Earth Systems Data and Models 1

Renji Remesan
Jimson Mathew

Hydrological Data Driven Modelling

A Case Study Approach

Springer

Earth Systems Data and Models

Volume 1

Aims and Scope

The book series Earth Systems Data and Models publishes state-of-the-art research and technologies aimed at understanding processes and interactions in the earth system. A special emphasis is given to theory, methods, and tools used in earth, planetary and environmental sciences for: modeling, observation and analysis; data generation, assimilation and visualization; forecasting and simulation; and optimization. Topics in the series include but are not limited to: numerical, data-driven and agent-based modeling of the earth system; uncertainty analysis of models; geodynamic simulations, climate change, weather forecasting, hydroinformatics, and complex ecological models; model evaluation for decision-making processes and other earth science applications; and remote sensing and GIS technology.

The series publishes monographs, edited volumes and selected conference proceedings addressing an interdisciplinary audience, which not only includes geologists, hydrologists, meteorologists, chemists, biologists and ecologists but also physicists, engineers and applied mathematicians, as well as policy makers who use model outputs as the basis of decision-making processes.

More information about this series at http://www.springer.com/series/10525

Renji Remesan · Jimson Mathew

Hydrological Data Driven Modelling

A Case Study Approach

Springer

Renji Remesan
Cranfield Water Science Institute
Cranfield University
Cranfield, Bedfordshire
UK

Jimson Mathew
Department of Computer Science
University of Bristol
Bristol
UK

ISBN 978-3-319-09234-8 ISBN 978-3-319-09235-5 (eBook)
DOI 10.1007/978-3-319-09235-5

Library of Congress Control Number: 2014947392

Springer Cham Heidelberg New York Dordrecht London

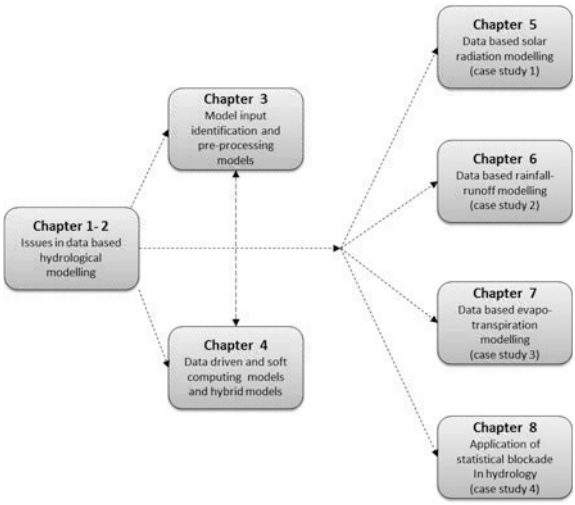
Cover image: pashabo/fotolia.com

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

A hydrological system is highly complex in nature with all processes within the system constituting dynamic and nonlinear interaction of several variables. Data-based soft computing techniques are emerging in the field of hydrology since the last couple of decades. Various developments in the recent data-driven soft computing models have shown their immense modelling capabilities amidst scarce and erroneous input space, such as models that are tolerant to imprecise and uncertain inputs. The notion and success of data-driven models depend on their learning capability from the given data set and on their capability to translate to useful information. Despite the growing advancements in data-driven approaches in hydrology, there are concerns on points like multi-collinearity, input selection, training data length selection, required data frequency for best modelling, model complexity control and modelling extreme values.



The aim of this book is a comparison of a number of state of art and traditional input selection approaches and specific data-driven models in different case studies considering the above-mentioned data-driven issues in hydrology. The structure of the book is as follows: Chap. 1 introduces modelling concepts and provides a review of data-driven modelling in modelling themes like rainfall-runoff dynamics, solar radiation and evaporation modelling. Chapter 2 starts with a brief detail of hydroinformatics and then addresses some of the data-based modelling issues in hydrology. This chapter reminds the need to evaluate existing hypothetic assumptions on various data-driven models including ANNs. Chapter 3 briefly discusses various novel approaches in data selection methods. The novel approach called the Gamma Test is described along with other mathematically sound techniques like Entropy Theory, Cluster Analysis, PCA, BIC and AIC. Towards the end of this chapter conventional data splitting and correlation approaches are described for the totality of the chapter. Chapter 4 includes details of data-driven artificial models and hybrid forms of these models that are intensively used in the field of hydrology, environment and other earth sciences. The chapter also describes conventional artificial intelligent techniques to investigate different aspects of the hydrological cycle. Conventional linear data-based techniques like ARX and ARMAX are described in the early part of the chapter. Following that, traditional ANN architecture is described along with different training algorithms adopted in this book for modelling. This chapter also deals with three other major nonlinear modelling techniques like Adaptive Neural Fuzzy Inference Systems (ANFIS, Support Vector Machines (SVMs) and Local Linear Regression (LLR). Discrete wavelet transforms (DWT) and its hybrid forms with ANNs, ANFIS and SVMs are briefly described.

Chapters 5–7 are case studies which incorporate all concepts and approaches described in Chaps. 2–4. Chapter 5 deals with a case study on data-based modelling on solar radiation estimation. This chapter draws different comparisons of working in data selection approaches (Gamma Test, Entropy theory, AIC (Akaike's information criterion) / BIC (Bayesian information criterion)) in solar radiation modelling. The modelling outputs of the proposed models and conventional models are discussed in detail after comparison with the observed measurements. The chapter deals with operations and applications of Conjugate Gradient ANNs and Levenberg–Marquardt ANNs along with other higher degree data-based models (ANFIS, SVM, wavelet hybrid models). Similar investigation is carried out in Chap. 6 in which the case study theme is rainfall-runoff modelling. This chapter also illustrates input redundancy checking and identifies the best data interval for rainfall-runoff dynamics modelling for the study region of the River Brue catchment. The next case study is evapotranspiration modelling; this is described in Chap. 7. In this theme we have tried to incorporate an analysis of different standard models of reference evapotranspiration along with data-based artificial intelligence models. The first few sections of this chapter are devoted to the results obtained from different Penman Montienth models in comparison with the newly proposed 'Copais Approach'. In later sections the results obtained from the data-based models are discussed in detail. We have also

introduced a unique final chapter, Chap. 8, in which we introduce a novel Monte Carlo (MC) technique called Statistical Blockade (SB), which focuses on significantly rare events in the tail distributions of data space and modelling. A case study modelling from a Himalayan river basin is introduced and compared with the results from that of ANNs and SVM in this chapter.

Cranfield, UK Renji Remesan
Bristol, UK Jimson Mathew

Acknowledgments

We express our immense gratitude to Prof. Dawei Han for his tremendous support, encouragement and valuable guidance during the research included in these case studies. We gratefully acknowledge Dr. Ian Holman, Dr. Michaela Bray, Dr. Ali Shamim and Dr. Azadeh Ahmadi for their professional support, responsiveness, encouragement and contributions. Thanks also to the following organizations for use of their data sets in the study of this book: British Atmospheric Data Centre (BADC) and NERC Hydrological Radar EXperiment (HYREX), National Centers for Environmental Prediction (NCEP) Climate Forecast System Reanalysis (CFSR) and Bhakra Beas Management Board (BBMB). Additional thanks are given to Dr. Alireza Moghaddamnia for providing evaporation data from the Chahnimeh reservoirs region of Iran. We express sincere thanks to Dr. Colin Clark for providing his lysimeter data for our research. Appreciation is also owed to Prof. Antonia J. Jones and her co-workers at University of Cardiff for providing WinGamma Tool.

Finally, we thank the Springer team for their support and fruitful collaboration during the preparation of the book. Special thanks to Dr. Robert K. Doe, Senior Publisher, Earth Sciences and Geography, and Naomi Portnoy, Springer project coordinator production, for their quick response and high-quality professional support. Special thanks to our proofreader Mr. Nick Freeman. Last but not least, special thanks are extended to Dr. Asha Nair for her invaluable help in compiling this book. We hope that the book will serve as an important intellectual resource for hydrological modeling researchers and push forward the boundaries of the field.

Renji Remesan
Jimson Mathew

Contents

Chapter 1
Introduction

Abstract In addition to classical physical/conceptual hydrological models at various complexity levels, data driven modelling tools are available in hydrological literature for last two decades to solve various complex issues in water resources and environmental science. "All models are wrong; some are useful." This quotation is meaningful in a data based hydrological modelling context due to the presence of different unsolved queries and deliberate assumptions. In a rush to develop interesting soft-driven models to solve differ processes issues, researchers often neglected or avoided in-depth researches on multi-collinearity, input selection, training data length selection, assuming that soft computing models have an intrinsic capability of managing extra errors caused by this negligence. Four case studies are illustrated in this book. These illustrate different model selection approaches and rigorously evaluate these approaches with state-of-art models through detailed and comprehensive experimentation and comparative studies. This chapter also aims to have a quick look into the critical points of current knowledge and or methodological approaches on data based modelling in hydrology and Earth sciences.

The hydrological cycle describes the natural phenomenon of continuous movement and changes of the state of water between the atmosphere and the earth. In modelling aspect, the hydrological cycle can be considered as a closed system because there are no external inputs or outputs of water to or from the system. The water movement from the earth's surface to the atmosphere is supported by solar radiation, while the water movement at and below the surface of the earth is mainly driven by gravity. The major processes in the hydrological cycle are illustrated in Fig. 1.1. The natural phenomena that make up the hydrological cycle are: the transfer of water, in its gaseous phase from land to the atmosphere (evapotranspiration), water transfer in its solid or liquid phases from the atmosphere to the land surface in the form of precipitation and land based phenomenon runoff, storage, infiltration and other processes shown in the Fig. 1.1. In short, solar radiation acts as a driving force of atmospheric phenomenon and gravity controls processes occurring at the land phase.

R. Remesan and J. Mathew, *Hydrological Data Driven Modelling*,
Earth Systems Data and Models 1, DOI 10.1007/978-3-319-09235-5_1

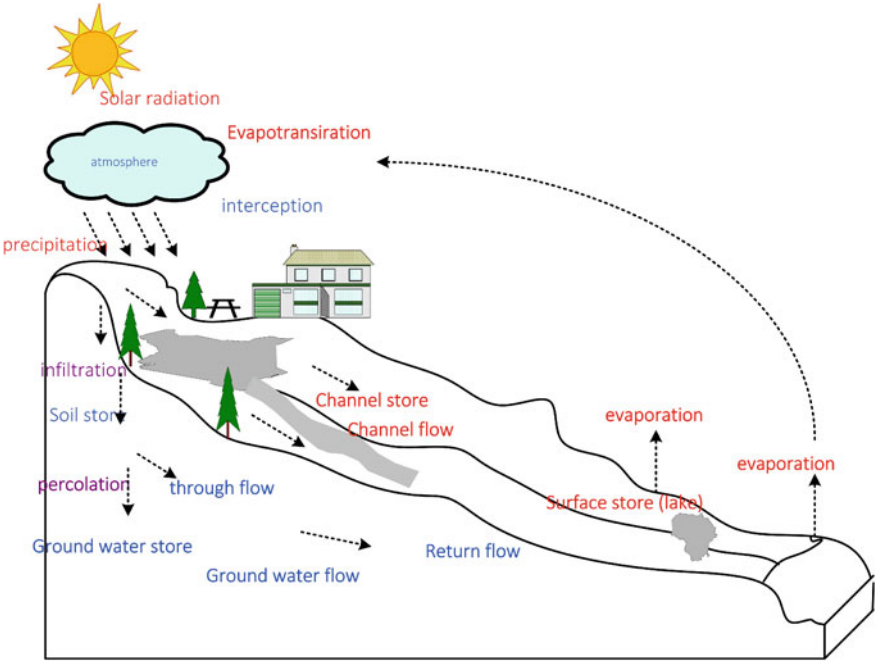


Fig. 1.1 Pictorial representation of hydrological cycle

The interdisciplinary science of hydrology is closely connected with human society and life of living beings, as water is the vital resource for survival. All kinds of personal, domestic, recreational, industrial, agricultural, luxurious needs of human society are closely connected with water usage and management. On the other hand, water poses serious threats to humanity in many forms, such as floods, drought, climatic change. This situation points to the necessity of a strong relationship between human society and water resources, through effective and practical applications of development, planning and management of this precious natural resource. Though, the process appears very simple from the outside, the subject is very vast and complex, due to the huge variety of processes involved, occurrence of these processes in different temporal and spatial scales and their interactive response on and with other environmental components. For the purposes of simplification and gaining a detailed insight into the hydrological details of the processes, hydrological models were introduced in the early second half of the 20th century with conceptual representations of the whole or a part of the hydrologic cycle.

1.1 Modelling in Hydrology

The known history of hydrology began around 5,000–6,000 BC. Evidence of this can be seen from the construction remains of canals, levees, dams, subsurface water conduits, and wells found in the Nile region of Egypt and Indus region of India. Nile river flow was monitored by the Egyptians as early as 3,800 years ago and one famous Indian scholar Kautilya used rainfall measuring instruments approximately 2,400 years ago [32]. The Roman architect Marcus Vitruvius made the first attempt to give a philosophical definition to the theory of the hydrologic cycle. More scientific studies on the hydrologic cycle were initiated in the seventeenth century, by the French scholars Pierre Perault and Edme Marriotte. By 1700, a British scientist Edmund Halley contributed to the work of Perault and Marriotte by estimating the quantity of water involved in the hydrologic cycle of the Mediterranean Sea and surrounding lands. The term "hydrology" received acceptance among scholars in its current meaning by around 1750 [68]. The eighteenth century witnessed the application of mathematics in hydrology and development of new dimensions of hydrology named fluid mechanics, and hydraulics by scientists like Pitot, Bernoulli, Euler, Chezy, and many more. The understanding of hydrological cycle and processes involved were solidified by the work of a British chemist John Dalton by the year 1800 [24]. The ground breaking innovation in hydrology occurred in the 18th century with the work of the Dutch-Swiss mathematician Daniel Bernoulli, which included the Bernoulli piezometer and Bernoulli's equation. The 19th century saw the development in groundwater hydrology, including Darcy's law, the Dupuit-Thiem well formula, and Hagen-Poiseuille's capillary flow equation. Hydrology had a close connection with civil engineering from the early days of development. During the 19th century, researchers started examining relationships between precipitation and stream flow. That information was used as a guideline for designing bridges and other structures. Daniel Mead published the first English-language text in hydrology in 1904 and Adolf Meyer followed with his text in 1919. Both publications were written for civil engineers. Rational analyses in hydrology began to replace empiricism in the 20th century and many organizations like International Union of Geodesy and Geophysics (1922), Hydrology Section of the American Geophysical Union (1930), which promotes researches in hydrology, were set up in the first half of the 20th century [55]. The second half of the 20th century witnessed the diversified application of techniques in the field of hydrology which includes statistical applications and artificial intelligence. Even now, in the 21st century, hydrology is developing day by day adding new concepts and approaches. The hydrological community is eagerly waiting for new breakthroughs and eminent deviations in hydrological science.

Hydrological models are mathematical representations of part or whole of the hydrological cycle in which processes involved in the transformation of climate inputs, such as precipitation, evapotranspiration, solar radiation and wind, through atmospheric, surface and subsurface transfers of water and energy into hydrological outputs like runoff, water level, etc. Singh and Woolhiser [83] define hydrological

modelling as the discipline that tries to quantitatively describe the land phase processes of the hydrological cycle. In a general conceptualized form, a hydrological model attempts to produce a desirable model response (e.g.: runoff) which describe the physical features of the hydrological system considered on a given forcing data (e.g.: rainfall, snow, temperature, solar radiation, evapotranspiration and wind velocity). The model may have two types of parameters: (i) physical parameters (which directly represent physical properties of the system and normally these quantities are measurable e.g.: catchment area, gradient, drainage length); (ii) process parameters (not directly measurable e.g.: depth of vadoze zone, soil depth, water holding capacity). The outputs or model responses are dependent upon the system defined by the modeller and the scope of the modelling. There are different kinds of models available, depending on which section of the hydrological cycle is taken as the subject of interest. Examples are, river runoff, and catchment overland flow, in the case of rainfall-runoff (RR) models, or groundwater flow and groundwater table elevation for groundwater (GW) models, extraterrestrial radiation and surface radiations in the case of solar models and evaporation and evapotranspiration in case of process models.

1.1.1 Model Classification

One can find different types of model classifications in literature depending on the different criteria of consideration. A comprehensive review of the existing and recent hydrological models can be found in Singh and Woolhiser [83]. However, as per degree of conceptualisation of the involving processes, models can be broadly classified into deterministic and stochastic models. Another meaningful classification is based on physically-based (white-box), conceptual (gray-box) and system theoretic (black-box) models. White box models are deterministic in nature and are made in a physically realistic manner, considering all internal sub processes and physical mechanisms involved in the phenomenon of the hydrological cycle. But, in most of the situations, practical reasons like data availability and calibration issues force the researchers to go for simple physically based or conceptual models with lumped representation of parameters or system [44]. This leads to another classification based on the spatial resolution at which the processes are described as: distributed, semi-distributed and lumped. The lumping could be a "structural lumping" of the study area or an "empirical lumping" of the dominant processes of interest [62]. The Sacramento soil moisture accounting (SACSMA) model of the US National Weather Service is the best example of a successful and widely used lumped model. In distributed models the hydrological processes are represented with a varying degree of high resolution in space and, in most cases, the model variables and parameters are also defined as functions of the space and time dimensions. But in the case of simple representation of lumped models, the hydrological system is represented as a unit block in which the varying properties are spatially averaged.

In semi-distributed models, the whole hydrological system is divided into different blocks, each represented by a lumped model. Even now, serious debate is taking place in the research community to establish the success rate of lumped models over complex distributed models and vice versa. Though, the conceptual and physics based models have given greater accuracy in terms of hydrograph modelling, there were still many issues to be further addressed by many researchers. Those difficulties include implementation and calibration difficulty, the vast amount of calibration data and the need of sophisticated tools etc. [29, 30, 87, 103].

1.2 Stochastic Modelling Case Studies in This Book

This book focuses on four major components in the hydrological cycle: solar radiation, precipitation (rainfall), evapotranspiration and runoff. These are illustrated as four case studies towards the end of this book. The following section gives the current modelling status in rainfall-runoff modelling, solar radiation modelling and evapotranspiration modelling.

1.2.1 Data Driven Rainfall-Runoff Modelling

Rainfall-runoff is a very complicated process due to its nonlinear and multidimensional dynamics. Hence it is difficult to model. The ASCE Committee on Surface-Water Hydrology (1965) introduced a new discipline incorporating statistical consideration into hydrology named "Stochastic Hydrology", defining "the manipulation of statistical characteristics of hydrologic variables to solve hydrologic problems on the basis of stochastic properties of the variables". This attempt has made a drastic change in the conventional direction of research and has encouraged many researchers to explore further the statistical and stochastic properties of hydrologic time series which have definite physical causes. An extensive review of the several types of stochastic models proposed for operational hydrology can be found in Lawrance and Kottegoda [58], Franchini and Todini [34] and Bras and Rodriguez-Iturbe [17]. Most of the black-box models include stochastic components and just relate outputs to inputs through a set of empirical functions, mathematical expressions or time series equations. The success rate of these data based stochastic models always encouraged hydrologists to implement simpler system theoretic models than the troublesome physics based or conceptual model, which demand more implementation and calibration effort but with the quality of the results comparable to the early mentioned stochastic models. In the early days, research concentrated more on Autoregressive (AR) and mixed Autoregressive and Moving Average (ARMA) models [16]. Later, linear time series models like ARX (auto-regressive with exogenous inputs) and ARMAX

(auto-regressive moving average with exogenous inputs) models [16] have gained more attention because of their satisfactory prediction performance and easy implementation procedure. The research conducted by [17, 88] has demonstrated the success of these linear models in different applications. Inability to represent the nonlinear dynamics inherent in hydrological processes was considered as the serious disadvantage of the above mentioned models [44]. The researchers quest for models which incorporate nonlinearity of the system with relatively short implementation effort, led hydrology to nonlinear pattern recognition and system control theory borrowed from electronics and communication engineering stream. In the early 1990s, much research was carried out in hydrology utilising the capabilities of advanced nonlinear system theoretical modelling approach called Artificial Neural Networks (ANN) [35, 50].

The advent of artificial intelligent techniques in hydrology brought a new dimension to flood modelling [18, 39, 40]. Among several artificial intelligence methods, artificial neural networks (ANN) holds a vital role and ASCE Task Committee Reports [11, 12] have accepted ANN as an efficient forecasting and modelling tool. Over the last decade, the artificial neural network has gained great attention and has evolved as the main branch of artificial intelligence that is now a recognized tool for modelling the underlying complexities in many artificial or physical systems including floods [2, 86]. Unlike traditional conceptual and physics based models, Artificial Neural Networks are able to mimic flow observations, without any mathematical descriptions of the relevant physical processes. A study by Jain et al. [46] demonstrated that the distributed structure of the ANN was able to capture certain physical properties. The success of hydrological forecasting systems depends on accurate predictions in the longer forecast lead time. Multi-step-ahead prediction is a challenging task which attempts to make predictions several time steps into the future. Dawson and Wilby [23] focused into neural network application on rainfall-runoff modelling and stream flow modelling. Maier et al. [61] provided a good review of neural network models used since 2000 for water quantity and quality modelling. Chang et al. [21] developed a two-step-ahead recurrent neural network for stream flow forecasting. Later, they explored three types of multi-step ahead (MSA) neural networks viz. multi-input multi-output (MIMO), multi-input single-output (MISO) and serial-propagated structure for rainfall-runoff modelling using data sets from two watersheds in Taiwan [22]. Nayak et al. [69] gave a detailed review of the application of ANFIS in rainfall runoff modelling. Mukherjee et al. [66] points out the advantages of support vector machines (SVMs) in making better predictions than other approximation methods such as polynomial and rational approximation, local polynomial techniques and artificial neural networks. A comprehensive review by Abrahart [1] provided two decades of neural network rainfall-runoff and streamflow modelling and suggested extended opportunities in this field.

1.2.2 Data Driven Solar Radiation Modelling

Solar radiation is one of the key inputs for most hydrological models in estimating reference evapotranspiration [93]. Moreover, daily solar radiation data is more popular than that of other time intervals for crop growth simulation models, hydrological and soil water balance models [14]. In spite of the great importance of solar radiation, many published studies pointed out the major challenges associated with solar radiation data collections. Lack of solar radiation data is quite common even in many developed countries, such as USA [42, 79] and Canada [49]. Many researchers pointed to the fact that solar radiation is an infrequently measured meteorological variable compared with temperature and rainfall [59, 102].

Over the past decades, many empirical and physical radiation models have been proposed [71, 72, 80, 97]. The Angstrom equation, which was proposed by Angstrom [10] and subsequently modified by Prescott [78], is considered as the most popular and widely used method for the estimation of monthly averaged daily (global) irradiation value. Later, several physical based empirical models appeared based on Chang [20], who reported that there was a good relation between net radiation and global solar radiation, since the latter is the principal source of energy. Based on this argument Bristow and Campbell [19], suggested an empirical relationship for daily global radiation, as a function of daily net radiation and the difference between maximum and minimum temperature. Later, Allen [9] suggested the use of a self-calibrating model to estimate mean monthly global solar radiation based on the work of Hargreaves and Samani [41]. His research suggested that the mean daily global radiation can be estimated as a function of net radiation, mean monthly maximum and minimum temperatures. The Bristow–Campbell model has been used in numerous hydrological related studies, and improvements have been developed over the past years [25]. The Campbell–Donatelli suggested method was implemented in many weather generators including MarkSim [48] and ClimGen [91]. Recently Donatelli et al. [27, 26] developed a windows based model named RadEst 3.00 which estimates and evaluates daily global solar radiation values at given latitudes. Some other interesting work has been done in the area of solar radiation prediction using ARMA and Fourier analysis [37, 67]. Furthermore, new approaches for predicting solar radiation series have been developed using ANN reported from different parts of the world, particularly from Turkey [74, 81, 95, 96], and authors from other places such as Negnevitsky and Le [70], Alawi and Hinai [4], Mohandes et al. [65], Kemmoku et al. [51] and Sfetsos and Coonick [82]. Mellit et al. [63] made an ANFIS-based prediction for monthly clearness index and daily solar radiation. A detailed review of ANFIS based modelling in solar radiation can be found in Mellit et al. [64]. Chen and Li applied support vector machine for the estimation of solar radiation from measured meteorological variables of 15 stations in China.

1.2.3 Data Driven Evapotranspiration Modelling

Evapotranspiration, termed ET for short, is a natural phenomenon which is the combined process of plant transpiration and soil evaporation. Though this study focusing on data based modelling with soft computing techniques, we have used some standard reference evapotranspiration equations for comparison. ET is considered as the most significant component of the hydrologic budget, apart from precipitation. Two commonly used evapotranspiration (ET) concepts are: potential evapotranspiration (ET_p) and reference evapotranspiration (ET_0). The ET_p concept was introduced in the late 1940s by Penman [75]. It defined as "the amount of water transpired in a given time by a short green crop, completely shading the ground, of uniform height and with adequate water status in the soil profile". In this definition of ET_p, the evapotranspiration rate is not related to a specific crop and therefore considered to be a shortfall. On the other hand ET_0 is defined as "the rate of evapotranspiration from a hypothetical reference crop with an assumed crop height of 0.12 m (4.72 in), a fixed surface resistance of 70 s m^{-1} and an albedo of 0.23, closely resembling the evapotranspiration from an extensive surface of green grass of uniform height, actively growing, well-watered, and completely shading the ground" [9]. In the late 1970s and early 1980s, the reference evapotranspiration concept was popularised among irrigation engineers and researchers which helped them to avoid ambiguities that existed in the definition of potential evapotranspiration.

The accurate estimation of reference evaporation is very critical in the context of many scientific and management issues; for example, irrigation system design, irrigation scheduling, hydrologic and drainage studies, crop production, management of water resources, evaluation of the effects of changing land use on water yields, and environmental assessment. The estimation of ET_0 depends on atmospheric variables, such as air temperature, solar radiation, wind speed, number of daylight hours, saturated vapour pressure and humidity. The Penman–Monteith approach recommended by FAO (FAO-PM) is considered as the standard to calculate reference evapotranspiration wherever the required input data are available [9]. Many researchers have made strong recommendations to consider FAO-PM as the standard method for evaluation of evapotranspiration through their comparative studies [8, 9, 43, 45, 84, 99]. Some studies also suggest that the ET estimation techniques are most appropriate for use in climatic regions similar to where they were developed [47, 75].

Other modifications of the Penman equation to estimate evapotranspiration from a hypothetical grass ET_0, are the CIMIS Penman equation [36, 85] and ASCE Penman equations. Doorenbos and Pruitt [28] added some modifications to the Penman combination equation, with a wind function that was developed at the University of California, Davis. This modification was adopted by California Irrigation Management Information System (CIMIS) for calculating hourly ET_0 and is

popularly known as CIMIS Penman equation [94]. ASCE-PM is a standardised calculation of reference evapotranspiration (ET) as recommended by the Task Committee on Standardization of Reference Evapotranspiration of the Environmental and Water Resources Institute of the American Society of Civil Engineers. Alexandris and Kerkides [5, 6] developed a new empirical equation for the hourly and daily estimation of evapotranspiration, using a limited number of readily available weather parameters and demonstrated the estimation of hourly values of ET_0 with a satisfactory degree of accuracy compared with the ASCE-PM estimation. The proposed equation is based on solar radiation, air temperature and relative humidity. The experiments had been conducted in an experimental field of The Agricultural University of Athens (Copais) in central Greece, using surface polynomial regression analysis. Thereafter the model was named the "Copais approach" for ET estimation. Even though, many equations have been developed and adapted for various applications based on available input data, there are still considerable amounts of uncertainty existing among engineers and environmental managers as to which method is to be adopted effectively in the calculation of ET_0 [7]. Several studies have been conducted by researchers for comparative evaluation of the most widely used and strongly recommended models for estimating hourly ET_0 like Penman–Monteith (FAO56-Penman–Monteith), CIMIS version of Penman (CIMIS-Penman), and the American Society of Civil Engineers version of Penman–-Monteith (ASCE-PM) [5, 28, 45]. In recent years several papers have evaluated hourly ET_0 equations (FAO-56 and ASCE Penman–Monteith, CIMIS Penman and Hargreaves) by comparing them with lysimetric measurements [15, 60]. Alexandris and Kerkides [5, 6] compared their model (Copais approach) performance with that of FAO-PM, ASCE-PM and CIMIS-PM for hourly and daily values ET_0 estimation using statistics and scatter plots.

Later data based approaches, including artificial intelligent techniques, have been applied in evapotranspiration estimation. Just as in the case of rainfall runoff modelling, ANN offered a promising alternative for modelling evapotranspiration in the case of data scarcity [53, 57]. The study by Sudheer et al. [92] used radial basis ANN in modelling ET_0 with limited climatic data. The study by Kumar et al. [57] used a multilayer perceptron (MLP) with back propagation training algorithm for estimation of ET_0 utilising various ANN architectures in data limited situations. Kisi [53] investigated the estimation of ET_0 using MLP. The results were compared with the above mentioned traditional approaches like Penman and Hargreaves' empirical models. Adaptive Neuro-fuzzy system (ANFIS) has been applied to evapotranspiration estimation by Kisi and Öztürk [54] to check the prediction capability. Wang and Luo [101] adopted Wavelet network model for reference crop evapotranspiration forecasting. A detailed study by El-Shafie et al. [31] suggests that ANN model is better than ARMA model for multi-lead ahead prediction of evapotranspiration.

1.3 Why Do You Read This Book?

Despite an abundance of studies on prediction and modelling of different hydrological processes in the hydrological cycle in the last few decades using nonlinear techniques like ANN, ANFIS and SVMs, there are still many questions that need to be answered. For example, to what extent do the inputs determine the output by a smooth model? Given an input vector x, how accurately can the output y be predicted? How many data points are required to make a prediction with the best possible accuracy? Which inputs are relevant in making the prediction and which are irrelevant? These questions have not been fully addressed adequately by the hydrological community [39]. The hydrological community acknowledged that issues like evaluation of available data, assessment of data adequacy and optimum decision on input selection are main challenges and potentially complicated questions in data based modelling. Although the performance of a model generally improves with addition of more information during model calibration, plateaus exist wherein new information adds little to a model's performance [77, 90]. In fact, systems accuracy can be reduced with increasing information during validation [90], usually because the additional variables produce models with overfitting problems [98]. An overfitted model is very specific to the training set and performs poorly on the test set. Overfitting is known to be a problem with multi-variate statistical methods when the data set contains too many predictor variables [98], which lead to excellent results on the training data but very poor results on the unseen test data. Therefore, an important question for modellers is which inputs are relevant in making the prediction and which are irrelevant.

However, due to the advancement of modern computing technology and a new algorithm from the computing science community called the Gamma Test [3, 56], it is possible that we could make significant progress in tackling these problems. A formal proof for the Gamma Test (GT) can be find in Evans and Jones [33]. It is accomplished by the estimation of the variance of the noise var(r) computed from the raw data using efficient, scalable algorithms. This novel technique, the Gamma Test, enables us quickly to evaluate and estimate the best mean squared error that can be achieved by a smooth model on unseen data for a given selection of inputs, prior to model construction. This technique can be used to find the best embedding dimensions and time lags for time series analysis. This information would help us determine the best input combinations to achieve a particular target output. The Gamma Test can avoid overtraining, which is considered as one of the serious weaknesses associated with almost all nonlinear modelling techniques including ANN. The Gamma Test is designed to solve this problem efficiently by giving an estimate of how closely any smooth model could fit the unseen data. Thus we can avoid the guesswork associated with the nonlinear curve fitting techniques. This book makes use the capabilities of these concepts in input selection and redundancy assessment when we have large number of input series for modelling.

Information theory is a widely used mathematical theory in electronics and communication. Information theory has two primary goals: (i) to develop the fundamental theoretical limits on the achievable performance when communicating a given information source over a given communications channel using coding schemes from within a prescribed class; (ii) to develop coding schemes that provide performance that is reasonably good in comparison with the optimal performance given by the theory. More detailed concept of Information Theory and Entropy can be found in Gray [38]. The capability mentioned in the first goal could be used for data quality assessment. In information theory, entropy is often referred as Shannon entropy which measures the uncertainty and randomness associated with a random variable. Capabilities of Shannon entropy to measure the average information content associated with input data series are explored in this book. Despite the claimed success of the methods from the aforementioned literature, there is a lack of comparison with conventional methods, data splitting approaches like cross validation approaches and cross correlation approaches. This book aims at comparison and assessment of model input selection based on the Gamma Test, entropy theory, AIC BIC and the above mentioned traditional benchmarking approaches in modelling.

This book makes an effort to comment on another rising debate in data based modelling in hydrology: should the input data be treated as signals with different frequency bands so that they could be modelled separately? Wavelet theory is a novel field of mathematics, which recently gained attention among scientists studying acoustics, fluid mechanics, and chemistry. The concept of wavelet transformation involves representation of a general signal or time series in terms of simpler, fixed building blocks of constant shape but at different scales and positions. Discrete Wavelet Transforms (DWT) can give very useful decomposition of time series in such a way that faint inherent temporal structure of the time series can be revealed and can be effectively handled by the above mentioned and used non-parametric models in this book. This capability has been used effectively in various fields of engineering for dealing issues in noise removal, object detection, image compression and structural analysis [89]. Unlike Fourier transforms, wavelets have an ability to elucidate simultaneously both spectral and temporal information within the signal whereas Fourier spectrum contains only globally averaged information. This property overcomes the basic shortcoming of Fourier analysis in modelling. Therefore, data pre-processing can be carried out by time series decomposition into its subcomponents using wavelet transform analysis [73]. The wavelets can express original signals as additive combination of wavelet coefficients at different resolution level. A study by Aussem et al. [13] was the first hybrid ANN-wavelet conjunction model in which they used it for financial time series forecasting. Later Zhang and Dong [104] proposed a short-term load forecast model based on multi-resolution wavelet decomposition with ANN model. The first application in hydrology was in 2003 when Wang and Ding [100] applied wavelet-network model to forecast shallow groundwater level and daily discharge. In the same year Kim and Valdes [52] applied this conjunction model concept in coupling dyadic wavelet transforms and ANNs to forecast droughts for the Conches river basin. Keeping the success stories of the aforementioned literature in mind, this book attempts to

couple DWT with nonlinear models like ANN, ANFIS and SVM and apply these novel hybrid schemes to three case studies. In some cases, a well calibrated data based model may not provide faultless forecast results over a longer time. To improve and make such dynamic conceptual models suitable for operational and long term real time predictions, integration of on-line or sequential data assimilation techniques could be used. Partial Recurrent Neural Networks (PRNN) is good example for such model with data assimilation principle. Kalman filtering is a most common data assumption exercise widely used in environmental application along with data based models with definite state space architecture. Application of data assimilation approaches and related case studies are beyond the scope of this book.

The relevant questions in data based modelling in hydrology are how useful is a model for predicting a particular component within the hydrological cycle and does a complex model work better? Though such debates are prominent in physics based modelling, related literature is almost non existant in the case of data based modelling. The usefulness of any model depends ultimately on the directional accuracy of its estimates, not on its ability to generate unassailably correct numerical values [76]. Critical evaluation on the usefulness of models based on sensitivity modelling error and complexity is essential in data based modelling. This study introduces an index of utility for critical evaluation of models in different modelling situations which utilises information like model sensitivity (response to changes in training data set), model complexity (changes in training time) and model error (closeness of simulation to measurement) for all used models in this book. Extreme value modelling is a challenging field in hydrology. This made an attempt to use state-of-art Statistical Blockade in extreme value modelling and compare the capabilities with other data driven approaches.

In short, this book aims to address the above mentioned issues in data based modelling by the following means:

1. The application of novel approaches in data and model selection to avoid the aforementioned difficulty associated with data based modelling;
2. A reliability check of the novel data selection approaches with conventional methods;
3. To propose and use new wavelet hybrid schemes with traditional data based intelligent systems;
4. To investigate the capabilities of popular and widely used artificial intelligent models with newly proposed hybrid schemes.
5. To introduce statistical blockade to hydrology and compare the capabilities with other models.

The above mentioned five objectives are accomplished through four case studies broadly dealing with data based modelling in respective fields.

Chapter 2 describes the modelling issues associated with data based modelling in hydrology. The chapter gives a detailed description on puzzling questions in hydrology like model selection, selection of model input architecture, selection of training data length, selection of best data interval etc. Another main goal of this chapter is to suggest an approach to identify and characterize modelling quality as a

function of the model complexity, model sensitivity and model error. Major studies made in this Book are conducted on the upper Brue catchment, Somerset, using the HYREX data set. However, for evapotranspiration estimation, we have used data from three other catchments, namely the Santa Monica station of the USA, the Chahnimeh reservoirs region of Iran and the Beas basin in India. The details of the catchments including location specification and data collection description are given in each case study. The detailed illustration of statistical parameters of the data used for the modelling is given in respective case study chapters. Different novel approaches in data selection methods are introduced and discussed in detail in Chap. 3. The novel approach called the Gamma Test has been described along with other mathematically sound techniques like Entropy Theory, Cluster Analysis, PCA, BIC and AIC and other traditional approaches. Chapter 4 gives details data driven models used in this study (ANNs, ANFIS, SVMs, and other hybrid forms). Chapters 5, 6 and 7 focuses different case studies on research themes like solar radiation modelling, rainfall-runoff dynamics and evapotranspiration modelling. Chapter 8 describes mathematical details of state-of-art Statistical Blockade and a river basin scale case study to illustrate its capability in extreme value modelling.

References

1. Abrahart RJ (2012) Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Prog Phys Geogr 36(4):480–513
2. Abrahart RJ, See LM (2007) Neural network modelling of non-linear hydrological relationships. Hydrol Earth Syst Sci 11:1563−1579. doi:10.5194/hess-11-1563-2007
3. Agalbjörn S, Končar N, Jones AJ (1997) A note on the gamma test. Neural Comput Appl 5 (3):131−133. http://dx.doi.org/10.1007/
4. Alawi SM, Hinai HA (1998) An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation. Renew Energy 14(1–4):199–204. doi:10.1016/S0960-1481(98)00068-8
5. Alexandris S, Kerkides P (2003) New empirical formula for hourly estimations of reference evapotranspiration. Agric Water Manage 60:181–198
6. Alexandris S, Kerkides P, Liakatas A (2006) Daily reference evapotranspiration estimates by the "Copais" approach Agric. Water Manage 82:371–386
7. Alkaeed O, Flores C, Jinno K, Tsutsumi A et al (2006) Comparison of several reference evapotranspiration methods for Itoshima Peninsula Area, vol 66. Memoirs of the Faculty of Engineering, Kyushu University, Fukuoka, Japan, p 1–14
8. Allen RG, Jensen ME, Wright JL, Burman RD et al (1989) Operational estimates of reference evapotranspiration Agron. J 81:650–662
9. Allen RG, Pereira LS, Raes D, Smith M et al (1998), Crop evapotranspiration: guidelines for computing crop water requirements. United Nations Food and Agriculture Organization, Irrigation and Drainage Paper 56 Rome, Italy
10. Angstrom A (1924) Solar and terrestrial radiation. Quart J Roy Meteorol Soc 50:121–125
11. ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000a) Artificial neural networks in hydrology—I: preliminary concepts. J Hydraul Eng ASCE 5 (2):115–123
12. ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000b) Artificial neural networks in hydrology—II: hydrologic applications. J Hydraul Eng ASCE 5 (2):124–137