

## Towards a Formal Approach to Calibration and Validation of Models Using Spatial Data

Jens Christian Refsgaard

### 13.1 INTRODUCTION

Spatially distributed models of catchment response tend to be highly complex in structure and contain numerous parameter values. Their calibration and validation is therefore extremely difficult but at the same time essential to obtain confidence in the reliability of the model (Chapter 3). Traditionally, calibration and validation has been mainly based on a comparison of observed versus simulated runoff at the outlet of the catchment of interest, but it has been pointed out numerous times that this is a very weak test of the adequacy of a model. Ideally, the internal behaviour of the models, in terms of simulated patterns of state variables and model output, should be tested, but examples of such internal tests are only relatively recent. In 1994 Rosso (1994, pp.18–19) pointed out that “Conventional strategies for distributed model validation typically rely on the comparison of simulated model variables to observed data for specific points representing either external boundaries or intermediate locations on the model grid.... Traditional validation based on comparing simulated with observed outflows at the basin outlet still remains the only attainable option in many practical cases. However, this method is poorly consistent with spatially distributed modeling...”. More recently, encouraging work has been done on demonstrating how observed spatial patterns of hydrologically relevant variables can be used for internal tests. Indeed, the case study chapters of this book (Chapters 6–12) have clearly illustrated the enormous value of detailed spatial data in developing and testing distributed representations of catchment hydrological processes. These chapters have used a plethora of different data types and models and are representative of the progress in this area within the scientific community. However, typically these studies have been performed in small, well instrumented research catchments. The models often have been developed or modified by the group of people who did the data collection and also were the users of the results; and the purpose of the model development was to obtain insight into spatial catchment processes and process interactions.

These conditions are quite different in practical applications. In a practical case, that is, when a model is used and/or developed to assist in making management or design decisions in a particular catchment, the catchments are usually much larger and one can often only rely on data from the standard network that are not nearly as detailed as those in research catchments. Often the standard data are of unknown and/or undocumented quality. This tends to make model calibration and validation in particular cases significantly more difficult and less accurate. In many practical cases there is then an issue of whether, with the given data, a spatially distributed model of catchment response can at all be considered to reliably portray catchment behaviour.

In practical applications, model users often use model codes they have not developed themselves and data that are provided by different agencies. It is sometimes not clear how reliable the code is and sometimes it is unclear to the user how the code exactly works. The lack of field experience also makes it more difficult to appreciate which processes operate in the catchment and what is the best model approach for representing them. The large scale of catchments often considered in practical applications of distributed models tends to cause scale problems similar to those discussed in Chapter 3 of this book. For example, it is not uncommon to use Richards' equation for elements that are as big as  $500 \times 500$  m – this is an area that is larger than the size of the whole catchment in many of the case studies in this book which have shown an enormous complexity that goes far beyond the processes represented by Richards' equation. The fact that model users, model builders, data providers, and clients are different groups and have differences in terminology, creates further problems. Currently there appears to exist no unique and generally accepted terminology on model validation in the hydrological community and the many different and contradictory definitions used tend to be confusing.

Finally, in practical applications, the purpose of the modelling is to make predictions rather than to gain insight into spatial catchment behaviour. What is considered a useful model for understanding catchment behaviour is not necessarily useful for practical applications. Many case studies in the chapters of this book used comparison of observed and simulated patterns not only to calibrate models and ensure that they are working for the right reasons, but also to identify from the pattern comparison, processes that the model cannot handle very well. These may be the subject of future research work. On the other hand, the situation is quite different in the practical case. What is needed in this case is a reliable model for the projected conditions of model application and the insights obtained are only important to the extent they can be used to improve model performance and/or interpretation of the results in terms of management or design decisions. It is important to realise that the type of model application i.e. investigative versus predictive, has profound implications on both model structure (predictive models often having a simpler structure) and model calibration/validation (predictive models often having a better defined range of applicability). The validation and calibration of distributed models in practical applications is therefore quite different from that in research type applications.

Unfortunately, in practical projects there is often not very thorough model testing due to data and resource constraints. It is not uncommon for predictions of spatial patterns to be made with models that have not been properly tested in terms of their spatial behaviour. For example, Kutchment et al. (1996) applied a distributed physically-based model to simulate the 3315 km<sup>2</sup> Ouse basin in the UK. They calibrated their model against runoff data only, but stated also that the model can give “hydrologically meaningful estimates of internal values”. Due to lack of data and lack of tests on internal variables, this statement appears as the authors’ own perception rather than a documented fact. This problem has been recognised by some authors who are a little more circumspect about the performance of distributed models. Jain et al. (1992) applied a distributed physically-based model to the 820 km<sup>2</sup> Kolar catchment in India, where the runoff data comprised the only available calibration and validation data. They concluded that “The resulting final model calibration is believed to give a reasonably good physical representation of the hydrological regime. However, a preliminary model set-up and calibration . . . resulted in an equally good hydrograph match, but on the basis of apparently less realistic soil and vegetation parameter values. Thus, it may be concluded that a good match between observed and simulated outlet hydrographs does not provide sufficient guarantee of a hydrologically realistic description.” However, the practical problems for which distributed models need to be applied remain, so the challenge is to better use the information available to us and to seek additional information, to help strengthen the confidence we can have in simulated responses from distributed models.

It is clear that proper validation and calibration of distributed models of catchment behaviour is of the utmost importance. This is obviously an uncertain endeavour. The primary role of model calibration and validation is to obtain a realistic assessment of this uncertainty – of what confidence we can place on the predictions of our model. In this chapter, these issues are addressed by proposing a framework for model validation and calibration. Also, issues of terminology will be clarified to develop a common language, and data issues relevant to distributed catchment models in practical applications will be discussed. The validation framework and data considerations will be illustrated in a case study for the 440 km<sup>2</sup> Karup catchment. The chapter concludes with a discussion on possible interactions between model builders, model users, and clients that could improve the understanding and treatment of uncertainty in practical applications of spatially distributed catchment models.

### 13.2 SOURCES OF SPATIAL DATA FOR PRACTICAL APPLICATIONS

Distributed hydrological models require spatial data. In this chapter, two different terms are used for the data, depending on its type: *parameter values* are those that do not vary with time while *variables* are time dependent. In a traditional model application the parameter values and the driving variables (typically climate data) are input data, while the other variables are simulated by the model.

An overview of typical data types and sources is given in Table 13.1 together with a characterisation of the typical availability of data from traditional sources and the potential for operational use of remote sensing data. Basically, traditional data sources provide point or vector data. Even for many of the exceptions to this, such as digital elevation maps or soil maps, the spatial data are inferred from originally measured point data. In general, it is possible to obtain such spatial data on catchment characteristics for use as model parameter input and by assuming relationships between these data and model parameters (e.g. between soil type and soil hydraulic properties, between vegetation types and water use etc.), model parameter values can be estimated (albeit with an accuracy determined by the validity of the assumed relationships – see discussion in Chapter 2, pp. 23–4, 41). It is also generally possible to get hydroclimatological time series for driving the model and for checking the overall catchment runoff. However, there is almost always a lack of data to check the detailed spatial patterns of internally simulated variables such as soil moisture, actual evapotranspiration and water depths. The only source for such data that can be characterised as realistic on scales above plots and small experimental catchments, is remote sensing data.

Remote sensing data have for a couple of decades been described as having a promising potential to supply spatial data to distributed hydrological models, e.g. Schulz (1988), Engman (1995) and De Troch et al. (1996). However, so far the success stories, at least in operational applications, are in practice limited to mapping of land use and snow cover, whereas scientific/technological breakthroughs are still lacking for assessing soil moisture, vegetation status, actual evapotranspiration and water depths. Chapters 5 and 6 in this book give examples of research applications where progress is clearly being made on representing variables such as soil moisture and saturated source areas, but as yet these methods are not available for practical application.

With progress made in recent years at the research level, we may foresee operational applications of remote sensing for practical modelling within the next decade in areas such as:

- Assessments of water depths and inundation areas at larger scales (> 1 km length) to be used for flood forecasting and flood mapping. This can today be done during cloud-free periods by use of thermal data, and appears promising in the future by use of SAR data. Furthermore, new high-resolution (few metres) visible satellite data are also promising.
- Assessment of vegetation and soil moisture status at field scale and above to be used for crop forecasting, irrigation management and meteorological forecasting.
- Assessment of vegetation status at field scale and below for supporting precision agriculture.
- Improved accuracy of RADAR derived precipitation.

A key point to remember about remote sensing data is that it is a surrogate measure – i.e. it depends on a relationship between properties of emitted or

**Table 13.1. Spatial data used in distributed catchment modelling**

<b>Data type</b>	<b>Function in model</b>	<b>Typical traditional data source</b>	<b>Typical availability of traditional data</b>	<b>Potential for operational use of remote sensing data</b>
Topography	Parameter – input data	Maps, DEMs	Very good	Interferometry
River network	Parameter – input data	Maps, derived from DEMs	Very good	
Geology	Parameter – input data	Geological surveys, maps	Good	
Soil	Parameter – input data	Maps, national databases	Good	
Land use/vegetation	Parameter – input data	Maps	Good	Well proven (Landsat, Spot, etc.)
Climate data (precipitation, temperature, wind speed, etc.)	Variable – input data	Meteorological databases	Usually exist, but some times difficult to get access to	Potential for rainfall data (Meteosat) Weather radar data in operational use, but quantitative accuracy so far not good
Snow cover	Variable – simulated		Very seldom	Well proven
Soil moisture	Variable – simulated	Ad hoc point measurements	Very seldom and only point values	Potential (microwave and SAR data), but so far no encouraging operational use
Vegetation status (leaf area, root depth)	Variable – input data or simulated		Very seldom and only point values	Potential, encouraging, but so far no operational use
Actual evapotranspiration/surface temperature	Variable – simulated		Very seldom and only point values	Potential
Water depth/inundation area	Variable – simulated	River gauging stations	Only at river gauging stations	Potential

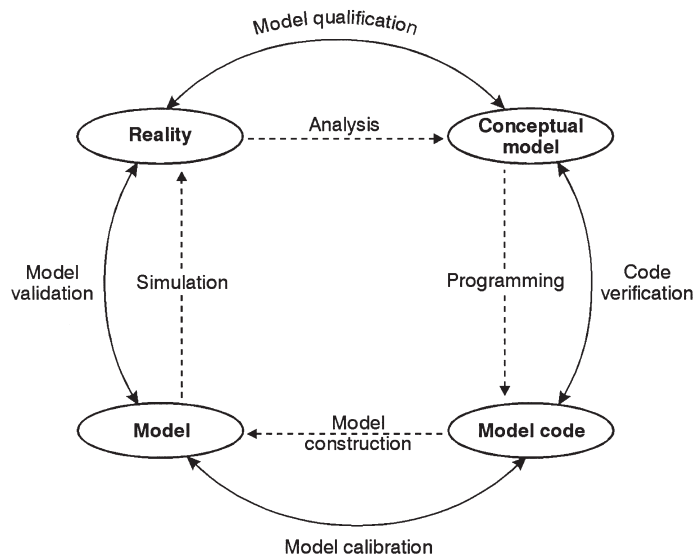
reflected radiation and a particular feature of interest such as soil moisture content. It is not a direct measure so, as with parameters like soil hydraulic properties estimated from soil type, the accuracy of derived measures is a function of the quality of the surrogate relationships (see Chapter 3, pp. 41–5).

### 13.3 ISSUES OF TERMINOLOGY

#### 13.3.1 Background

Before presenting a practical methodology for model calibration and validation, it is worth reflecting on the more fundamental question of whether models at all can be validated, and issues of terminology. Konikow and Bredehoeft (1992) argued that the terms validation and verification are misleading and their use should be abandoned: "... the terms validation and verification have little or no place in ground-water science; these terms lead to a false impression of model capability". The main argument in this respect relates to the anti-positivistic view that a theory (in this case a model) can never be proved to be generally valid, but can on the contrary be falsified by just one example. De Marsily et al. (1992) argued in a response to Konikow and Bredehoeft (1992) for a more pragmatic view: "... using the model in a predictive mode and comparing it with new data is not a futile exercise; it makes a lot of sense to us. It does not prove that the model will be correct for all circumstances, it only increases our confidence in its value. We do not want certainty; we will be satisfied with engineering confidence." Part of the difference of opinion relates to interpretations of the terminology used.

Konikow (1978) and Anderson and Woessner (1992) use the term verification with respect to the governing equations, the code or the model. According to Konikow (1978) a model is verified "if its accuracy and predictive capability have been proven to lie within acceptable limits of errors by tests independent of the calibration data". The term model verification is used by Tsang (1991) in the meaning of checking the model's capability to reproduce historical data. Anderson and Woessner define model validation as tests showing whether the model can predict the future. As opposed to the authors above, Flavelle (1992) distinguishes between verification (of computer code) and validation (of site-specific model). Oreskes et al. (1994), using a philosophical framework, state that verification and validation of numerical models of natural systems theoretically is impossible, because natural systems are never closed and because model results are always non-unique. Instead, in their view models can only be "confirmed". Within the hydraulic engineering community attempts have been made to establish a common methodology (IAHR, 1994). The IAHR methodology comprises guidelines for standard validation documents, where validation of a software package is considered in four steps (Dee, 1995; Los and Gerritsen, 1995): conceptual validation, algorithmic validation, software validation and functional validation. This approach concentrates on what other authors call code verification, while schemes for validation of site-specific models are not included.



**Figure 13.1.** Elements of a modelling terminology and their interrelationships. Modified after Schlesinger et al. (1979).

The terminology and methodology proposed below has evolved from a background of more than twenty years' experience with research, development and practical applications of hydrological models. The proposed terminology and methodology is aimed at being pragmatic and does not claim to be in full accordance with scientific philosophy. Thus, it operates with the terms verification and validation, which are being used on a routine basis in the hydrological community, although with many different meanings. On the other hand, the term model validation is not used carelessly here but within a rigorous framework where model validation refers to site specific applications and to pre-specified performance (accuracy) criteria. Thus, in agreement with past practical experience and in accordance with philosophical considerations, a model code is never considered generally valid.

### 13.3.2 Definition of Terminology

The following terminology is inspired by the generalised terminology for model credibility proposed by Schlesinger et al. (1979), but modified and extended to suit distributed hydrological modelling. The simulation environment is divided into four basic elements as shown in Figure 13.1. The inner arrows describe the processes which relate the elements to each other, and the outer arrows refer to the procedures which evaluate the credibility of these processes. The most important elements in the terminology and their interrelationships are defined as follows:

**Reality.** The natural system, understood here as the hydrological cycle or parts of it.

**Conceptual model.** A conceptual description of reality, i.e. the user's perception of the key hydrological processes in the catchment and the corresponding simplifications and numerical accuracy limits which are assumed acceptable in the model in order to achieve the purpose of the modelling.

**Model code.** Generic software program.

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

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

**Simulation.** Use of a validated model to gain insight into reality and obtain predictions that can be used by water managers.

**Model qualification.** An estimate of the adequacy of the conceptual model to carry out the desired application within the acceptable level of accuracy.

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

**Model calibration.** The procedure of adjustment of parameter values of a model to reproduce the response of a catchment under study 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.

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

In the above definitions the term conceptual model should not be confused with the word conceptual used in the traditional classification of hydrological models (lumped conceptual rainfall-runoff models).

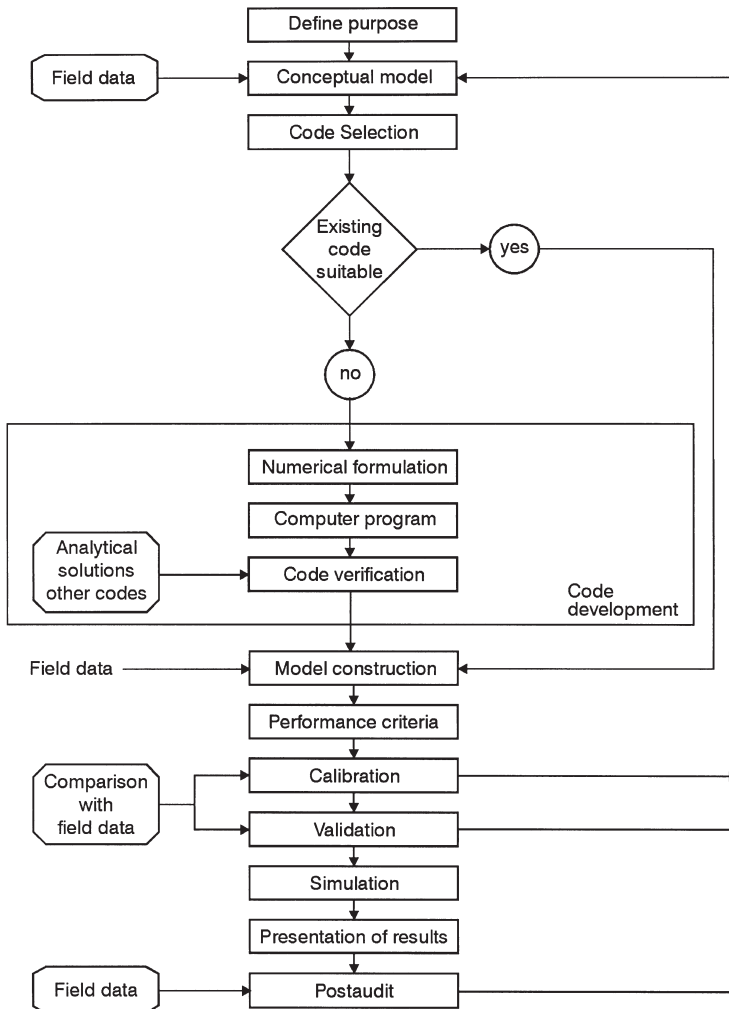
## 13.4 PROPOSED METHODOLOGY FOR MODEL CALIBRATION AND VALIDATION

### 13.4.1 Modelling Protocol

The protocol described below is a translation of the general terminology and methodology defined above into the field of distributed hydrological modelling. It is furthermore inspired by the modelling protocol suggested by Anderson and Woessner (1992), but modified concerning certain steps.

The protocol is illustrated in Figure 13.2 and described step by step in the following.

1. The first step in a modelling protocol is to *define the purpose* of the model application. An important element in this step is to give a first assessment of the desired accuracy of the model output.
2. Based on the purpose of the specific problem and an analysis of the available data, the user must establish a *conceptual model*.
3. After having defined the conceptual model, a suitable computer program has to be selected. In principle, the computer program can be prepared specifically for the particular purpose. In practice, a *code* is often *selected* among existing generic modelling systems. In this case it is important to ensure that the selected code has been successfully verified for the particular type of application in question.
4. In case no existing code is considered suitable for the given conceptual model a *code development* has to take place. In order to substantiate that the code solves the equations in the conceptual model within acceptable limits of accuracy a *code verification* is required. In practice, code verification involves comparison of the numerical solution generated by the code with one or more analytical solutions or with other numerical solutions.
5. After having selected the code and compiled the necessary data, a *model construction* has to be made. This involves designing the model with regard to the spatial and temporal discretisation, setting boundary and initial conditions and making a preliminary selection of parameter values from the field data. In the case of distributed modelling, the model construction generally involves reducing the number of parameters to calibrate (i.e. reducing the “degrees of freedom”, Chapter 3, pp. 75–6) e.g. by using representative parameter values for different soil types.
6. The next step is to define *performance criteria* that should be achieved during the subsequent calibration and validation steps. When establishing performance criteria, due consideration should be given to the accuracy desired for the specific problem (as assessed under step 1) and to the realistic limit of accuracy determined by the field situation and the available data (as assessed in connection with step 5). If unrealistically high performance criteria are specified, it will either be necessary to modify the criteria or to obtain more and possibly quite different field data.



**Figure 13.2.** The different steps in a hydrological model application – a modelling protocol. (From Refsgaard, 1997; reproduced with permission.)

7. *Model calibration* involves adjustment of parameter values of a specific model to reproduce the observed response of the catchment within the range of accuracy specified in the performance criteria. It is important in this connection to assess the uncertainty in the estimation of model parameters, for example from sensitivity analyses.
8. *Model validation* involves conduction of tests which document that the given site-specific model is capable of making sufficiently accurate predictions. This requires using the calibrated model, without changing the parameter values, to simulate the response for a period other than the calibration period. The model is said to be validated if its accuracy and predictive capability in the validation period have been proven to lie within acceptable limits or to provide acceptable errors (as specified in

the performance criteria). Validation schemes for different purposes are outlined below.

9. *Model simulation* for prediction purposes is often the explicit aim of the model application. In view of the uncertainties in parameter values and, possibly, in future catchment conditions, it is advisable to carry out a predictive sensitivity analysis to test the effects of these uncertainties on the predicted results (see Chapter 11 for one such procedure).
10. *Results* are usually *presented* in reports or electronically, e.g. in terms of animations. Furthermore, in certain cases, the final model is transferred to the end user for subsequent day-to-day operational use.
11. An extra possibility of validation of a site-specific model is a so-called *postaudit*. A postaudit is carried out several years after the modelling study is completed and the model predictions can be evaluated against new field data.

#### 13.4.2 Scheme for Construction of Systematic Model Validation Tests

Distributed hydrological models contain a large number of parameters, and it is nearly always possible to produce a combination of parameter values that provides a good agreement between measured and simulated output data for a short calibration period. However, as discussed in Chapter 3, this does not guarantee an adequate model structure nor the presence of optimal parameter values. The calibration may have been achieved purely by numerical curve fitting without considering whether the parameter values so obtained are physically reasonable. Further, it might be possible to achieve multiple calibrations or apparently equally satisfactory calibrations based on different combinations of parameter values (see also Chapter 11). Ideally, the ultimate purpose of calibration is not to fit the calibration data but to fit reality. If the other error sources, including the effects of non-perfect model structure and data uncertainties, are not somehow considered, there is a danger of overfitting.

In order to assess whether a calibrated model can be considered valid for subsequent use it must be tested (validated) against data different from those used for the calibration. According to the methodology established above, model validation implies substantiating that a site-specific model can produce simulation results within the range of accuracy specified in the performance criteria for the particular study. Hence, before carrying out the model calibration and the subsequent validation tests, quantitative performance criteria must be established. In determining the acceptable level of accuracy a trade-off will, either explicitly or implicitly, have to be made between costs, in terms of data collection and modelling work, and associated benefits of achieving more accurate model results. Consequently, the acceptable level of accuracy will vary from case to case, and should usually not be defined by the modellers, but by the water resources decision makers. In practice, however, the decision maker often only influences this important issue very indirectly by allocating

a project budget and requesting the modeller to do as good as possible within this given frame.

The scheme proposed below is based on Klemeš (1986b) who states that a model should be tested to show how well it can perform the kind of task for which it is specifically intended. The four types of test correspond to different situations with regard to whether data are available for calibration and whether the catchment conditions are stationary or the impact of some kind of intervention has to be simulated.

The *split-sample test* is the classical test, being applicable to cases where there is sufficient data for calibration and where the catchment conditions are stationary. The available data record is divided into two parts. A calibration is carried out on one part and then a validation on the other part. Both the calibration and validation exercises should give acceptable results. This approach was taken in Chapters 6, 7, 10 and 11.

The *proxy-basin test* should be applied when there is not sufficient data for a calibration of the catchment in question. If, for example, streamflow has to be predicted in an ungauged catchment Z, two gauged catchments X and Y within the region should be selected. The model should be calibrated on catchment X and validated on catchment Y and vice versa. Only if the two validation results are acceptable and similar can the model command a basic level of credibility with regard to its ability to simulate the streamflow in catchment Z adequately.

The *differential split-sample test* should be applied whenever a model is to be used to simulate flows, soil moisture patterns and other variables in a given gauged catchment under conditions different from those corresponding to the available data. The test may have several variants depending on the specific nature of the modelling study. If, for example, a simulation of the effects of a change in climate is intended, the test should have the following form. Two periods with different values of the climate variables of interest should be identified in the historical record, such as one with a high average precipitation, and the other with a low average precipitation. If the model is intended to simulate streamflow for a wet climate scenario, then it should be calibrated on a dry segment of the historical record and validated on a wet segment. Similar test variants can be defined for the prediction of changes in land use, effects of groundwater abstraction and other such changes. In general, the model should demonstrate an ability to perform through the required transition regime.

The *proxy-basin differential split-sample test* is the most difficult test for a hydrological model, because it deals with cases where there is no data available for calibration and where the model is directed to predicting non-stationary conditions. An example of a case that requires such a test is simulation of hydrological conditions for a future period with a change in climate and for a catchment where no calibration data presently exist. The test is a combination of the two previous tests.

Examples of the four tests are given by Refsgaard and Knudsen (1996), and Styczen (1995) provides an example of a test procedure based on the same prin-

ciples for the validation of pesticide leaching models for registration purposes. A general point related to all the tests, is that if the accuracy of the model for the validation period is much worse than for the calibration period, it is an indication of “overfitting”; that is, there is likely to be a problem with the model structure causing the parameters to be specific to the conditions used for calibration. The ratio of accuracy during the calibration period to accuracy during the validation period is sometimes used as a measure of the degree of overfitting.

It is noted that, according to this scheme, a distributed model cannot claim a predictive capability in simulation of spatial patterns unless it has been specifically tested for this purpose. Thus, if a model has only been validated against discharge data, which is very commonly the case, there is no documentation on its predictive capability with regard to, for example, simulation of spatial patterns of soil moisture and groundwater levels at grid scales. Claims on predictive capabilities with regard to, for example, soil moisture variation in time and space, require successful outputs from a validation test designed specifically with this aim. When designing validation tests the following additional principles must be taken into account:

- The *scale* of the measurements used must be appropriate for the scale of the model elements (see Chapter 2, pp. 19–20). The scales need not be identical, but the field data and the model results must be up/downscaled, so that they are directly comparable.
- The *performance criteria* must be specified, keeping the spatial patterns in mind.
- The *validation test* must be designed in accordance with a special emphasis on the spatial patterns and the distributed nature of the model.

For illustrative purposes, a hypothetical example is given in the following. Suppose that one purpose of a model application is to predict the patterns of soil moisture in the topsoil, and that the validation data consist of a remote sensing based SAR data set with a 30 m spatial resolution at four times during a given period. Suppose that the SAR data set has been successfully calibrated/validated against ground truth data and that the error can be described statistically (mean, standard deviation, spatial correlations). Suppose finally that the hydrological model uses a horizontal grid of 60 m and can match the vertical depth of the topsoil measured by the SAR. The above three principles could be implemented as follows:

- *Scaling*. The comparisons between model and data should be carried out at a minimum scale of 60 m. The data could also be aggregated to larger scales if that were sufficient for the subsequent model application. In any case the error description of the SAR data must be corrected to correspond to the selected scale, implying that the standard deviation of the error is reduced due to the aggregation process (see Chapter 2, p. 23).
- The *performance criteria* could, for example, be chosen to reflect various aspects such as:

- Capability to describe correct overall levels. This may include criteria on comparison of mean values and standard deviations of SAR data and model results.
- Capability to describe spatial patterns. This may include comparison of correlation lengths or division of the entire area into subareas and comparison of statistics within each subarea.
- Capability to describe temporal dynamics. This may include criteria on comparison of SAR and model time series for selected areas, such as regression coefficient and model efficiency (Nash and Sutcliffe, 1970).

In general, the numerical values of the performance criteria should depend on the uncertainty of the data, as described by the error statistics, and the purpose of the modelling study.

- The *type of validation test* should be decided from the same principles as outlined above. For instance, if SAR data were available for just part of the model area, a proxy-basin split sample test can be applied where model results, calibrated without SAR data, were compared in the overlapping area. The performance criteria would then be assumed to indicate how well the model will perform for the area not covered by SAR data.

### 13.4.3 Use of Spatial Data

Distributed hydrological models are structured to enable the spatial variations in catchment characteristics to be represented by having different parameter and variable values for each element. Often model applications require several thousands of elements, meaning that the number of parameters and variables could be two or three orders of magnitude higher than for a lumped model of the same area. Obviously, this generates different requirements for lumped and distributed models with regard to parameterisation, calibration and validation procedures.

A critique expressed against distributed models by several authors concerns the many parameter values which can be modified during the calibration process. Beven (1989, 1996) considers models which are usually claimed to be distributed physically-based as in fact being lumped conceptual models, just with many more parameters. Hence, according to Beven (1996) a key characteristic of the distributed model is that “the problem of overparameterisation is consequently greater”.

To address this problem in practical applications of distributed models, it is necessary to reduce the “degrees of freedom” by inferring spatial patterns of parameter values so that a given parameter only reflects the significant and systematic variation described in the available field data. This approach is exemplified by the practice of using representative parameter values for individual soil types, vegetation types or geological layers along with patterns of these types (see also the

discussion in Chapter 3, pp. 75–6 and examples of using this approach in Chapters 6, 9 and 10). This approach reduces the number of free parameter coefficients that need to be adjusted in the subsequent calibration procedure. The following points are important to consider when applying this approach (Refsgaard and Storm, 1996):

- The parameter classes (soil types, vegetation types, climatological zones, geological layers, etc.) should be selected so that it becomes easy, in an objective way, to associate parameter values. Thus the parameter values in the different classes should, to the highest possible degree, be assessable from available field data.
- It should explicitly be evaluated which parameters can be assessed from field data alone and which need some kind of calibration. For the parameters subject to calibration, physically acceptable intervals for the parameter values should be estimated.
- The number of real calibration parameters should be kept low, both from a practical and a methodological point of view. This can be done, for instance, by fixing a spatial pattern of a parameter but allowing its absolute value to be modified through calibration.

Reducing the number of free parameters in this way helps to avoid methodological problems in the subsequent phases of model calibration and validation. An important benefit of a small number of free parameters adjustable through calibration is that the whole parameter assessment procedure becomes more transparent and reproducible. The quality of the results, however, depends on the adequacy and accuracy of the imposed patterns. It is also important to consider to what extent the imposed pattern dominates the pattern of the simulated output (see Chapter 3, p. 76, and Figures 6.13, 9.12 and 10.15). Another example is the use of Thiessen polygons for representing spatial precipitation patterns where it is possible that soil moisture may be dominated by precipitation quantities making the simulated spatial patterns just reflect the Thiessen polygons. It may often be required to aggregate both model results and field data to a scale where the imposed pattern does not dominate.

Thus, the challenge for the hydrologist has been expanded beyond the task of tuning parameter values through calibration to the art of defining hydrologically sound methods for reducing the number of parameters to be calibrated.

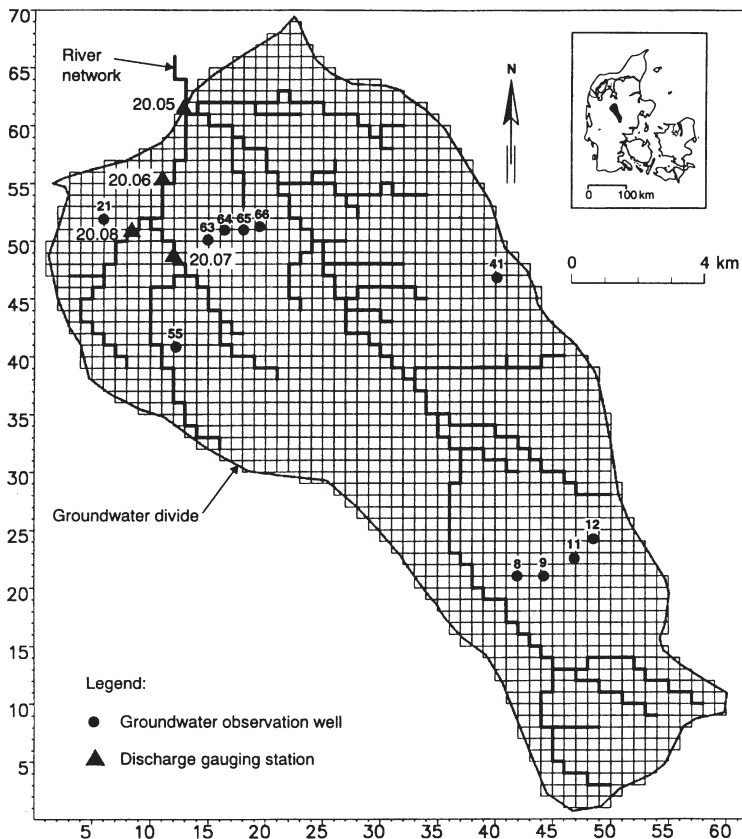
### 13.5 CASE STUDY

The above methodology is illustrated step by step in the study described in Refsgaard (1997). The first seven steps are summarised with rather brief descriptions in Section 13.5.2, while the model validation step is addressed more thoroughly in Section 13.5.3.

### 13.5.1 The Karup Catchment

The 440 km<sup>2</sup> Karup catchment is located in a typical outwash plain in the western part of Denmark. From a geological point of view the area is relatively homogeneous, consisting of highly permeable sand and gravel with occasional lenses of moraine clay. The depth of the unsaturated zone varies from 25 m at the eastern groundwater divide to less than 1 m in the wetland areas along the main river. The aquifer is mainly unconfined and of glacial deposits. The thickness of the aquifer varies from 10 m in the western and central parts to more than 90 m to the east. The catchment has a gentle sloping topography and is drained by the Karup River and about 20 tributaries.

The Karup catchment has been subject to several hydrological studies (e.g. Stendal, 1978; Miljøstyrelsen, 1983; Styczen and Storm, 1993) and a comprehensive database exists both for surface water and ground water variables. The catchment area and the measurement sites referred to in the following are shown in Figure 13.3.



**Figure 13.3.** The 440 km<sup>2</sup> Karup catchment with the river network in a 500 m model grid together with the location of the discharge gauging stations and groundwater observation wells referred to in the text. (From Refsgaard, 1997; reproduced with permission.)

### 13.5.2 Establishment of a Calibrated Model – the First Steps of the Modelling Protocol

#### *Step 1. Definition of Purpose*

The overall objectives of the case study are to illustrate the parameterisation, calibration and validation of a distributed model and to study the validation requirements with respect to simulation of internal variables and to changing spatial discretisation. In this context the purpose of the model is to simulate the overall hydrological regime in the Karup catchment, especially the spatial pattern of discharges and groundwater table dynamics.

#### *Step 2. Establishment of a Conceptual Model*

The assumptions made regarding the hydrological system are described in detail in Refsgaard (1997). The main components of the conceptual model can be characterised as follows:

- The groundwater system is described by an unconfined aquifer comprising one main aquifer material with the same hydraulic parameters throughout the catchment and five minor lenses with distinctly different parameters. The aquifer system is modelled as two-dimensional.
- The unsaturated zone is described by one-dimensional vertical flows. The soil system is via maps and profile descriptions described by two soil types characterised by different hydraulic parameters.
- Four vegetation/cropping classes are assumed: agriculture, forest, heath and wetland.
- The main river system and the tributaries which could be accommodated within the 500 m spatial model discretisation are included in the model. The wetland areas are assumed to be drained by ditches and tile drainpipes. The stream–aquifer interaction is assumed to be governed by the head differences in the river and the main aquifer and controlled by a thin, low permeability layer below the riverbed.
- Daily values, averaged over the catchment, of precipitation, potential evapotranspiration and temperature are used.

#### *Step 3. Selection of Model Code*

The code selected for the study was MIKE SHE (Refsgaard and Storm, 1995). In the present case the following modules were used: two-dimensional overland flow (kinematic wave), one-dimensional river flow (diffusive wave), one-dimensional unsaturated flow (Richards' equation), interception (Rutter concept), evapotranspiration (Kristensen and Jensen concept), snowmelt (degree-day concept) and two-dimensional saturated flows (Boussinesq).

#### *Step 4. Code Verification*

As MIKE SHE is a well proven code with several verification tests as well as many large scale engineering applications, including prior tests on the present

area and on similar cases, no additional code verification was required in this case.

#### ***Step 5. Model Construction***

The details regarding spatial discretisation of the catchment, setting of boundary and initial conditions and making a preliminary selection of parameter values from the field data are described in Refsgaard (1997). The number of parameters to be calibrated was reduced to 11 by, for example, subdividing the domains based on soil classes with uniform soil parameters in each subdivided area. Three of the parameters related to the aquifer properties and stream-aquifer interaction, while the eight remaining ones were soil hydraulic parameters. Thus the degrees of freedom in describing the spatial pattern of ground water levels in practice reduces to three parameters that can be fitted through calibration. One of the costs of such simplification is that one (spatially constant) value for aquifer hydraulic conductivity may not be sufficient to adequately describe the spatial patterns in groundwater flows and groundwater levels. A previous calibration of groundwater transmissivities for the same aquifer (Miljøstyrelsen, 1983) suggests that the transmissivities vary substantially more than can be explained by the variation in aquifer thickness in the present model.

#### ***Step 6. Performance Criteria***

The performance criteria were related to the following variables:

1. Discharge simulation at station 20.05 Hagebro (the outlet of the catchment) with a graphical assessment of observed and simulated hydrographs supported by the following two numerical measures:
  - average discharges of observed and simulated records,  $OBS_{ave}$  and  $SIM_{ave}$ , and
  - model efficiency,  $R^2$ , calculated on a daily basis (Nash and Sutcliffe, 1970).
2. Groundwater level simulations at observation wells 21, 41 and 55 located in the downstream part of the catchment and also used by Styczen and Storm (1993) plus observation wells 8, 9, 11, 12 representing a cross-section at the upstream part of the catchment.

These criteria were used for the calibration and the first part of the validation tests. For the second part of the validation tests, focussing on the capability to describe internal variables and spatial patterns, additional criteria were defined (see below).

#### ***Step 7. Model Calibration***

Most of the parameter values were assessed directly from field data or transferred from experience in other similar catchments. The remaining eleven parameter values were assessed during calibration through a trial-and-error process. The model calibration was carried out on the basis of data for the period 1971–

74. Altogether, the calibration results are of the same accuracy as the results in Styczen and Storm (1993), and are, as such, considered acceptable.

### 13.5.3 Model Validation

The validation tests have been carried out in two steps:

- **Step 8a.** Validation on the same station/wells as used for the calibration.
- **Step 8b.** Validation on additional data representing internal variables not utilised during the calibration process.

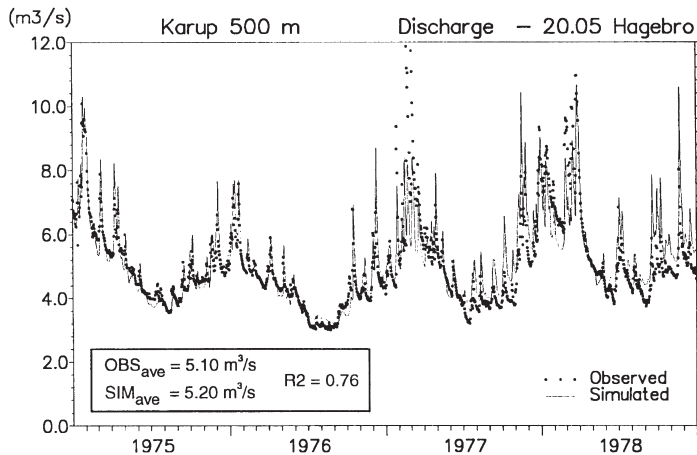
The validation tests were designed in accordance with the three guiding principles, outlined in Section 13.4.2. The *validation test type* is a traditional split-sample test for Step 8a and a proxy-basin split-sample test for Step 8b.

*Scaling.* The discharge data are aggregated values integrating the runoff over the respective catchments. This applies both to the field data and the model simulations, so no scaling inconsistency occurs here. The groundwater level data from observation wells are point data as opposed to the model simulations which represent average values over 500 m grids. The groundwater levels are known to vary typically by 1–2 m over a 500 m distance (Stendal, 1978). Hence, observed and simulated groundwater levels cannot be expected to match more closely than 1–2 m with regard to levels but somewhat better with regard to temporal dynamics.

The *performance criteria* take the various aspects into account as follows:

- The overall levels are expected to match within 10 % with regard to average discharge and within 2 m with regard to groundwater levels. The 2 m criteria should be seen in view of a typical variation in observed groundwater levels of  $1\frac{1}{2}$ –2 m within 500 m (Stendal, 1978).
- The temporal dynamics is expected to match reasonably well. No specific, numerical criteria have been identified for this purpose, but the visual inspection will focus on amplitude and phasing of the annual fluctuations.
- The capability to describe internal spatial patterns has been tested by using additional data for the following stations, for which data were not used at all during the calibration process:
  - discharge values at the three stations 20.06 Haderup (98 km<sup>2</sup>), 20.07 Stavlund (50 km<sup>2</sup>) and 20.08 Feldborg (17 km<sup>2</sup>) (Figure 13.3).
  - groundwater tables at observation wells 63, 64, 65 and 66, located in the area between the main river and the tributary with the three discharge stations 20.06, 20.07 and 20.08 (Figure 13.3).

Key results from this validation test are shown in Figures 13.4 and 13.5. These results from one discharge station and seven groundwater observation wells were comparable to the results from the calibration period and have been assessed as acceptable. Hence the model has now been successfully validated for simulation

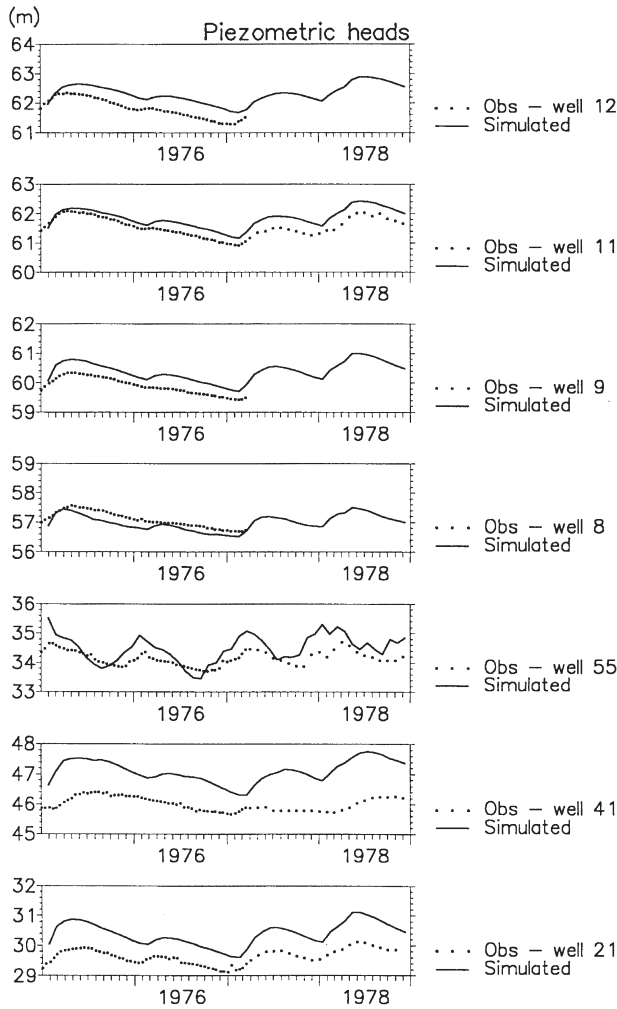


**Figure 13.4.** Simulated and observed discharge for the entire catchment for the validation period together with figures for average observed and simulated flows,  $OBS_{ave}$  and  $SIM_{ave}$ , and model efficiency on a daily basis,  $R^2$ . (From Refsgaard, 1997; reproduced with permission.)

of catchment discharge and groundwater levels in these seven observation wells with the expected accuracy similar to those shown in Figures 13.4 and 13.5.

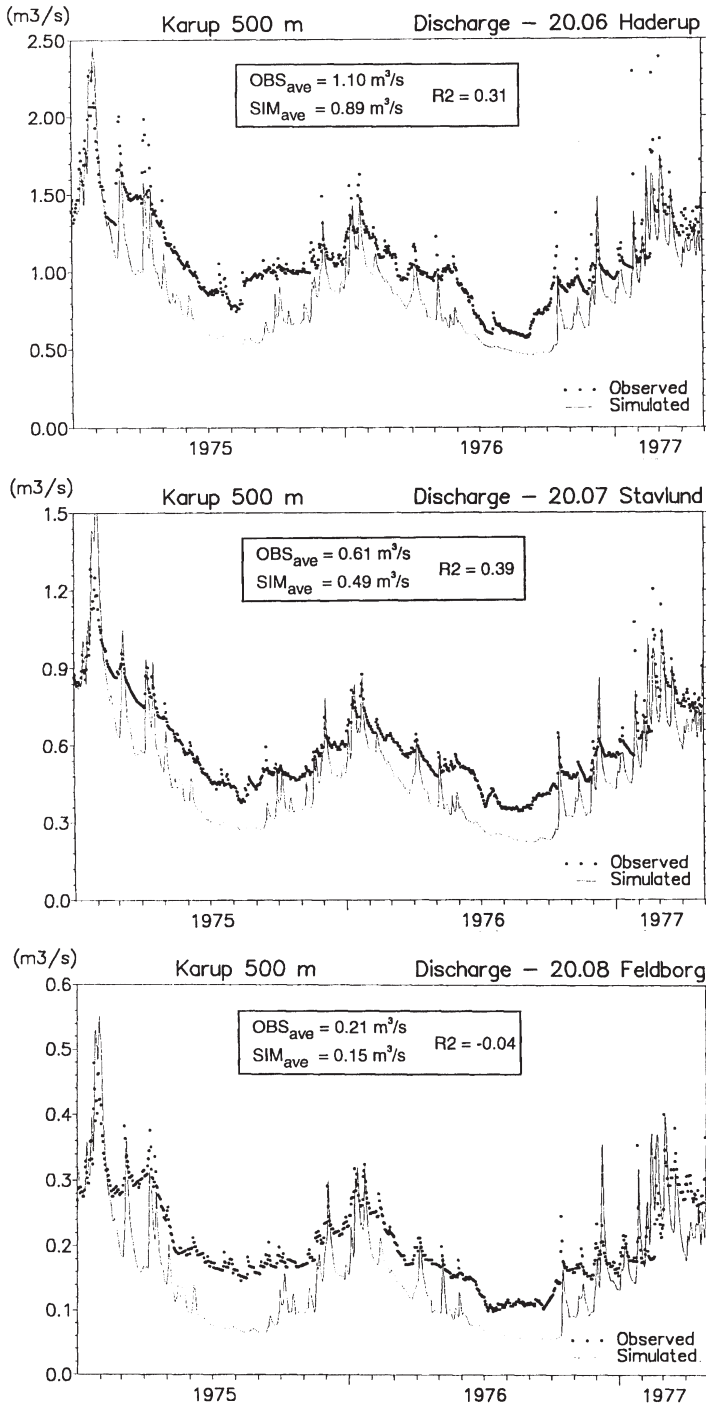
The interesting question now is how reliable is the model for simulation of internal variables and spatial patterns. This was addressed during step 8b. Results from the first 28 months of the validation period, where data are available for all the above stations, are seen in Figures 13.6 and 13.7 for discharge and groundwater tables, respectively. As can be seen from the hydrographs, the water balance, and the model efficiency, the simulation results are significantly less accurate than for the calibrated stations. The simulated discharges at the three tributary stations are significantly poorer than for the calibrated station 20.05 in two respects. Firstly, there is a clear underestimation of the baseflow level and the total runoff for the three tributary stations, where the 10% accuracy on the water balance performance criteria is not fulfilled for any of the three stations. Secondly, the simulation shows a significantly more flashy response than the observed hydrographs. The simulated groundwater tables (Figure 13.7) show correct dynamics, but have problems with the levels. The groundwater level error at well no. 64 is just above the 2 m specified as the accuracy level in the performance criteria on groundwater levels. Taking into account that the gradient between wells 64 and 65, which are located in two neighbouring grids, is wrong by about 3 m, the model simulation of groundwater levels can not be claimed to have passed the performance criteria in general. The primary reason for the differences in baseflow levels appear to be that the internal groundwater divide between the main river and the main tributary is not simulated correctly, with the result that the three tributary stations according to the model are draining smaller areas than they do in reality.

From the conducted validation tests, it may be concluded that the model cannot be claimed to be valid for discharge simulation of subcatchments, nor

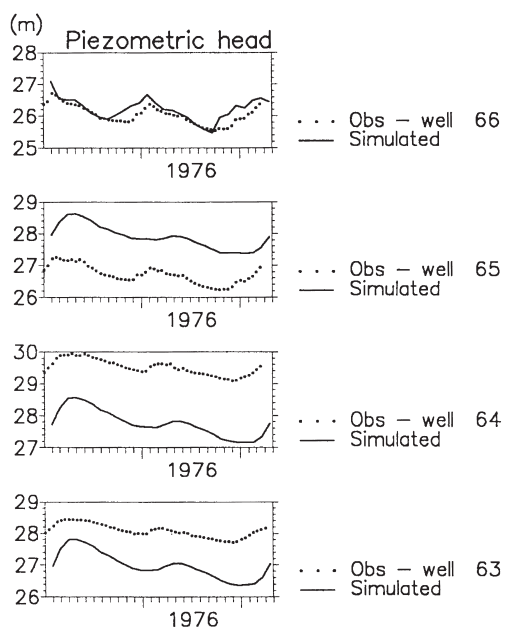


**Figure 13.5.** Simulated and observed piezometric heads at seven well sites for the validation period. The locations of the wells are shown in Figure 13.3. (From Refsgaard, 1997; reproduced with permission.)

for groundwater levels in general over the entire catchment area. Following the methodology represented in Figure 13.2, we should now re-look at the model conceptualisation and model construction, and perform additional calibration and tests to improve confidence. To do this we would need additional data. This could be more discharge data from other subcatchments, or additional data on groundwater levels to derive more detailed spatial patterns of ground water response. It would be hoped that using new and more detailed data will improve the model so that it meets the desired performance criteria. In the context of a practical application, more resources would need to be obtained to carry out these improvements. This is where the interaction with managers regarding acceptable accuracy and available budgets becomes critical.



**Figure 13.6.** Simulated and observed discharges, average flows,  $SIM_{ave}$  and  $OBS_{ave}$ , and model efficiencies,  $R^2$ , from the validation period for three internal discharge sites 20.06 (98 km<sup>2</sup>), 20.07 (50 km<sup>2</sup>) and 20.08 (17 km<sup>2</sup>), which have not been subject to calibration. The locations of the discharge stations are shown in Figure 13.3. (From Refsgaard, 1997; reproduced with permission.)



**Figure 13.7.** Simulated and observed piezometric heads from the validation period for four well sites for which no calibration has been made. The locations of the wells are shown in Figure 13.3. (From Refsgaard, 1997; reproduced with permission.)

### 13.6 DISCUSSION AND CONCLUSIONS

The need for model validation in distributed hydrological modelling was discussed in Chapter 3. Loague and Kyriakidis (1997) also concluded that the hydrologists need to establish rigorous model evaluation protocols. Gupta et al. (1998) argued that the whole nature of the calibration problem is multi-objective with a need to include not only streamflow but also other variables.

While some attention has been paid to systematic validation of lumped hydrological (rainfall-runoff) models (e.g. WMO, 1975, 1986, 1992), very limited emphasis has so far been put on the far more complicated task of validation of distributed hydrological models, where spatial variation of internal variables also has to be considered. Based on a review of some of the few studies focussing on validation of distributed models with respect to internal variables and multiple scales, the following conclusions can be drawn:

- Distributed models are usually calibrated and validated only against runoff data, while spatial data are seldom available.
- In the few cases, where model simulations have been compared with field data on internal variables, these test results are generally of less accuracy than the results of the validation tests against runoff data.
- Authors, who have not been able to test their models' capabilities to predict internal spatial variables, often state that their distributed models provide physically realistic descriptions of spatial patterns of internal variables.

In summary, it is possible to simulate spatial patterns at a quite detailed level and produce nice colourgraphics results; but due to lack of field data it is in general not possible to check to which extent these results are correct. This fact is pre-

sently one of the most severe constraints for the further development of distributed hydrological modelling. It is believed that although predictions of spatial patterns both by distributed models and by remote sensing are subject to considerable uncertainties, the possibilities of combining the two may prove to be of significant mutual benefit. There is an urgent need for more research on this interface. A recent example of such an exercise by Franks et al. (1998), who combined SAR based estimates of saturated areas with TOPMODEL simulations, shows encouraging results.

A particular area, where limited work has been carried out so far, is on the establishment of validation test schemes for the situations where the split-sample test is not sufficient. The only rigorous and comprehensive methodology reported in the literature is that of Klemeš (1986b). It may correctly be argued that the procedures outlined for the proxy-basin and the differential split-sample tests, where tests have to be carried out using data from similar catchments, from a purely theoretical point of view are weaker than the usual split-sample test, where data from the specific catchment are available. However, no obviously better testing schemes exist. Hence, this will have to be reflected in the performance criteria in terms of larger expected uncertainties in the predictions.

One of the important practical fields of application for distributed models is prediction of the effects of land use changes (Ewen and Parkin, 1996; Parkin et al., 1996). Many such studies have been reported; however, most of them can be characterised as hypothetical predictions, because the models have not been subject to adequate validation tests (Lørup et al., 1998). In this case it would be necessary to apply a differential split sample test but the data requirements are considerable and will be difficult to meet in practical applications without detailed information on patterns of hydrological response under different land use and climatic conditions.

It must be realised that the validation tests proposed in this chapter are so demanding that many applications today would fail to meet them. This does not imply that these modelling studies are not useful, only that their output should be realised to be somewhat more uncertain than is often stated and that they should not make use of the term 'validated model'.

Success criteria need to be clearly articulated for the model calibration and validation that focus on each model output for which it is intended to make predictions. Hence, multisite calibration/validation is needed if predictions of spatial patterns are required, and multi-variable checks are required if predictions of the behaviour of individual sub-systems within the catchments are needed. Thus, as shown also in the case study, a model should only be assumed valid with respect to outputs that have been explicitly validated. This means, for instance, that a model which is validated against catchment runoff cannot automatically be assumed valid also for simulation of erosion on a hillslope within the catchment, because smaller scale processes may dominate here; it will need validation against hillslope soil erosion data. Furthermore, it should be emphasised that with the present generation of distributed model codes, which do not contain adequate up- or down-scaling methodologies, separate calibra-

tion and validation tests have to be carried out every time the element size is changed.

As discussed above, the validation methodologies presently used, even in research projects, are generally not rigorous and far from satisfactory. At the same time models are being used in practice and daily claims are being made on validity of models and on the basis of, at the best, not very strict and rigorous test schemes. An important question then, is how can the situation be improved in the future? As emphasised by Forkel (1996), improvements cannot be achieved by the research community alone, but requires an interaction between the three main “players”, namely water resources managers, code developers and model users.

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

The model user has the responsibility for selection of a suitable code as well as for construction, calibration and validation of the site-specific model. In particular, the model user is responsible for preparing validation documents in such a way that the domain of applicability and the range of accuracy of the model are explicitly specified. Furthermore, the documentation of the modelling process should ideally be done in enough detail that it can be repeated several years later, if required. The model user has to interact with the water resources manager on assessments of realistic model accuracies. Furthermore, the model user must be aware of the capabilities and limitations of the selected code and interact with the code developer with regard to reporting of user experience such as shortcomings in documentation, errors in code, market demands for extensions, etc.

The key responsibilities of the developer of the model code are to develop and verify the code. In this connection it is important that the capabilities and limitations of the code appear from the documentation. As code development is a continuous process, code maintenance and regular updating with new versions, improved as a response to user reactions, become important. Although a model code should be comprehensively documented, doubts will, in practice, always occur once in a while on its functioning, even for experienced users. Hence, active support to and dialogue with model users are crucial for ensuring operational model applications at a high professional level.

Although the different players have different roles and functions, a special responsibility lies with the research community. Unless we take a lead in improving the situation within our own community, the overall credibility of hydro-

logical modelling is at risk. Thus a major challenge for the coming decade is to further develop suitable rigorous validation schemes and impose them to all hydrological modelling projects. Part of this challenge lies in the collection and use of spatial patterns of key model inputs, parameters and outputs so that the calibration and validation exercises can fully quantify the model capabilities.