Chapter 2
Hydroinformatics and Data-Based Modelling Issues in Hydrology

Abstract This chapter highlights and addresses some basic issues associated with data-based modeling. The chapter starts with a brief description of emergence and development of hydroinformatics as a potent segment of mainstream hydrology and proceeds to the ignored or least considered modeling queries existing in hydrology, e.g., how much benefit could be gained by increased complexity in data-based models or whether increased complexity adversely affects model performance. The chapter reminds one of the need to evaluate existing hypothetic assumptions on various modeling properties.

Data-based modeling is subject to different types of uncertainties and ambiguities because of the presence of different unsolved queries and deliberately over-simplified assumptions. Several studies in hydrology have pointed out the contradictory fact that, under certain circumstances, a poor model may give acceptable results, while, under other circumstances, a good, refined model may fail to give better and more reliable answers. The main reason for this is that developers view modeling as a rigorous mathematical exercise rather than as a subjective activity [56]. Previously successful model in one phase of the hydrological cycle might not give relevant results in a new situation. Proper vision or insight into the working of the actual hydrological system is necessary, even in data-based modeling; else modeling results from even a sophisticated mathematical model would be irrelevant or misguiding with regard to the behavior of the actual system. In data-based modeling, the model trusts the quality of the data which should be inherent in the actual behavior of the system. Savenije [60] emphasized the need to change the modeling process to a "top-down" approach, i.e., learning from the data to the physical theory rather than giving a lower preference to the strength of the data. Barnes [11] suggests that an "adequate" model is a model which represents all the information contained in the data (so that there is effectively no residual information). As per this definition, all data-based and artificial intelligence models fall into that category. Another required aspect of a model is how efficient the model is in tackling a particular phenomenon or situation in hydrology. The problems of overparameterization and scaling issues have gained much attention and invited detailed studies in the context of processes modeling on a wide catchment and regional scale.

R. Remesan and J. Mathew, *Hydrological Data Driven Modelling*,
Earth Systems Data and Models 1, DOI 10.1007/978-3-319-09235-5_2

However, the modeling issues in data-based models are not addressed properly in hydrology. It is a fact that there is no such thing as a 'perfect data-based model' in any field. As hydrologists, our aim is to approach to more realistic and "better" model framework through (1) selecting an "appropriate" model structure and (2) selecting "correct" inputs through avoiding "redundant" inputs. This chapter addresses a number of issues in traditional and artificial intelligence data-based techniques adopted for hydrological modeling.

2.1 Hydroinformatics

Hydroinformatics is one interdisciplinary field of technology which embraces the application of information technology in different aspects of the water sector, focusing on integrating information technology with hydrologic, hydraulic, and environmental science and engineering. The scientist who coined the term 'Hydroinformatics,' Professor Abbott [1], defines the term as *the study of the flow of information and the generation of knowledge related to the dynamics of water in the real world, through the integration of information and communication technologies for data acquisition, modeling and decision support, and to the consequences for the aquatic environment and society and for the management of water based systems.* It includes many state-of-the-art applications of modern information technologies in water management and decision making. Hydroinformatics focuses on:

- New themes such as computational intelligence, control systems, and their application in data-driven hydrological modeling
- Optimization and real-time control of models
- Flood modeling for management of module integrating modeling theory, hydraulics, and flood simulation
- Decision support systems module integrating system analysis, decision support system theory, and model integration

From 1993, hydroinformatics has started developing as a strong stream in hydrology after the introduction of a Section on Hydroinformatics by the International Association for Hydraulic Research (IAHR). In 1998 the International Water Association (IWA) then established a Specialist Group on Hydroinformatics [65]. IAHR-IWA-IAHS Joint Committee on Hydroinformatics was formed in 1998 by The International Association of Hydrological Sciences (IAHS) in collaboration with IAHR and IWA. Recently, in 2005, the European Geosciences Union (EGU) established hydroinformatics as a subdivision of hydrological sciences. The first international hydroinformatics conference was held at IHE, Delft in 1994, and, thereafter, bi-annual conferences were conducted all over the world in places such as Zürich, Copenhagen, Iowa City, Cardiff, Singapore, Nice, and Chile, the latest in 2014 being held in New York, USA. The theme of the 11th conference is "Informatics and the Environment: Data and Model Integration in a Heterogeneous Hydro World". Over the last 20 years the hydroinformatics stream has shown its

capabilities through success stories published in a wide range of articles. Abbott and Vojinovic [2] introduced a new role for hydroinformatics in its sociotechnical environment, developing the concept introduced by Abbott [3]. However, researchers working in hydroinformatics are still struggling to get full-scale acceptance within the hydrological community, which is dominated by larger groups of traditionalists who care less about data and more about physics. Some argue against this section of hydrology, saying it adds no scientific knowledge or improved understanding to the field of physical modeling of hydrology. However, many studies have clearly shown the capabilities of hidden nodes of artificial neural networks to communicate the real physics involved in the process [37, 81]. The capabilities of new concepts such as Genetic Programming are worth mentioning on this occasion, having great potential to provide us with new hydrological knowledge [24]. Some traditional hydrologists argue over generally adopted thumb rules and assumptions during training and modeling, an obstacle to the wider acceptance of this new stream. Although hydroinformatics and data-driven modeling have been in use for more than two decades, it is struggling to find full acceptance within the hydrological community, which is dominated by large groups of traditional hydrologists because of inherent problems in these models (e.g., chances of overfitting, redundancy of input, lack of modeling rigor, lack of transparency in reproducing results, uncertainty issues, etc.). Some studies [25, 49, 71] suggested better modeling frameworks and guidelines in data-based modeling. Some of the modeling shortcomings and ambiguity in such data-based models are discussed below. Elshorbagy et al. [24] argue that most data-based studies are a 'less-than-comprehensive approach' focusing on (1) one or two data sets or application models [6] and (2) random realization of the three subsets for modeling; which makes the generalization ability of that model questionable. Elshorbagy et al. [22], See and Openshaw [64] and Abrahart et al. [6] have reminded the hydroinformatics research community of the need to maintain scientific rigor in the application and use of data-driven techniques in hydrology and environmental sciences. The fundamental means to assess the capability of any novel approach or modeling technique is to evaluate it against other modeling techniques or approaches under different modeling conditions or data sets. Elshorbagy et al. [23] has noted that most modeling comparative studies in the literature of data-based modeling hydrology are highly impaired due to the less-than-comprehensive approaches adopted. Single realization of the data set and single case study makes it difficult to assess the actual capability of the novel concept such as the Gamma Test. All new techniques should be evaluated against available basic models (linear regression) and complex models (SVMs or wavelet SVMs).

2.2 Why Overfitting and How to Avoid

Overfitting or overtraining is a statistical phenomenon associated with nonlinear data-based models when a model is generally complex with too many degrees of freedom in relation to the amount of data available. The predictive models used in

this book, such as artificial neural networks (ANNs), and other flexible nonlinear estimation methods, such as kernel regression models (SVM) and smoothing splines, are susceptible to either overfitting or underfitting. Underfitting is mainly because of design or training incompetence of the modeler. A network which is not sufficiently complex to handle nonlinear data processing may fail to detect the full characteristics of the signal which leads to underfitting. If networks are too complex, it would lead to a dangerous situation called overfitting, which would give better predictions in the training data and poor predictive results to future values. The complexity normally connected with the complexity of a network is related to both the size of the weights and the number of hidden units and layers. Apart from that, model input selection and training data length influence the overfitting of nonlinear models. Overtraining can be detected during training by the use of a test set. However, the disadvantage of this split technique is that the size of the training set reduces considerably in limited data cases, and thereby spoils the final performance. Another easily adoptable approach is to rotate parts of the available data sets as the training set and the test set. In some cases, strong nonlinearity of the problem may lead to overfitting, which is easily noticeable from the size of the weights. Figure 2.1 presents an example of variation of model performance under scenarios such as overfitting, underfitting, and a reasonable model during calibration (training) and testing (validation) phases.

However, there are some standard techniques to tackle overfitting to some extent, which are briefly described here.

1. *Proper selection of model input structure and training data length*: Tackled in this book and discussed through case studies and in the next chapter.
2. *Jittering*: Somewhat similar to data enrichment. In this method, an artificial noise deliberately added to the inputs during training. A good example in

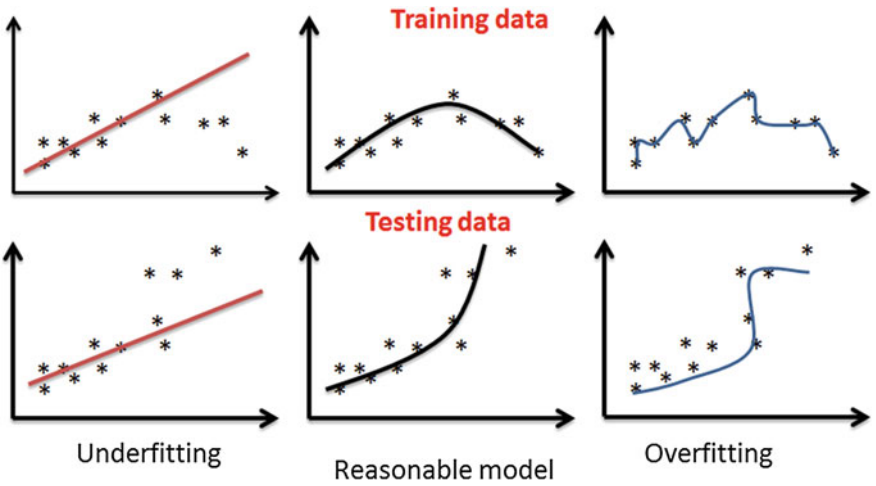

Fig. 2.1 Illustration of performance of an overfitted model, underfitted model and a reasonable model

hydrology is Olden and Poff [54]. They have applied Jittering to tackle redundancy in hydrologic indices of long-term flow records from 420 sites from across the continental USA. Jittering is also related to regularization methods such as weight decay and ridge regression.

3. *Early stopping*: In this approach, we need to stop training processes just before an adaptation to the noise starts. The optimal stopping time can be found using test data. In other words, the modeler requires three subsets of data (training, test, and verification). At much later stages of the modeling process the prediction accuracy of the model may start worsening for the test set. This is the stage when the model should cease to be trained to overcome the over-fitting problem.
4. *Weight decay*: Weight-decay reduces the effect of noise associated with the inputs.
5. *Bayesian learning*: The conventional training statistical approaches are replaced by Bayesian statistics.
   This approach involves modification of general objective functions, such as the mean sum of squared network errors (MSE or E_m):

$$F = E_m = MSE_m = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 \tag{2.1}$$

$$F = \beta E_m + \alpha E_d \tag{2.2}$$

   The modification of MSE is to improve the generalization capability to avoid overfitting. The above equation will be modified to a new F value by adding a new E_d term. In (2.2) the parameters β and α are to be optimized by a Bayesian framework. It is usually assumed that the weights and biases of the network are random variables following Gaussian distributions, with enormous computations required.
6. *Use of small networks*: If parameters in the training network are less than the objects in the training set, it cannot be overtrained unless the tackled case is too complex.
7. *Pruning*: The method for reducing the size of a network just after the training process. This approach helps to detect redundant neurons which cause delay in modeling. In the case of pruning network modeling, the training process starts with a large, densely connected network, and then examines the trained network's performance to assess the relative importance of network weights. After that, the pruning algorithm removes the least important weight/node from the network and performs analysis on the new pruned network. This procedure continues till the modeler is happy with the results.
8. *Data enrichment*: An approach for artificially enlarging the training set by artificial data. This process is not found to be effective in all cases.
9. *Regularization*: In this method, we add a penalty term to the optimization criterion for networks with large weights, as these are related to strong nonlinearity.

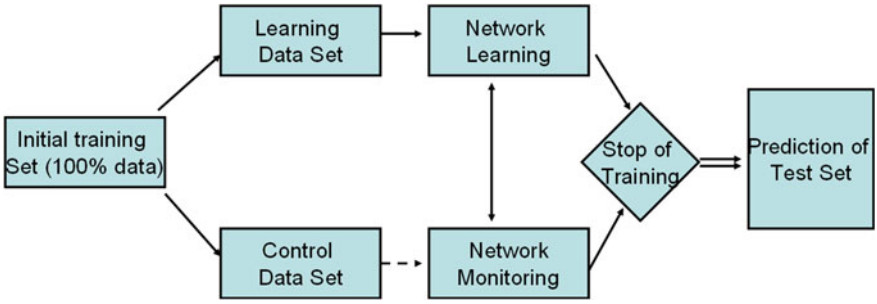

Fig. 2.2 Partition procedure adopted conventionally in learning to prevent overfitting

The traditional data partitioning method adopted to prevent overfitting in most of the literature is shown in Fig. 2.2

2.3 Input Variable (Data) Selection

One of the serious problems encountered in data-based modeling in hydrology using either traditional or intelligent approaches is the choice of independent variables or data series from the available data pool for inclusion in the predictive model. Although overfitting issues in the previous section are addressed in the context of neural models, studies have shown that overfitting is not confined just to neural models with hidden units. Overfitting can occur even in generalized linear models with no hidden nodes or layers because of improper selection of inputs. "Multicollinearity" is another weakness in data-based modeling. It is a statistical situation from the presence of input variables or data series in the input architecture, which are highly correlated with each other. Although this is the situation in most of the studies in water resources using data-based models, very little attention or no attention is being given to the selection process of better input model structure [49]. The lack of methodological approach in selecting the significant inputs may lead to the modeling issues listed below:

1. Increase in input dimensionality: when we use all available inputs indiscriminately, this would cause computational complexity and memory insufficiency
2. Presence of more local minima in the error surface due to inclusion of irrelevant data points
3. Early convergence and divergence due to the presence of irrelevant data: this will lead to poor model accuracy.

Maier and Dandy [49] have highlighted the fact that issues relating to the optimal division of the available data, data pre-processing, and the choice of appropriate model inputs are seldom considered in data-based modeling and

artificial intelligence models for application in hydrology. They have reviewed more than 43 journal papers in hydrology and pointed out that in most cases the inputs were chosen arbitrarily without any scientific reasoning and some studies used a trial and error approach or validation data. A study by Bowden et al. [19] gives an extensive review of the background and methodology adopted in input determination for neural network models in water resource applications. They have classified major attempts in input data selection in water resources into five categories:

1. *Relying on prior knowledge of the system*
   ASCE Task Committee on Application of Artificial Neural Networks in Hydrology [9] has pointed out the predominant use of a priori knowledge of the system as an indicator of model selection in hydrology. Studies such as Campolo et al. [20] and Jayawardena et al. [38] have used modelers' expert knowledge on the system and the study condition to select the influencing inputs. Some studies have benefited from the combination of a priori knowledge and analytical approaches [47, 48]. Factors such as large dependency on an expert's knowledge, very subjective nature, and case dependency are considered as disadvantages of such methods.
2. *Based on linear cross-correlation*
   Cross-correlation methods are the most common and popular analytical techniques for selecting appropriate inputs [35, 36, 67]. In this approach research normally depends on linear cross-correlation analysis values to determine the strength of the relationship between the input time series and the output time series at various lags [31]. The disadvantage associated with this method is its inability to capture any nonlinear dependence that may exist between the inputs and the output. The cross-correlation method works on linear dependence between two variables, so there is a good chance of the omission of important inputs that are related to the output in a nonlinear fashion.
3. *Based on heuristic approach*
   In this approach, various models are trained using different subsets of inputs. In this method, some researchers often employ stepwise selection of inputs such as forward selection and backward elimination to avoid total enumeration [75]. The concepts of these two approaches are self-explanatory from the name itself. Forward selection is the most common approach, in which we try to find the best single input and select it for the final model [48]. Backward elimination works in the opposite way; it starts modeling with a set of all inputs, and sequentially removes the input set which reduces performance least. Most of the heuristic approaches are computationally intensive trial and error procedures and there is no guarantee that they will find the globally best subsets.
4. *Methods that extract knowledge contained within trained ANNs*
   In this type of method, researchers mostly depend on sensitivity analyses to extract information from a trained ANN [45, 61]. Abrahart et al. [5] used a novel concept known as saliency analysis to disaggregate a neural network solution in

terms of its forecasting inputs. The saliency analysis was achieved by setting one input data stream at a time to zero and then performing the modeling, replacing the input data stream after the computation, then repeating this process on the next data set, and so on. This approach determines the relative importance of each input by examining the change in forecasting error and the plots from the flood hydrograph. Abrahart et al. [5] claimed superiority of the saliency approach over sensitivity analysis, as sensitivity analysis does not investigate the rate of change of one data variable with respect to the change in another. Bowden et al. [19] suggests the disadvantages of their approach are (1) lack of retraining the ANN after removing each input and (2) the possibility of producing nonsensical outputs due to the presence of zero inputs.

5. *Methods that use various combinations of the above four approaches*
   Some studies have used effective combinations of the above-mentioned methods in data selection [4, 61, 67]. Abrahart et al. [4] used a genetic algorithm (GA)-based approach to optimize the inputs to an ANN model used to model runoff. Approaches such as Pearson correlation, stepwise forward regression analysis, and sensitivity analysis were used by Schleiter [61] to select appropriate inputs for water quality modeling.

The above-mentioned approaches are widely used, even in multiple linear regression models, although many disadvantages are associated with them [46]. Another possible approach associated with models with nodes is the method cited in the previous section, the "pruning approach." There are very powerful pruning algorithms available which are used effectively for input variable selection in other fields of engineering [74]. The study by Livingstone et al. [46] has pointed out the relevance of selection of effective modeling of the responsive variables (data series) to the success of nonlinear models, which is dictated by the data.

Bowden et al. [19] proposed state-of-the-art methods such as the Partial mutual information algorithm (applied for calculation of dependence in the case of multiple inputs), the Self-organizing map (SOM) (used to reduce the dimensionality of the input space and obtain independent inputs), the GA, and the General regression neural network (GRNA) (applied to determine which inputs have a significant relationship with the output (dependent) variable) for input selection of ANN models.

2.4 Redundancy in Input Data and Model

Hydrologists often face challenges in identifying redundant input data during the preprocessing period as the sets of possible inputs into a hydrological system are huge. This process becomes more challenging in the modeling of some hydrological processes, as all measurable variables are highly nonlinear in dynamics and have multiple interrelations. The normal practice for data-based model practitioners

is to present a large number of inputs to the model and rely on the network to identify the critical model inputs. Usually, not all of the available data pool will be equally supportive for effective modeling, since some may be redundant with very close correlation with another; some may have predominant noise over the information or may not have any appreciable relationship with the target variable of the expected study. So such a practice normally causes adverse effects on modeling results. Another serious issue is that the redundancy of a network is related to both the number of weights and the size of the weights. Selection of appropriate and effective models connects with the number of weights, and hence the number of hidden units and layers. Until now there has been no exact solution for questions such as how many hidden layers and how many hidden nodes there should be in node-based modeling [51, 72]. The selection of hidden neurons is the tricky part in ANN modeling, as it relates to the complexity of the system being modeled and is usually set by the user. There should be an effective way to decide on the number of nodes, considering many factors such as number of input and output units, number of training data points, amount of noise in the targets, complexity of the function or classification or learning algorithm, topology of the model, type of hidden unit activation function, and regularization.

There are many practical rules of thumb that are available in the literature to facilitate a decision on the number of hidden nodes. Blum [17] reports that the number of nodes in the hidden layer—somewhere between the input layer nodes and the output layer node size—is appropriate for modeling. Hecht-Nielsen [32] proposes that the maximum number of elements in the hidden layer be twice the input layer dimension plus one. Another study by Maren et al. [50] recommends using the number of nodes equal to the geometric average between the input and output node dimension. Mechaqrane and Zouak [52] have used a feed-forward network with the size of the hidden layer equal to the size of the input layer. Some companies working on commercial neural network software development adopt a rule of thumb of the sum of input and output nodes multiplied by 2/3 as the indicator to choose the number of hidden neurons. Swingler [73] suggests that, for networks with one hidden layer, the model give better performance if we use twice the number of input nodes in the hidden layer. At the same time, Berry and Linoff [13] note that the number of hidden nodes should never be more than double the nodes in the input layer. Boger and Guterman [18] used principle component analysis to find the number of hidden nodes and they suggested using the same number of components which express 70–90 % of the variance of the input data. However, our experience shows that, in general, these rules give only some indication for the hidden layer dimension and none are properly right or wrong.

The above-mentioned Hecht-Nielsen suggestion has more scientific authenticity as the method is based on the Kolmogorov theorem in node based computation. There were strong arguments against this recommendation by many researches [33, 34], saying that it is sufficient to use a single hidden layer when using regular transfer functions (e.g., sigmoidal) but the number of required hidden nodes can be as high as the number of training samples, and justifying their arguments through

valid proofs. Huang [33] made some recommendations for the two-hidden-layer case. He suggested the number of hidden nodes sufficient to train N samples with a reasonable minimum error is

$$N_{\text{hid}} = 2\sqrt{(M+2)}N \tag{2.3}$$

The sufficient number of hidden nodes in the first layer is

$$N1_{\text{hid}} = 2\sqrt{(M+2)}N + 2\sqrt{\frac{N}{(M+2)}} \tag{2.4}$$

The sufficient number of hidden nodes in the second layer is

$$N2_{\text{hid}} = M\sqrt{\frac{N}{(M+2)}} \tag{2.5}$$

In all these equations, M = output neurons and N = training data points.

Stathakis [72] suggests the most accurate structure will have fewer nodes than the one suggested by Huang [33] and this high structure leads to redundancy in structure and over-fitting of the training data. Stathakis [72] proposed a near-optimal solution approach with a GA to find a better topology.

Traditionally, identification of topology has been based on trial and error, on heuristic approaches, on heuristics sometimes followed by trial and error, and on pruning or constructive methods.

Trial and error: This is the most traditional and primitive way of assessment. It may yield severely suboptimal structures, especially when adopted by inexperienced users.

Heuristic methods: Several approaches are found in the literature [8, 59, 79] which are all based on the objective to devise a formula which estimates the number of nodes in the hidden layers as a function of the number of input and output nodes. However, most of the heuristics lack the theoretical evidence to support the discovery of an optimal structure, so they are commonly used in subsequent search by trial and error.

Exhaustive search: This is one of the perfect but impracticable approaches in real life applications, as the number of search alternatives is exceedingly large, computationally intensive, and with longer computation time. Yao [82] illustrates the difficulties in exhaustive searching to find hidden neurons; he identified that the major complication is due to the noisy fitness evaluation problem.

Pruning and constructive algorithms: These are developed with the objective of devising an effective network topology by incrementally adding or removing links (weights) to the redundant or simple structures, respectively. Optimal Brain Damage [44] and Optimal Brain Surgeon [29] are two commonly used algorithms.

2.5 Data-Based Modeling—Complexity, Uncertainty, and Sensitivity

Two major modeling themes focusing on modeling errors are upward or mechanistic approaches (associated issues are overparameterization, equifinality) and downward or data-driven approaches (associated issue is lack of a priori defined model structure) with different complexities [78]. Are more complex models better? Should the increasing complexity of the existing model add any benefit to the model users? These issues are not properly addressed in hydrology and data-based modeling although in abundance in the many competing artificial intelligence models in the literature. These questions can be answered by tackling the complexity of a model's structure and the uncertainty associated with its output. It is often difficult in hydrology to decide which model should be used for a particular purpose, and the decision is often made on the basis of familiarity rather than the appropriateness and effectiveness of the model. Another major concern is overparameterization of the model to represent an uncertain process over limited and noisy data. Comparing different models just in terms of their better accuracy in predicting numerical values is often ludicrous; there are many other aspects which need to be taken into account before declaring one model a success with entirely different mathematical concepts over the other. The best model is not necessarily the most complex, or the one which overtly reflects the most sophisticated understanding of the system [11]. There is a hypothesis that more complex models simulate the processes better but with high variability in sensitivity and relatively less error [68]. However, a study by Oreskes et al. [55] argues that there is no strong evidence that simple models are more likely to produce more accurate results than complex ones. Case studies in this book use a simple index of utility which evaluates in terms of model complexity (we used training time as the indicator of complexity), model sensitivity (response to changes in input), and model error (closeness of simulation to measurement). Perrin et al. [57] performed an extensive comparative performance assessment of the structures of 19 daily lumped models, carried out on 429 catchments, and suggested that the main reason why complex models lack stability is that the structure, i.e., the way components are organized, is not suited to extracting information available in hydrological time series. Complex models in their study face considerable difficulties in parameter estimation and structure validation. Gregory et al. [27] have applied Akaike's information criterion (AIC) [7] and Bayes information criterion (BIC) (Schwartz [63] model selection model complexity problems in rainfall time series modeling, and similar approaches were applied in groundwater modeling [42]. Cherkassky and Mulier [21] have developed structural risk minimization (SRM) as an alternative model complexity control method. Schoups et al. [62] used the above-mentioned three models. We compare three model complexity control methods for hydrologic prediction. Information theory could be selected as the central framework to evaluate information content in

training data and associated predictions. Weijs et al. [80] proposed that hydrological system should be based on information-theoretical scores.

2.5.1 Modeling Uncertainty

In hydrology and water resources research, there are two major bases of uncertainty attitudes; one is based on stochasticity as a necessary factor and the other on the deterministic nature of the system. The definition of the uncertainty is much more uncertain about the modeled numerical values; it relates much deeper processes and pertains to the governing mechanisms of the model. Distinguishable uncertainties in hydrology are data uncertainties (mainly associated with measurements), sample uncertainties (e.g., number of data for calibration), and model uncertainty [58]. Klir [41] made an attempt to consider uncertainty in terms of the complexity of the model. He found both categories have a conflictive nature, i.e., if complexity decreases, uncertainty grows. Halfon [28] also addressed the issue of modeling in the context of Lake Ecosystem models; he evaluated the performance of several models of varying complexity. There are many ways to assess roughly the modeling uncertainty and these methods range from the use of statistical parameters such as standard deviation to analytical calculations to find the propagation of error. Recently, very powerful tools such as Monte Carlo analysis have been used for sensitivity estimation of more complicated methods. Other criteria such as fractals, Bayesian fuzzy-sets, and random fields have been applied successfully to solve uncertainty problems in hydrology and other fields such as applied mathematics, physics, systems theory, etc. Wagener et al. [77] applied a Monte Carlo analysis toolbox, combining a number of analysis tools to investigate parameter identifiability, model behavior, and prediction uncertainty to establish a sensible relationship between model parameters and catchment characteristics. Beck [12] points out valid reasons to concentrate more on the uncertainty of model structure as an important area of study. Mizumura [53] combined a conceptual tank model and a fuzzy logic model to yield satisfactory results with minimum uncertainty issues. Kindler and Tyszewski [39] acknowledged the applicability of fuzzy theory to a diagnostic approach of problem solving and uncertainty assessment. Feluch [26] applied non-parametric estimation methods to two classes of hydrological problems. Various studies have been carried out to ascertain the ‘best and right’ model in environmental chemistry using modern statistical approaches [43, 76], considering uncertainty. Beven [15, 16] introduced the concept of equifinality which is related to the uncertainty associated with parameters. Equifinality arises when, in a hydrological model, many different parameter sets are equally good at reproducing an output signal.

2.5.2 Model Complexity

In the last 20 years, the study of complexity in modeling systems has emerged as a recognized field in statistics. However, the initial attempts to formalize the concept of complexity go back even further, to Shannon's inception of Information theory [66]. The complexity of a model is closely related to the uncertainty of the system, which can be defined in terms of model properties such as model sensitivity and modeling error. The general hypothesis of model complexity and its influence during training and testing phases is shown in Fig. 2.3. The general hypothesis states that more complex models can simulate reality better than simpler models (i.e., less prediction error), and with a greater variance and low bias during training phase. Less complex models provide a relatively approximate simulation (i.e., with more prediction error), but with less variance and higher bias. However, the case is somewhat different in the testing phase; highly complex models won't give the best test results as the graph is parabolic, with a minimum somewhere in the middle.

Figure 2.4 displays the hypothesis which shows the variation of different model parameters, particularly with bias-variance interaction during the test phase.

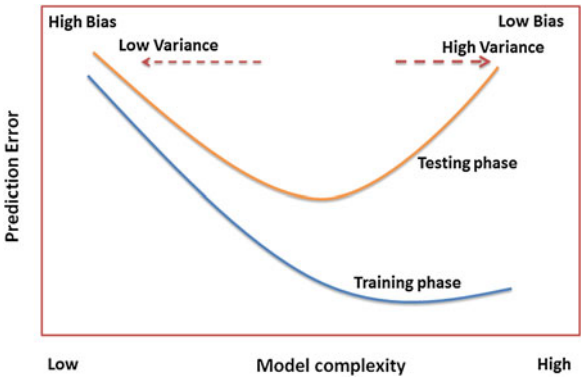

Fig. 2.3 Hypothesis showing the effect of complexity during training and testing [30]

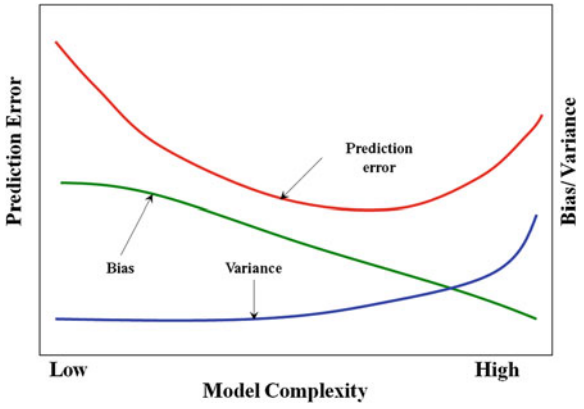

Fig. 2.4 Hypothesis showing effect of complexity on bias-variance interaction

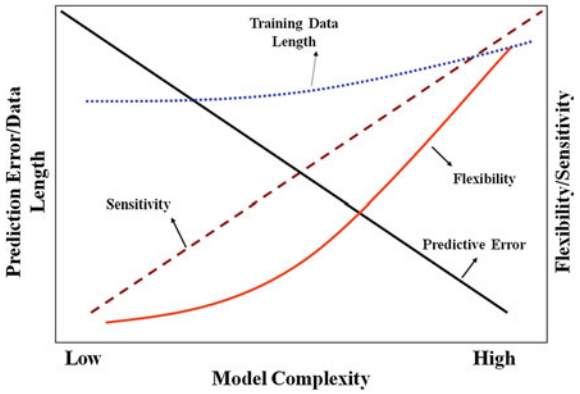

Fig. 2.5 Hypothetical relation of model complexity with sensitivity, flexibility, data requirements and predictive error

Models of different complexity may show different modeling properties, such as sensitivity, flexibility, error, and data requirements based upon their structure. Figure 2.5 illustrates the hypothetical relationship between model sensitivity, modeling error, model flexibility, training data requirement, and model complexity.

2.5.3 Training Data Requirements

More complex models may have more parameters, state variables, or linkages, and therefore the hypothesis is that such models require more data. However, for node-based modeling systems, if the training data length causes overfitting then the hypothetical relation depicted above may not be true. There is a need to investigate the optimal length of the training data phase as too little data for training may lead to poorly trained model and too much data may lead to overfitting. The hypothetical relation of model data requirements for node based modeling systems is as shown in Fig. 2.6.

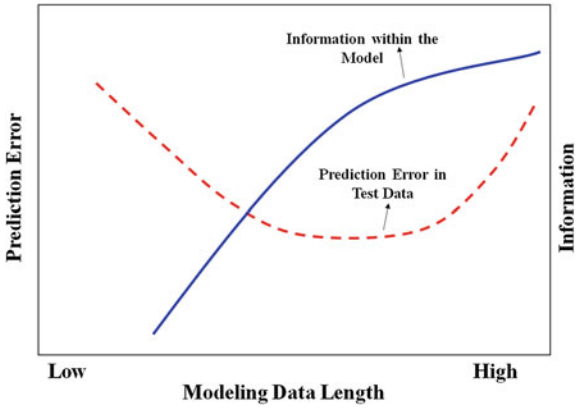

Fig. 2.6 Hypothetical relation of modeling data length in node based models

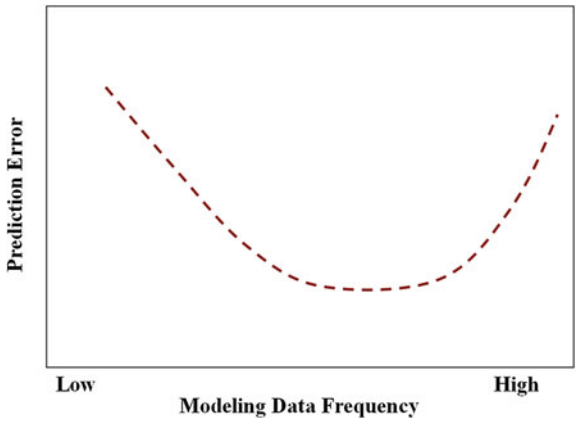

Fig. 2.7 Hypothetical relation of data interval on modeling error

Data interval for modeling: The case studies in this book also address a mostly ignored area in data-based hydrological modeling—data time interval for models. Modern data collection and telecommunication technologies can provide us with very high resolution data with extremely fine sampling intervals. We hypothesized that both too large and too small time intervals were detrimental to a model's performance, which has been illustrated in this book with particular reference to rainfall-runoff modeling. The data time interval is a major factor affecting forecast performances of node-based models, particularly neural network models. The performance of neural network models is highly time-dependent [10]. Very large and small data time intervals could have negative effects on modeling results. The hypothetical condition for the effect of data time interval on modeling is shown in Fig. 2.7.

2.5.4 Flexibility for a Model

The flexibility of a model increases as the number of parameters goes up. However, the modeler should be careful in increasing the flexibility of a model by addition of extra parameters, and in most cases it may cause irrationality. Flexibility of a model is dependent on the assumptions and rules employed during its development. The hypothesis is that less complex models are less flexible because of adoption of tough restricting assumptions to reduce the parameters. In general, complex models have minimal assumptions, and thus they are more flexible and applicable to a wide range of scenarios. Because of this flexibility factor, applications of less complex models are limited only to situations where the assumptions are valid. The flexibility of a model is invariably accompanied by extensibility of the mathematical code which is determined by the level of complexity of the model. In data-based modeling, the modeler should always be aware of five basic aspects before modeling: (1) has the selected model sufficient rigor to represent the process; (2) adequacy of the selected model for simulation of hydrological processes; (3) flexibility

of the model; (4) model design and optimization method; and (5) computational capabilities and complexities of the code.

2.5.5 Sensitivity of a Model

Sensitivity of a model is a major factor which decides its reliability in real situations. It often refers to the amount of change in model output resulting from a change in model input. As this book deals with data-based training models, the variations in modeling results are assessed with a certain percentage change in each input data series. Actually, this process tests the robustness of model results of a model under uncertain inputs. However, for physical models in general, the sensitivity of a model refers to the changes in its individual parameters. The overall sensitivity would be a cumulative result of the effects of all parameters in the system model. The general hypothesis is that sensitivity increases with increasing complexity because of the presence of more parameters or links. However, variation of sensitivity is dependent on many factors, so this hypothesis is much generalized. To resolve uncertainty-sensitivity issues, different kinds of optimization algorithms have been developed, namely the variance-based Sobol' method [69, 70] and the GLUE procedure [14]. Sensitivity analyses are valuable tools for identifying important model parameters to test model conceptualization and model structure.

2.5.6 Predictive Error of a Model

Prediction error is a generalized indicator of the performance of a model. The true predictive error is the sum of training error and training optimism. It is often referred to as the quality of the output, and the way the model performance should be interpreted and assessed. Training optimism is a measure of how bad our model can be over unseen data in comparison to training data. The more optimistic we are, the better our training error will be compared to what the true error is, and the worse our training error will be as an approximation of the true error. The hypothesis is that highly complex models simulate the real systems and give least prediction error. The inadequacies of simple models in most cases are because of the presence of simplifying assumptions.

2.5.7 Identifiability of a Model

Identifiability is a measure of how well the system is defined by the model, which is not directly assessable. This quantity tells the model whether the model 'over-defines' the system, which normally happens when the degree of freedom of the model is

higher than that of the real system. One can find discussions of hydrological model identifiability from the late 1980s [12, 40] in the hydrological literature.

Defining modeling uncertainty as a function of the model properties above (particularly model sensitivity and modeling error), it is important to investigate the relationship between modeling uncertainty and model complexity.

2.6 Index of Model Utility (U)

This book adopts an index of model utility to make a decision about which is the 'best and right' model for any hydrological modeling exercise. The adopted approach is a somewhat modified version of Snowling and Kramer [68] for suitability in data-based modeling. Statistically, the proposed 'index of model utility' of a model can be defined as a scaled distance from the origin on a graph of sensitivity versus modeling error of different models to the point corresponding to that model in the graph. Mathematically it can be written as

$$U_i = 1 - \sqrt{\frac{K_s S_i^2 + K_e E_i^2}{(K_s + K_e)}} \tag{2.6}$$

where

U_i is the utility index for model I,

S_i is the sensitivity value for model i (relative to the maximum sensitivity), in this study the value obtained from the mean value of slope of all sensitivity curves obtained from all inputs,

E_i is the error value for model i (relative to the maximum error; in this study we have adopted root mean squared error as the indicator of model error), and

K_s and K_e are weighting constants for sensitivity and error, respectively.

The value of U varies between 0 and 1 and if the value of U is larger, the model has higher utility. The values of S and E for each model should be normalized to satisfy the equation, which is the reason for dividing all values by the maximum sensitivity and error value. The values of K_s and K_e depend on how the model values error and sensitivity. If error and sensitivity are valued equally, then K_s and K_e should both be set to 1. In this study, both values were set to 1. In this book, the model utility indexes (U) were calculated for the three case studies and are illustrated in Chaps. 5–7. The purpose of this equation is to explore the usefulness of several statistical models, considering their complexity and sensitivity in hydrologic prediction in a simpler way. Further research can be accomplished by considering varying proportions of K_s and K_e values.

2.7 Conclusions

This chapter summarizes some of the data modeling issues where one can find major over-simplified assumptions and unsolved issues. It covers the relatively simple and neglected topics of training data length, data redundancy, and assumptions in neuron selection in ANN modeling.

References

1. Abbott MB (1991) Hydroinformatics: information technology and the aquatic environment. Ashgate, Aldershot
2. Abbott MB, Vojinovic Z (2013) Towards a hydroinformatics praxis in the service of social justice. J Hydroinform (in press) doi:10.2166/hydro.2013
3. Abbott MB (1996) The sociotechnical dimensions of hydroinformatics. In: Proceedings of the second international conference on hydroinformatics, Balkema, Rotterdam, pp 3–18
4. Abrahart RJ, See L, Kneale PE (1999) Using pruning algorithms and genetic algorithms to optimise network architectures and forecasting inputs in a neural network rainfall-runoff model. J Hydroinform 1(2):103–114
5. Abrahart RJ, See L, Kneale PE et al (2001) Investigating the role of saliency analysis with a neural network rainfall-runoff model. Comput Geosci 27:921–928
6. Abrahart R, See L, Dawson C (2008) Neural network hydroinformatics: maintaining scientific rigour. In: Abrahart R, See L, Solomatine D (eds) Practical hydroinformatics. Computational intelligence and technological develop-ments in water applications. Springer-Verlag, Berlin, Heidelberg, Germany, pp 33−47
7. Akaike H (1970) Statistical predictor identification. Ann Inst Statist Math 22:203−217
8. Anellopoulos I, Wilkinson G (1997) Strategies and best practice for neural networkimage classification. Int J Remote Sens 18:711–725
9. ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000) Artificial neural networks in hydrology-I: preliminary concepts. J Hydraul Eng ASCE 5(2): 115–123
10. Avci E (2007) Forecasting daily and sessional returns of the ISE-100 index with neural network models. Doğuş Üniversitesi Dergisi 8(2):128–142
11. Barnes CJ (1995) Commentary: the art of catchment modelling: what is a good model? Environ Int 21(5):747–751
12. Beck M (1987) Water quality modelling: a review of the analysis of uncertainty. Water Resour Res 23:1393–1442
13. Berry MJ, Linoff G (1997) Data mining techniques: for marketing, sales, and customer support. Wiley, New York
14. Beven K, Binley A (1992) The future of distributed models—model calibration and uncertainty prediction. Hydrol Process 6(3):279–298
15. Beven KJ (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. Adv Water Resour 16:41−51
16. Beven KJ (2001) How far can we go in distributed hydrological modelling? Hydrol Earth Syst Sci 5(1):1−12
17. Blum A (1992) Neural networks in C ++. Wiley, NY, p 60
18. Boger Z, Guterman H (1997) Knowledge extraction from artificial neural network models. In: IEEE Systems, Man, and Cybernetics Conference, Orlando, FL

19. Bowden GJ, Graeme C, Dandyb Holger R, Maier et al (2005) Input determination for neural network models in water resources applications. Part 1—Background and methodology. J Hydrol 301:75–92
20. Campolo M, Soldati A, Andreussi P et al (1999) Forecasting river flow rate during low-flow periods using neural networks. Water Resour Res 35(11):3547–3552
21. Cherkassky V, Mulier F (2007) Learning from data, 2nd edn. John Wiley & Sons, New York
22. Elshorbagy A, Corz G, Srinivasulu S, Solomatine DP (2009) Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology—Part 1: concepts and methodology. Hydrol Earth Syst Sci Discuss 6:7055–7093
23. Elshorbagy A, Corz G, Srinivasulu S, Solomatine DP (2009) Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology—Part 2: application. Hydrol Earth Syst Sci Discuss 6:7095–7142
24. Elshorbagy A, Corzo G, Srinivasulu S, Solomatine DP (2010) Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology—Part 2: application. Hydrol Earth Syst Sci 14:1943−1961. doi:10.5194/hess-14-1943-2010
25. Elshorbagy A, Parasuraman K (2008) Toward bridging the gap between data-driven and mechanistic models: cluster-based neural networks for hydrologic processes. In: Abrahart R, See L, Solomatine D (eds) Practical hydroinformatics. Computational intelligence and technological developments in water applications. Springer, Berlin, pp 389–403
26. Feluch W (1995) Nonparametric estimation of multivariate density and nonparametric regression. In: Kundzewicz ZW (ed) New uncertainty concepts in hydrology and water resources. Cambridge University Press, New York
27. Gregory JM, Wigley TML, Jones PD (1992) Determining and interpreting the order of a two-state Markov chain: Application to models of daily precipitation. Water Resour Res 28:1443−1446
28. Halfon E (1983) Is there a best model structure? 1. Modelling the fate of a toxic substance in a lake. Ecol Modelling 20:135–152
29. Hassibi B, Stork DG (1993) Second order derivatives for network pruning: optimal brain surgeon. In: Hanson SJ, Cowan JD, Giles CL (eds) Advances in neural information processing systems 5:164−171, San Mateo, CA, Morgan Kaufmann
30. Hastie T, Tibshirani R, Friedman J et al (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
31. Haugh LD, Box GEP (1977) Identification of dynamic regression (distributed lag) models connecting two time series. J Am Statist Assoc 72(397):121–130
32. Hecht-Nielsen R (1990) Neurocomputing. Addison-Wesley, Reading
33. Huang GB (2003) Learning capability and storage capacity of two-hidden-layer feedforward networks. IEEE Trans Neural Networks 14(2):274–281
34. Huang GB, Babri HA (1997) General approximation theorem on feed forward networks. In: IEEE Proceedings of International Conference on Information, Communications and Signal Processing, pp 698–702
35. Huang W, Foo S (2002) Neural network modelling of salinity variation in Apalachicola River. Water Res 36: 356−362
36. Imrie CE, Durucan S, Korre A et al (2000) River flow prediction using artificial neural networks: generalisation beyond the calibration range. J Hydrol 233:138–153
37. Jain A, Sudheer KP, Srinivasulu S et al (2004) Identification of physical processes inherent in artificial neural network rainfall runoff models. Hydrol Processes 18:571–581
38. Jayawardena AW, Fernando DAK, Zhou MC et al. (1997) Comparison of multilayer perceptron and radial basis function networks as tools for flood forecasting IAHS Publication (International Association of Hydrological Sciences), p 239
39. Kindler J, Tyszewski S (1995) On the value of fuzzy concepts in hydrology and water resources management. In: Kundzewicz ZW (ed) New uncertainty concepts in hydrology and water resources. Cambridge University Press, New York
40. Kleissen F, Beck M, Wheater H (1990) The identifiability of conceptual hydrochemical models. Water Resour Res 26(12):2979–2992

41. Klir GJ (1989) Methodological principles of uncertainty in inductive modelling: a new perspective. In: Erickson GJ, Smith CR (eds) Maximum-entropy and Bayesian methods in science and engineering, vol 1. Kluwer Academic Publishers, Dordrecht
42. Knotters M, De Gooijer JG (1999) TARSO modeling of water table depths. Water Resour Res 35. doi:10.1029/1998WR900049
43. Kohler M, Curtis GP, Kent DB, Davis JA et al (1992) Experimental investigation and modelling of uranium(VI) transport under variable chemical conditions. Water Resour Res 32:3539–3551
44. Le Cun Y, Denker JS, Solla S, Howard RE, Jackel LD (1990) Optimal brain damage. In: Touretzky D (ed) Neural Information Processing Systems vol 2. Denver, Morgan Kaufman
45. Liong SY, Lim WH, Paudyal GN et al (2000) River stage forecasting in Bangladesh: neural network approach. J Comput Civil Eng 14(1):1–8
46. Livingstone DJ, Manallack DT, Tetko IV et al (1997) Data modelling with neural networks: advantages and limitations. J Comput Aid Mol Des 11:135–142
47. Maier HR, Dandy GC (1997) Determining inputs for neural network models of multivariate time series. Microcomput Civil Eng 12(5):353–368
48. Maier HR, Dandy GC, Burch MD et al (1998) Use of artificial neural networks for modelling cyanobacteria Anabaena spp. in the River Murray, South Australia. Ecol Model 105:257–272
49. Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. Environ Model Softw 15:101−124
50. Maren AJ, Harston CT, Pap RM et al (1990) Handbook of neural computing applications. Academic Press, San Diego
51. Mas JF, Flores JJ (2008) The application of artificial neural networks to the analysis of remotely sensed data. Int J Remote Sens 29:617–663
52. Mechaqrane A, Zouak M (2004) A comparison of linear and neural network ARX models applied to a prediction of the indoor temperature of a building. Neural Comput Applic 13:32–37
53. Mizumura K (1995) Application of fuzzy theory to snowmelt-runoff. In: Kundzewicz ZW (ed) New uncertainty concepts in hydrology and water resources. Cambridge University Press, New York
54. Olden JD, Poff NL (2003) Redundancy and the choice of hydrologic indices for characterizing stream flow regimes. River Res Applic 19:101–121
55. Oreskes N, Shrader-Frechette K, Belitz K et al (1994) Verification, validation, and confirmation of numerical models in the earth sciences. Science 26:641–646
56. Penrose R (1988) The Emperor's new mind. Oxford University Press, New York
57. Perrin C, Michel C, Andreassian V (2001) Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. J Hydrol 242:275−301
58. Plate EJ, Duckstein L (1987) Reliability in hydraulic design. In: Duckstein L, Plate EJ (eds) Engineering reliability and risk in water resources. NATO ASI Series, Series E: Applied Sci., No. 124. Nijhoff, Dordrecht
59. Ripley BD (1993) Statistical aspects of neural networks. In: Barndorff-Nielsen OE, Jensen JL, Kendall WS (eds) Networks and chaos—statistical and probabilistic aspects. Chapman & Hall, London, pp 40–123
60. Savenije HHG (2009) HESS Opinions. The art of hydrology. Hydrol Earth Syst Sci 13:157–161. doi:10.5194/hess-13-157-2009
61. Schleiter IM (1999) Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. Ecol Model 120(2–3):271–286
62. Schoups G, van de Giesen NC, Savenije HHG et al. (2008) Model complexity control for hydrologic prediction. Water Resour Res 44 W00B03, doi:10.1029/2008WR006836
63. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 62(2):461–464
64. See L, Openshaw S (2000) Applying soft computing approaches to river level forecasting. Hydrol Sci J 44(5):763–779

65. See L, Solomatine D, Abrahart R, Toth E et al (2007) Hydroinformatics: computational intelligence and technological developments in water science applications—editorial. Hydrol Sci J 52(3):391–396. doi:10.1623/hysj.52.3.391
66. Shannon CE (1949) The mathematical theory of communication. University of Illinois Press, Urbana
67. Silverman D, Dracup JA (2000) Artificial neural networks and long-range precipitation in California. J Appl Meteorol 31(1):57–66
68. Snowling SD, Kramer JR (2001) Evaluating modelling uncertainty for model selection. Ecol Model 138:17–30
69. Sobol' IM (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simulat 55(13):271–280
70. Sobol' IM (1990) On sensitivity estimation for nonlinear mathematical models. Matematicheskoe Modelirovanie 2:112–118
71. Solomatine DP, Ostfeld A (2008) Data-driven modelling: some past experiences and new approaches. J Hydroinform 10(1):3–22
72. Stathakis D (2009) How many hidden layers and nodes? Int J Remote Sens 30(8):2133–2147
73. Swingler K (1996) Applying neural networks: a practical guide. Academic Press, London
74. Tetko IV, Villa AEP, Livingstone DJJ et al (1996) Chem Inf Comput Sci 36:794
75. Tokar AS, Johnson PA (1999) Rainfall-runoff modeling using artificial neural networks. J Hydrol Eng 4(3):232–239
76. Usunoff E, Carrera J, Mousavi SF et al (1992) An approach to the design of experiments for discriminating among alternative conceptual models. Adv Water Resour 15:199–214
77. Wagener T, Lees M, Wheater HS et al (2001) A toolkit for the development and application of parsimonious hydrological models. In: Singh VP, Frevert DK (eds) Mathematical models of watershed hydrology. Water Resources Publications, Highlands Ranch, pp 91–140
78. Wagener T, Sivapalan M, Troch PA, Woods R (2007) Catchment classification and hydrologic similarity. Geogr Compass 1:901–931. doi:10.1111/j.1749-8198.2007.00039.x
79. Wang(1994) The use of artificial neural networks in a geographical information system for agricultural land-suitability assessment. Environ Plann A 26:265–284
80. Weijs SV, Schoups G, van de Giesen N et al (2010) Why hydrological predictions should be evaluated using information theory. Hydrol Earth Syst Sci 14:2545–2558. doi:10.5194/hess-14-2545-2010
81. Wilby RL, Abrahart RJ, Dawson W et al (2003) Detection of conceptual model rainfall—runoff processes inside an artificial neural network. Hydrolog Sci J 48(2):163–181. doi:10.1623/hysj.48.2.163.44699
82. Yao X (1993) Evolutionary artificial neural networks. Int J Neural Sys 4:203−222

Springer

http://www.springer.com/978-3-319-09234-8

Hydrological Data Driven Modelling
A Case Study Approach
Remesan, R.; Mathew, J.
2015, XV, 250 p. 172 illus., 59 illus. in color., Hardcover
ISBN: 978-3-319-09234-8