at watershed scale and therefore combine the metric and the conceptual paradigm [*Wheater et al*., 1993]. All structures consist of a moisture accounting and a routing module.

MCAT is a collection of analysis and visualisation functions integrated through a graphical user interface. The toolbox can be used to analyse the results from Monte-Carlo parameter sampling experiments or from model optimisation methods that are based on population evolution techniques, for example, the SCE-UA [*Duan*, this volume] or the MOCOM-UA [*Gupta et al*., this volume, "Multiple …"] algorithms. Although this toolbox has been developed within the context of ongoing hydrological research, all functions can be used to investigate any dynamic mathematical model. Functions contained in MCAT include an extension of the Regional Sensitivity Analysis [RSA, *Spear and Hornberger*, 1980] by *Freer et al*. [1996], various components of the Generalised Likelihood Uncertainty Estimation method [GLUE, *Freer et al*., this volume], options for the use of multiple-objectives for model assessment [*Gupta et al*., 1998; *Boyle et al*., 2000], and plots to analyse parameter identifiability and interaction.

Both toolboxes are implemented in the Matlab [*Mathworks*, 1996] programming environment.

A large variety of lumped parsimonious model structures can be found in the literature [e.g. *Singh*, 1995]. However, the range of components on which these structures are based is relatively small. Some of the most commonly found components are selected here in a component library shown in Figure 6. Further details about these components can be found in *Wagener et al*. [2001b; and in the references given here].

The soil moisture accounting components used are:

- The catchment moisture deficit [cmd, *Evans and Jakeman*, 1998]. A conceptual bucket with a bottom outlet to sustain drainage into the summer periods.
- The catchment wetness index [cwi, *Jakeman and Hornberger*, 1993]. A metric approach based on the Antecedent Precipitation Index [API, e.g. *Shaw*, 1994].
- The probability distributed soil moisture stores [pd3 and pd4, *Moore*, 1999]. A probability distribution of conceptual buckets based on a Pareto distribution. Evapotranspiration is either at the potential rate, as long as soil moisture is available, or at a rate declining linearly with soil moisture content.
- A simple bucket type structure (buc), evaporating at the potential rate as long as soil moisture is available.
- The Penman storage model [*Penman*, 1949]. A layered structure of two conceptual buckets connected by an overflow mechanism. Evapotranspiration occurs at potential rate from the upper layer, similar to the root zone, and at a reduced rate, 12% of PE, from the bottom layer. An additional bypass mechanism diverts a fraction of the rainfall from the SMA component to contribute to the effective rainfall at time-steps where rainfall exceeds PE.

The routing components used are:

- Conceptual reservoirs in various combinations and in linear and non-linear form [e.g. *Wittenberg*, 1999].

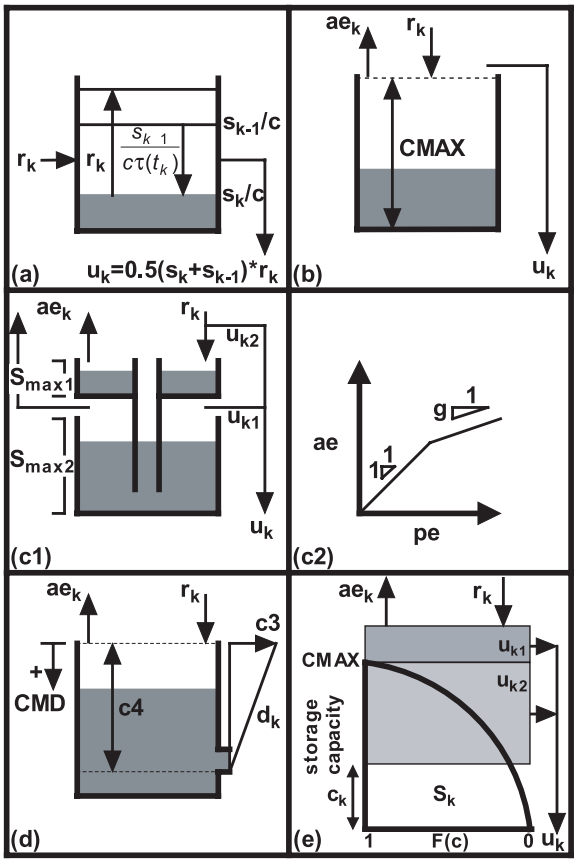


**Figure 6.** Table showing the soil moisture accounting 'component library' used in the application example. The components are: (a) catchment wetness index (cwi), (b) simple bucket (buc), (c1) and (c2) Penman structure (ic1), (d) catchment moisture deficit (cmd), and probability distribution of soil moisture stores (pdX).

*4.2. Data*

The river selected for this study is the Lower Medway at Teston (1256.1 $km^2$) located in South Eastern England. Six years (10/04/1990 – 14/07/1996) of data (daily naturalised flows, precipitation, potential evapotranspiration (PE) and temperature) are available. The Medway watershed is characterised by a mixture of permeable (chalk) and impermeable (clay) geologies subject to a temperate climate with an average annual rainfall of 772 mm and an average annual PE of 663 mm (1990-1996).

*4.3. Methodology*

Multi-objective (MO) analysis and DYNIA are performed, based on the results of Monte Carlo sampling procedures. For the MO analysis, 20000 parameter sets, i.e. models, are randomly sampled from the feasible parameter space for each individual model structure, based on a uniform distribution.

For each of these models, five OFs are calculated. These are the overall RMSE and four OFs derived for different response modes of the watershed. The segmentation applied is based on an approach by *Wagener and Wheater* [2001] which uses the slope of the hydrograph and an additional threshold as segmentation criteria to split the hydrograph into different response modes. The slope separates periods when the watershed is wetting up or is "driven" [*Boyle et al*., this volume] by rainfall, i.e. positive slope, and when the watershed is draining, i.e. falling slope. A threshold is used to separate periods of high and low flow, i.e. the mean flow during driven and 50% of the mean flow during drainage periods. Four OFs are therefore derived when the residuals during the different periods are aggregated separately using the RMSE criterion: FDH, "driven" flow during high flow, FDL, "driven" flow during low flow, FQ, quick drainage (high flows), and FS, slow drainage (low flows). This is a modification of the initial approach by *Boyle et al*. [2000], which was based on the analysis of flow and rainfall. However, the approach presented here has been shown to be more suitable for British watersheds as modelled in the example presented here. These OFs are based on the assumption that different processes are dominant during periods of high and low flow, and during periods of watershed wetting-up and drainage. The residuals, i.e. the differences between observed and simulated flows are calculated and summarised in form of the root mean squared error for each period. The performance and identifiability analysis is based on these measures.

The resulting parameter populations are used to rank all models or model structures, with respect to their performance and identifiability, using the measures introduced earlier. The best model structures are retained and a more thorough analysis using DYNIA is performed. DYNIA is based on a random sampling procedures using 2500 parameter sets collected from a uniform distribution. The smaller sample size is due to computational limitations of the current DYNIA application in the Matlab [*Mathworks*, 1996] environment.

*4.4. Results and Discussion*

The main results of the MO analysis as shown in Figure 7 are as follows:

- At a general level for the SMA modules (Figure 7, top): the probability distributions of storage elements (pd3 and pd4) seem to perform best, followed by the simple bucket (buc), and the cmd and cwi modules.
- The cm1, i.e. a cmd that always evaporates at the potential rate, performs much more poorly than the rest with respect to those objective functions which mainly describe periods of high flow, RMSE(total), FDH and FQ. This is also the case for the cmd module, but not as pronounced. However, the cmd and cm1 modules do very well during low flow periods. This is caused by the bottom outlet of the bucket, which sustains the production of effective rainfall even during periods of severe moisture deficits in the SMA module.
- The overall result of the performance analysis is that the pd3 and pd4 SMA modules in combination with 2pll or 2pln routing modules are superior. The cmd is a useful component when the modelling purpose demands the accurate prediction of low flow periods and periods of high flows are of minor importance.
- A detailed analysis of the routing components shows that the use of a non-linear conceptual reservoir in parallel with a linear one (2pln), performs better at the peaks (RMSE(total) and FDH), see Figure 7(top).
- The uncertainty analysis (Figure 7, bottom) however reveals that the identifiability of the cmd parameters is very low and this module is rejected here on this basis. For some applications, this aspect might be of minor importance, however.

The pd3 and the pd4 SMA components are retained for further analysis with the DYNIA approach. Assuming that our interest is in low flows, e.g. for water resources purposes, only a linear parallel routing structure (2pll) is considered. A non-linear component would be advisable for high flow periods.

The results of the DYNIA are shown in Figures 8 and 9, for the structures pd3-2pll and pd4-2pll. This reveals some problems with the pd3 SMA module.

Figure 8 shows the dynamic identifiability of the five parameters of the pd3-2pll structure. These are: (1) cmax, the

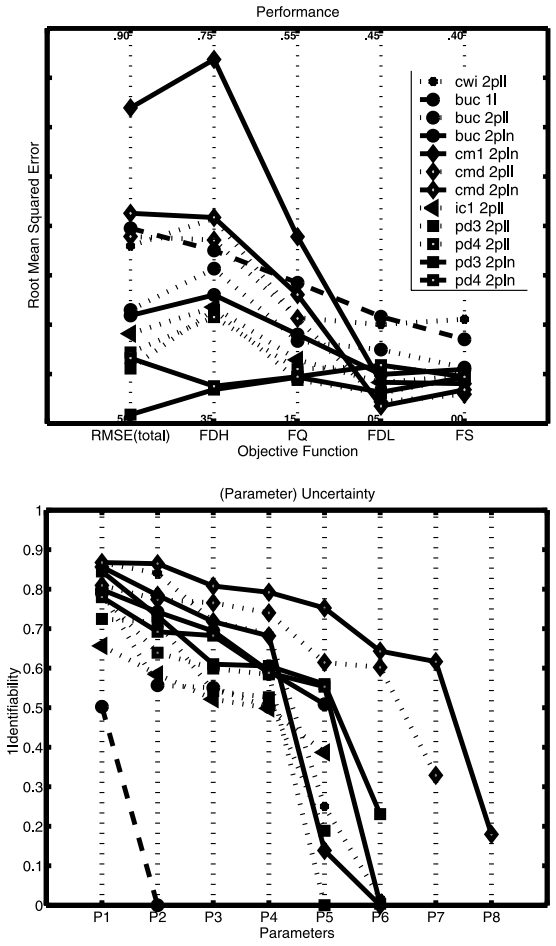


**Figure 7.** Results of the model structure comparison.

maximum storage capacity, (2) b, the shape parameter of the Pareto distribution of storage capacities, (3) k(quick), the residence time of the quick linear reservoir, (4) alpha, the fraction of flow going through the quick flow component, and (5) k(slow), the residence time of the slow flow linear reservoir.

The plot for the parameter cmax exposes some ambiguity about the optimum values for this parameter. The confidence limits (cfls) narrow into two different parts of the parameter space, towards low values after wet periods and towards high values during periods of wetting up, indicating inadequacies within the model structure. Similarly, but much less pronounced, the parameter b shows a slight shift of optimum after the wet period, i.e. the lower cfls go up. It is mainly identifiable during low flow events (e.g. dark areas just before time step 700). The residence times of the routing component show the expected behaviour, i.e. the cfls of k(quick) narrow down on the quick falling limbs of the hydrograph, while darker areas appear for k(slow) during the long recessions. The cfls for k(slow) hardly narrow during periods of identifiability, suggesting that the peaks on the response surface are quite small, and that the difference between different values for this parameter is not large. Values for this parameter are therefore still widespread, since the top 10% are selected here. The example of the two residence times also demonstrates the need for different window sizes. A small size (11 time steps) is required for k(quick), whose influence is only very local, while a much larger window (81 time steps) is need to capture the effect of k(slow). Finally, the parameter alpha is most identifiable during periods where the split between quick and slow response is occurring. However, further investigations, which are outside the scope of this example, are required to explain the behaviour of this parameter. In general, this structure is too simplistic to reproduce all aspects of the hydrograph with one parameter set. This is especially reflected in the results for cmax.

The difference between pd3 and pd4 is that, while pd3 always evaporates at the potential rate, pd4 decreases the evapotranspiration with decreasing soil moisture content in a linear manner. However, without adding an additional (scaling) parameter, i.e.

$$AE_t = S_t / S_{max} \cdot PE_t \quad (1)$$

The effect of this change can be seen in the dynamic results shown in Figure 9. The ambiguity with respect to cmax is removed and the cfls only narrow towards larger values indicating a better structure.

It is interesting to remember that the MO performance analysis had shown that the pd3 component actually performed better. The reason is that the pd4 component puts an additional constraint on the behaviour of the watershed system. The result is that the structure becomes less flexible. The pd3 component can therefore perform better with respect to the different OFs. However, this is due to the expense of a larger variation in parameter values as shown in the dynamic analysis. This indicates that pd4 is actually the better SMA component and should be retained, while pd3 should be rejected. This result supports the statement by *Gupta et al*. [2001] that consistency in a model is more important than optimality.

## 5. SUMMARY AND CONCLUSIONS

*Test everything. Hold on to the good. Avoid every kind of evil.*
1 Thessalonians 5, 21:22, New International Version

The identification of suitable conceptual rainfall-runoff (CRR) models is a difficult problem. It has been increased by the recent awareness of the influence of model structural inadequacies.

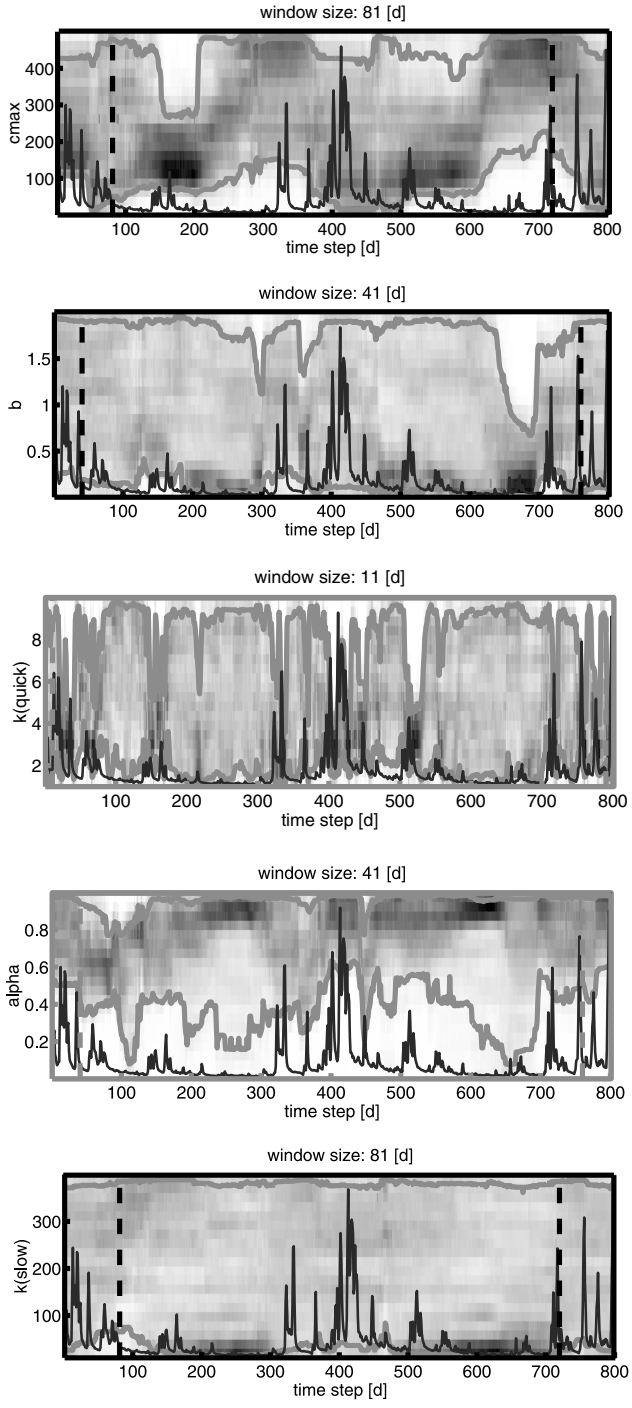


**Figure 8.** DYNIA results for pd3-2pll.

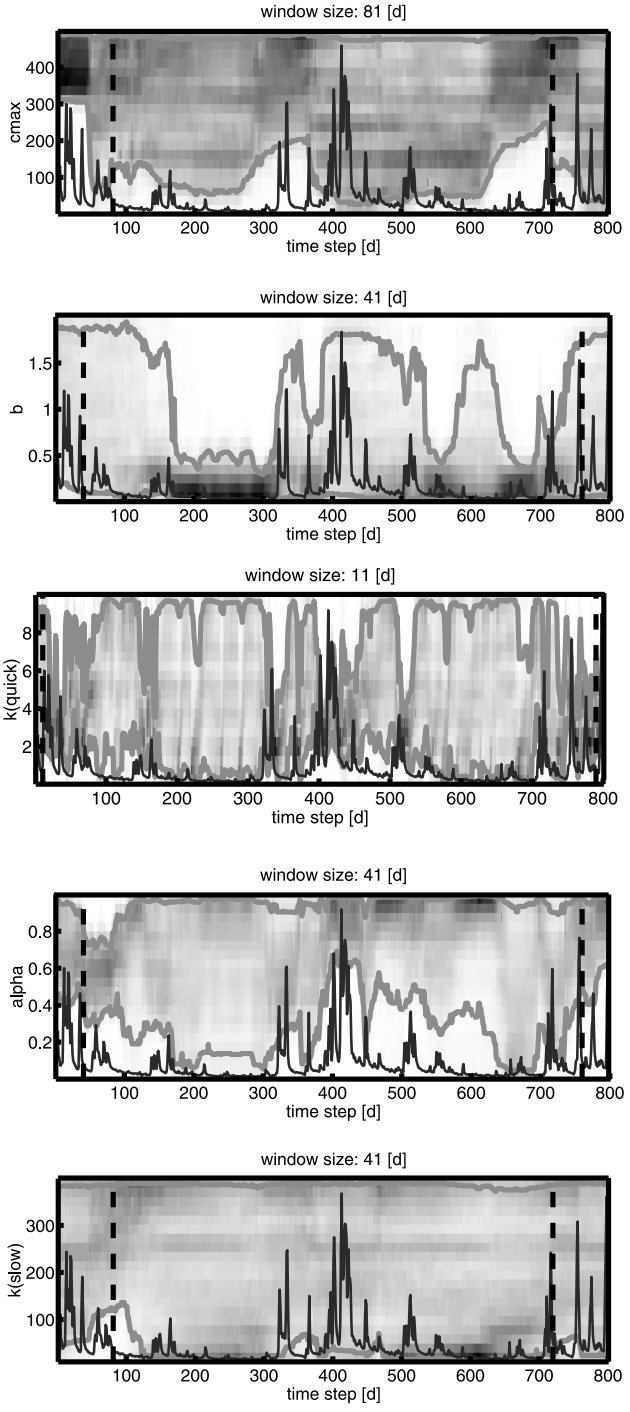


**Figure 9.** DYNIA results for pd4-2pll.

A framework of corroboration and rejection is presented to embed the identification problem into a scientific method as outlined by *Popper* [2000]. The framework uses multi-objective and novel dynamic approaches to the evaluation of CRR models and model structures. The theory and methods underlying this framework are described and an application example is presented. It demonstrates that a range of approaches is required for an objective analysis of the suitability of models and model structures.

DYNIA is an attempt to develop an approach to complement traditional calibration methods resulting in increased discriminative power. Advantages of the approach are its simplicity and its general applicability (for example, an application to a solute transport model can be found in *Wagener et al*. [2001d]). Possible areas of application of DYNIA are [see *Wagener et al*., 2001c for details]: (1) the pure estimation of parameters, (2) the analysis of model structures, (3) relating model parameters and response modes, and (4) to investigate data outliers and anomalies.

Current work is focusing on the extension of this framework to include the identification of CRR models at ungauged sites, using parameter regionalisation approaches.

The RRMT and MCAT toolboxes are available for download free of charge for non-commercial use from the Environmental and Water Resource Engineering Section Web-site on http://ewre.cv.ic.ac.uk/software.

*Acknowledgments.* This project is funded by NERC under grant GR3/11653. We thank Southern Water for providing the data used in the example application. We also thank Matthew J. Lees and Neil McIntyre for constructive criticism on the presented work.

## REFERENCES

Anderson, M. G., and T. P.Burt, Modelling strategies, in *Hydrological forecasting*, edited by M. G. Anderson and T. P. Burt, pp. 1-13, John Wiley and Sons, Chichester, UK, 1985.

Bard, Y., *Non-linear parameter estimation*, Academic Press, 1974.

Bashford, K., and K. J. Beven, Model structures, observational data and predictive uncertainty: explorations using a virtual reality, in Proceedings 7th National Hydrology Symposium, edited by C. Kirby, Newcastle, UK, pp. 3.31-3.37, 2000.

Beck, M. B., Structures, failure, inference and prediction, in *Identification and System Parameter Estimation*, edited by M. A. Barker and P. C. Young, Proceedings IFAC/IFORS 7th Symposium Volume 2, July 1985, York, UK, pp. 1443-1448, 1985.

Beck, M. B., Water quality modelling: a review of the analysis of uncertainty, *Water Resour. Res.*, 23, 1393-1442, 1987.

Beck, M. B., A. J. Jakeman and M. J. McAleer, Construction and evaluation of models in environmental systems, in *Modelling change in environmental systems*, edited by A. J. Jakeman, M. B. Beck and M. J. McAleer, John Wiley and Sons Ltd., USA, pp. 3-35, 1993.

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

Beven, K. J., *Rainfall-runoff modelling – The primer*, John Wiley and Sons Ltd, Chichester, UK, 2000.

Beven, K. J., and A. M. Binley, The future of distributed models: Model calibration and uncertainty in prediction, *Hydrol. Proc.*, 6, 279-298, 1992.

Beven, K. J., R. Lamb, P. Quinn, R. Romanovicz, and J. Freer, Topmodel, in *Computer models of watershed hydrology*, edited by V. P. Singh, Water Resources Publishers, pp. 627-668, 1995.

Beveridge, W. I. B., *The art of scientific investigation*, 3rd edition, William Heinemann Ltd, Melbourne, Australia, 1957.

Boogard, H. F. P. van den, M. S. Ali, and A. E. Mynett, Self organising feature maps for the analysis of hydrological and ecological data sets, in *Hydroinformatics'98*, edited by V. Babovic and L. C. Larsen, Balkema, Rotterdam, NL, pp. 733-740, 1998.

Boyle, D.P., H. V. Gupta, and S. Sorooshian, Towards improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, 36, 3663-3674, 2000.

Boyle, D.P., H. V. Gupta, S. Sorooshian, V. Koren, Z. Zhang, and M. Smith, Towards improved streamflow forecasts: The value of semi-distributed modelling, *Water Resour. Res*., 37, 2739-2759, 2001.

Chappell, N.A., S.W. Franks and J. Larenus, Multi-scale permeability estimation for a tropical catchment, *Hydrological Processes*, 12, 1507-1523, 1998.

Duan, Q., V. K. Gupta, and S. Sorooshian, Effective and efficient global optimisation for conceptual rainfall-runoff models, *Water Resour. Res*., 28, 1015-1031, 1992.

Dunne, S. M., Imposing constraints on parameter values of a conceptual hydrological model using baseflow response, *Hydrol. Earth Syst. Sci.*, 3, 271-284, 1999.

Ehrgott, M., *Multicriteria optimization*, Springer-Verlag, Berlin, Germany, 2000.

Evans, J. P., and A. J. Jakeman, Development of a simple, catchment-scale, rainfall-evapotranspiration-runoff model, *Env. Model. Softw.*, 13, 385-393, 1998.

Franks, S. W., P. Gineste, K. J. Beven, and P. Merot. On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas in the calibration process, *Water Resour. Res.*, 34, 787-797, 1998.

Freer, J., K. J. Beven, and B. Ambroise, Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res.*, 32, 2161-2173, 1996.

Gershenfeld, N., *The nature of mathematical modeling*, Cambridge University Press, Cambridge, UK, 1999.

Goldberg, D. E., *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, USA, 1989.

Gupta, H. V., Penman Lecture, paper presented at 7th BHS National Symposium, Newcastle-upon-Tyne, UK, 2000.

Gupta, V. K., and S. Sorooshian, The relationship between data and the precision of parameter estimates of hydrologic models, *J. Hydrol.*, 81, 57-77, 1985.

Gupta, H. V., S. Sorooshian, and P. O. Yapo, Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, 34, 751-763, 1998.

Gupta, H. V., S. Sorooshian, and D. P. Boyle, Assimilation of data for the construction and calibration of watershed models, Keynote paper presented at the International Workshop on Catchment Scale Hydrologic Modelling and Data Assimilation, Wageningen, September 2001, NL, 2001.

Harlin, J., Development of a process oriented calibration scheme for the HBV hydrological model, *Nordic Hydrol.*, 22, 15-36, 1991.

Harvey, J. W., and B. J. Wagner, Quantifying hydrologic interactions between streams and their subsurface hyporheic zones, in

*Streams and ground waters*, edited by J. A. Jones and P. J. Mulholland, Academic Press, San Diego, USA, pp. 3-44. 2000.

Hornberger, G. M., and R. C. Spear, An approach to the preliminary analysis of environmental systems, *J. Env. Management*, 12, 7-18, 1981.

Hornberger, G. M., K. J. Beven, B. J. Cosby, and D. E. Sappington, Shenandoah watershed study: Calibration of the topography-based, variable contributing area hydrological model to a small forested catchment, *Water Resour. Res.*, 21, 1841-1850, 1985.

Jakeman, A. J. and G. M. Hornberger, How much complexity is warranted in a rainfall-runoff model? *Water Resour. Res.*, 29, 2637-2649, 1993.

Kleissen, F. M., Uncertainty and identifiability in conceptual models of surface water acidification, Ph.D. thesis, Imperial College of Science, Technology and Medicine, London, UK, 1990.

Kleissen, F. M., M. B. Beck, and H. S. Wheater, The identifiability of conceptual hydro-chemical models, *Water Resour. Res.*, 26, 2979-2992, 1990.

Kuczera, G., and M. Mroczkowski, Assessment of hydrologic parameter uncertainty and the worth of multiresponse data, *Water Resour. Res.*, 34, 1481-1489, 1998.

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

Legates, D. R., and G. J. McCabe Jr., Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, *Water Resour. Res.*, 35, 233-241, 1999.

Magee, B., *Popper*, 6th Impression, Fontana/Collins, Glasgow, Scotland, 1977.

Mathworks, *Matlab – reference guide*, The Mathworks Inc., Natick, M.A., 1996.

Moore, R. J., Real-time flood forecasting systems: Perspectives and prospects, in *Floods and landslides: Integrated risk assessment*, edited by R. Casale and C. Margottini, Springer, Berlin, pp. 147-189, 1999.

Mous, S. L. J., Identification of the movement of water in unsaturated soils: the problem of identifiability of the model, *J. Hydrol.*, 143, 153-167, 1993.

Mroczkowski, M., G. P. Raper, and G. Kuczera, The quest for more powerful validation of conceptual catchment models, *Water Resour. Res.*, 33, 2325-2335, 1997.

Nash, J. E., and J. V. Sutcliffe, River flow forecasting through conceptual models, I, A discussion of principles, *J. Hydrol.*, 10, 282-290, 1970.

NWS, 2001. Calibration of the Sacramento model structure, http://hsp.nws.noaa.gov/oh/hrl/calb/ workshop/parameter.htm. Accessed 15th February 2001.

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

Penman, H. L., The dependence of transpiration on weather and soil conditions, *J. Soil Sci.*, 1, 74-89, 1949.

Piñol, J., K. J. Beven, and J. Freer, Modelling the hydrological response of Mediterranean catchments, Prades, Catalonia. The use of distributed models as aid to hypothesis formulation, *Hydrol. Proc.*, 11, 1287-1306, 1997.

Popper, K., *The logic of scientific discovery*, first published 1959 by Hutchinson Education, Routledge, UK, 2000.

Seibert, J., Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrol. Earth Syst. Sci.*, 4, 215-224, 2000.

Shamseldin, A. Y., K. M. O'Connor, and G. C. Liang, Methods of combining the outputs of different rainfall-runoff models, *J. Hydrol.*, 197, 203-229, 1997.

Shaw, E. M., *Hydrology in practice*, 3rd Edition, Chapman and Hall, London, UK, 1994.

Singh, V. P., *Computer models of watershed hydrology*, Water Resources Publishers, USA, 1995.

Sorooshian, S., and V. K. Gupta, The analysis of structural identifiability: Theory and application to conceptual rainfall-runoff models, *Water Resour. Res.*, 21, 487-495, 1985.

Spear, R. C., and G. M. Hornberger, Eutrophication in Peel Inlet, II, Identification of critical uncertainties via generalized sensitivity analysis, *Water Res.*, 14, 43-49, 1980.

Stigter, J. D., M. B. Beck, and R. J. Gilbert, Identification of model structure for photosynthesis and respiration of algal populations, *Water Sci. Technol.*, 36, 35-42, 1997.

Torrence, C., and G. P. Compo, A practical guide to wavelet analysis, *Bull. American Meteorol. Soc.*, 79, 61-78, 1998.

Uhlenbrock, S., J. Seibert, C. Leibundgut, and A. Rohde, Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure, *Hydrol. Sci. J.*, 44, 779-797, 1999.

Van Straten, G., and K. J. Keesman, Uncertainty propagation and speculation in projective forecasts of environmental change: a lake-eutrophication example, *J. Forec.*, 10, 163-190, 1991.

Wagener, T. and H. S. Wheater, On the evaluation of conceptual rainfall-runoff models using multiple-objectives and dynamic identifiability analysis, in *Continuous river flow simulation: Methods, applications and uncertainties*, edited by I. Littlewood, and J. Griffin, British Hydrological Society - Occasional Papers, Wallingford, UK, 45-51, 2001.

Wagener, T., M. J. Lees, and H. S. Wheater, A generic rainfall-runoff modelling toolbox, *Eos Trans. AGU*, 80, Fall Meet. Suppl., F203, 1999.

Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian, A framework for the development and application of hydrological models, *Hydrol. Earth Syst. Sci.*, 5(1), 13-26, 2001a.

Wagener, T., M. J. Lees, and H. S. Wheater, A toolkit for the development and application of parsimonious hydrological models, in *Mathematical models of large watershed hydrology – Volume 1*, edited by V. P. Singh and D. Frevert, Water Resources Publishers, USA, pp. 87-136, 2001b.

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta, Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrol. Proc.*, in press, 2001c.

Wagener, T., L. A. Camacho, M. J. Lees, and H. S. Wheater, Dynamic parameter identifiability of a solute transport model, in *Advances in design sciences and technology*, edited by R. Beheshti, europia, Delft, April 2001, NL, pp. 251-264, 2001d.

Wagner, B. J., and J. W. Harvey, Experimental design for estimat-

ing parameters of rate-limited mass transfer: Analysis of stream tracer studies, *Water Resour. Res.*, 33, 1731-1741, 1997.

Wheater, H. S., K. H. Bishop, and M. B. Beck, The identification of conceptual hydrological models for surface water acidification, *Hydrol. Proc*., 1, 89-109, 1986.

Wheater, H. S., A. J. Jakeman, and K. J. Beven, Progress and directions in rainfall-runoff modelling, in *Modelling change in environmental systems*, edited by A. J. Jakeman, M. B. Beck and M. J. McAleer, Wiley, Chichester, UK, pp. 101-132, 1993.

Wittenberg, H., Baseflow recession and recharge as non-linear storage processes, *Hydrol. Proc.*, 13, 715-726, 1999.

Yan, J., and C. T. Haan, Multiobjective parameter estimation for hydrologic models – Weighting of errors, *Trans. Am. Soc. Agric. Eng.*, 34(1), 135-141, 1991.

Yapo, P. O., H. V. Gupta, and S. Sorooshian, Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *J. Hydrol.*, 18, 23-48, 1996.

Young, P. C., S. Parkinson, and M. J. Lees, Simplicity out of complexity in environmental modeling: Occam's razor revisited, *J. Appl. Stat*., 23, 165-210, 1996.

---

Hoshin V. Gupta. SAHRA, NSF STC for the Sustainability of semi-Arid Hydrology and Riparian Areas, Department of Hydrology and Water Resources, Harshbarger Bldg. 11, University of Arizona, Tucson, AZ 85721, USA

(hosh_stc@sahra.arizona.edu)

Thorsten Wagener, now at: SAHRA, NSF STC for the Sustainability of semi-Arid Hydrology and Riparian Areas, Department of Hydrology and Water Resources, Harshbarger Bldg., University of Arizona, Tucson, AZ 85721, USA

Howard S. Wheater, Department of Civil and Environmental Engineering, Imperial College of Science, Technology and Medicine, Imperial College Road, SW7 2BU, London, UK

(t.wagener@ic.ac.uk)

(h.wheater@ic.ac.uk)