CS/AR-16/1999-2000

RAINFALL-RUNOFF MODELING USING ARTIFICIAL NEURAL NETWORK TECHNIQUE

NATIONAL INSTITUTE OF HYDROLOGY
JAL VIGYAN BHAWAN
ROORKEE - 247 667 (UTTARANCHAL)

1999-2000

Preface

An artificial neural network (ANN) is a flexible mathematical structure, which is capable of identifying complex non-linear relationships between input and output data sets. ANN models have been found useful and efficient, particularly in problems for which the characteristics of the process are difficult to describe using physical equations. The success with which ANNs have been used to model dynamic system in other fields of science and engineering, suggests that the ANN approach may prove to be an effective and efficient way to model the rainfall runoff process. Further, for hydrological applications, ANN models can take advantage of their capability to reproduce the unknown relationship existing between a set of input variables descriptive of the system, such as rainfall and river flow.

This report, titled *'Rainfall-runoff modeling using artificial neural network technique'*, presents a research study conducted to develop a rainfall-runoff model using ANN approach and has been trained and validated for the Baitarani River Basin, Orissa. The study demonstrates the applicability of ANN approach in developing effective non-linear models of Rainfall Runoff process without the need to explicitly represent the internal hydrologic structure of the watershed. The study has been done by Sri. K. P. Sudheer, Scientist 'B', with the assistance of Sri. P. C. Nayak, Scientist 'B' and Sri. D. Mohan Rangan, Technician Gr. II. Dr. Ramasastri, Scientist 'F' and Co-ordinator, supervised the research work.

(K.S. RAMASASTRI)
DIRECTOR

Acknowledgements

We express our deep sense of gratitude to Dr. S. M. Seth, Director, National Institute of Hydrology, Roorkee for providing us with the infrastructure and support for conduct of this research work. We are grateful to Dr. K. S. Ramasastri, Scientist 'F' and Co-ordinator for his valuable guidance during the formulation and completion of this study.

We gratefully acknowledge the valuable suggestions rendered by Dr. A. K. Gosain, Professor, Civil Engineering, Indian Institute of Technology, Delhi during many discussions for developing the methodology adapted in this work.

We are indebted to Mr. B. B. Sing Samanth, then Director, Orissa Water Planning Organization, Bhubaneswar (currently working as Chief Engineer, Designs) for providing us the data used in this study. The co-operation extended by Mr. N. C. Mohanty, Deputy Director, and Mr. Abhimanyu Behra, Assistant Director, OWPA, Bhubaneswar is also duly acknowledged.

We do express our sincere gratitude to the Scientists and staff members of Deltaic Regional Centre, Kakinada for providing support in various means throughout the course of the study.

List of Figures

Contents

List of Tables

ABSTRACT

The artificial neural network (ANN) methodology has been reported to provide reasonably good solutions for circumstances where there are complex systems that may be poorly defined or understood using mathematical equations, problems that deal with noise involve pattern recognition, and situations where input data are incomplete or ambiguous by nature. Because of these characteristics, it was believed that ANN could be applied to model the daily rainfall runoff relationship. Accordingly, a research study was conducted by employing ANN computing approach to forecast daily runoff as a function of daily precipitation and previous values of runoff. The model was trained and tested for the data of the Baitarani River Basin, Orissa. Two ANN algorithms were considered while developing the model, namely back error propagation network (BPN) and radial basis function network (RBF). The sensitivity of the prediction accuracy to the number of hidden layer neurons in a back error propagation algorithm was investigated. Based on this analysis, two BPN models were selected to represent the rainfall-runoff transformation. These two BPN models and the RBF model were compared for their performance using various statistical indices. The performance ANN model for Baitarani river basin was compared with that of existing models. The study demonstrates the applicability of ANN approach in developing effective non-linear models of Rainfall Runoff process without the need to explicitly represent the internal hydrologic structure of the watershed. The developed ANN model was found performing to a good degree of accuracy compared to other models in use.

Chapter 1
Introduction

Floods are indeed a part of the earth's natural water cycle and have been occurring right from the beginning. In fact, earth's geography has time and again been altered by floods and changing courses of major river systems. However, the damage due to floods had tended to increase with time due to greater interference by man in natural processes and encroachment of flood plain zones and even riverbeds. The problem of floods faced by India is unique in several respects due to varied climate and rainfall patterns in different parts of the country. Of the country's total geographical area of about 328 million hectares (m. ha.), about 41 m.ha. (nearly one eighth) is considered flood prone. There are occasions when one part of the country is experiencing floods while another is in the grip of a severe drought. Forewarning of floods can indeed go a long way in preventing much of the potential damage due to floods. For many years, hydrologists have attempted to understand the transformation of rainfall to runoff, in order to forecast stream flow for purposes such as water supply, flood control, irrigation, drainage, water quality etc.

The rainfall runoff transformation is one of the most complex hydrologic phenomena to comprehend due to the tremendous spatial and temporal variability of watershed characteristics and precipitation patterns and number of variables involved in the mathematical modeling of the physical processes. Since 1930's numerous rainfall runoff (R-R) models have been developed to forecast stream flow. Conceptual R-R models are designed to approximate with their structures (in some physically realistic manner) the general internal sub processes and physical mechanisms which govern the hydrologic cycle. These conceptual models, usually incorporate simplified forms of physical laws and are generally non linear, time invariant and deterministic with parameters that are representative of watershed characteristics. Until recently, for practical reasons (data availability, calibration problems etc.), most conceptual watershed models assumed lumped representation of the parameters.

While conceptual models are of importance in the understanding of hydrologic process, there are many practical situations such as stream flow forecasting where the main concern is with making accurate predictions at specific watershed locations. In such a situation, a hydrologist may prefer not to spend the time and effort required to develop and implement a conceptual model, and instead implement a simple system theoretic model (some times referred to as black box). In these models, difference equations or differential equation based models are used to identify a direct mapping between the inputs and outputs without detailed consideration of the internal structure of the physical processes. The linear time series models such as ARMAX (auto regressive moving average with exogenous inputs) models developed by Box and Jenkin (1976) have been most commonly used and have been found to provide satisfactory predictions in many applications. However, such models do not attempt to represent the non linear dynamics inherent in the transformation of rainfall to runoff and therefore may not always perform well.

Owing to the difficulties associated with non-linear model structure identification and parameter estimation, very few truly non-linear watershed models have been reported. In most cases, linearity or piecewise linearity has been assumed. The model structural errors that arise from such assumptions can, to some extent, be compensated for by allowing model parameters to vary with time. For example, real time identification techniques such as recursive least squares and state space Kalman filtering have been applied for adaptive estimation of model parameters, with generally acceptable results.

Recently significant progress in the field of non-linear pattern recognition and system control theory have been made possible through advances in a branch of non linear system theoretic modeling called artificial neural network (ANN). ANN is a non-linear mathematical structure, which is capable of representing arbitrarily complex non-linear processes that relate the inputs and outputs of any system. The success with which ANNs have been used to model dynamic system in other fields of science and engineering, suggests that the ANN approach may prove to be an effective and efficient way to model the rainfall runoff processes in situations where explicit knowledge of the internal hydrologic sub processes is not required.

ANN was first developed in the 1940's and in recent decades, considerable interest has been raised over their practical applications, because the current algorithms overcome the limitations of early networks. There are a wide variety of ANN algorithms, however the main function of all ANN paradigms is to map a set of inputs to a set of outputs. An ANN is described as an information processing system that is composed of many non-linear and densely interconnected processing elements of neurons. ANNs are proven to provide better solutions when applied to (i) complex systems that may be poorly described or understood; (ii) problems that deal with noise or involve pattern recognition, diagnosis, abstraction and generalization and (iii) situation where input is incomplete or ambiguous by nature. ANN has the ability to extract the patterns in phenomena and overcome the difficulties due to the selection of model form such as linear, power or polynomial. An ANN algorithm is capable of modeling the rainfall runoff relationship due to its ability to generalize pattern in noisy and ambiguous input data and to synthesize a complex model without a priori knowledge or probability distributions.

In this study, ANN algorithms were used to model the daily rainfall runoff relationship for the Baitarani river basin, Orissa, India. The study demonstrates the applicability of ANN approach in developing effective non-linear models of rainfall runoff process without the need to explicitly represent the internal hydrologic structure of the watershed. The study also aims at identifying the best ANN algorithm/structure to represent the rainfall runoff process effectively. The performance of the ANN model for Baitarani river basin was compared with that of existing models.

Chapter 2

Study area and data

The Baitarani river basin, covering an area of 14,218 sq. km is one of the three major basins in the Orissa State. The index map of Baitarani basin in shown in Fig 2.1. Although comparatively it is smaller than that of Mahanadi and Brahmani basins, it brings heavy flow and creates havoc in lower reaches during monsoon. Out of its drainage area of 14,218 sq.km, 736 sq.km. lies in Singhbhum district of Bihar and the rest lies inside the state of Orissa.

The Keonjhar district of Orissa covers the major portion of the basin area whereas Mayurbhanj, Sundargarh, Dhenkanal, Cuttack and Balasore districts cover the rest. The river, after traversing in hilly regions, enters the plains at Anandpur. Further below it meets the deltaic region at Akhuapada where it branches off and bifurcates. Further below it meets the river Brahmani and assumes the name Dhamara and joins the Bay of Bengal. The basin map, up to Anandpur is shown in Fig. 2.2.

The River System

The river Baitarani originates from Guptaganga hills in Keonjhar district of Orissa, about 2 km from Gonasika village, at an elevation of 900m at Latitude 21^0 31' North and Longitude 85^0 33' East. Initially the river flows in a northern direction for about 80 Km and then takes a sudden right-angled turn. In this reach the river serves as a boundary between Bihar and Orissa states for a certain length that is up to the confluence of Kongira river. The river while flowing towards south enters the plains at Anandpur and further downstream meets the deltaic zone at Akhuapada. The river after travelling a total distance of 360 km joins the bay of Bengal. There are, in all 64 tributaries of Baitarani river out of which 35 join in the left and 29 join in the right side. The prominent tributaries are Kangira, Khairi Bandhan, Deo, Kanjhari, Sita, Kusei and Salandi.

Basin Characteristics

Climate

The Baitarani Basin is having 14,218 Sq.Km drainage area. The basin consists of Singhbhum district of Bihar and Keonjhar, Dhenkanal, Mayurbhanj, Sundargarh, Cuttack and Balasore districts of Orissa. The upper Baitarani is about 700m above msl and therefore the climate of upper Baitarani is of extreme nature. The middle Baitarani basin is partly hilly and partly plain, and the lower Baitarani basin in Coastal area. The effect of the sea is very much felt in Lower Basin of Coastal plain. Keonjhar is the district headquarters of

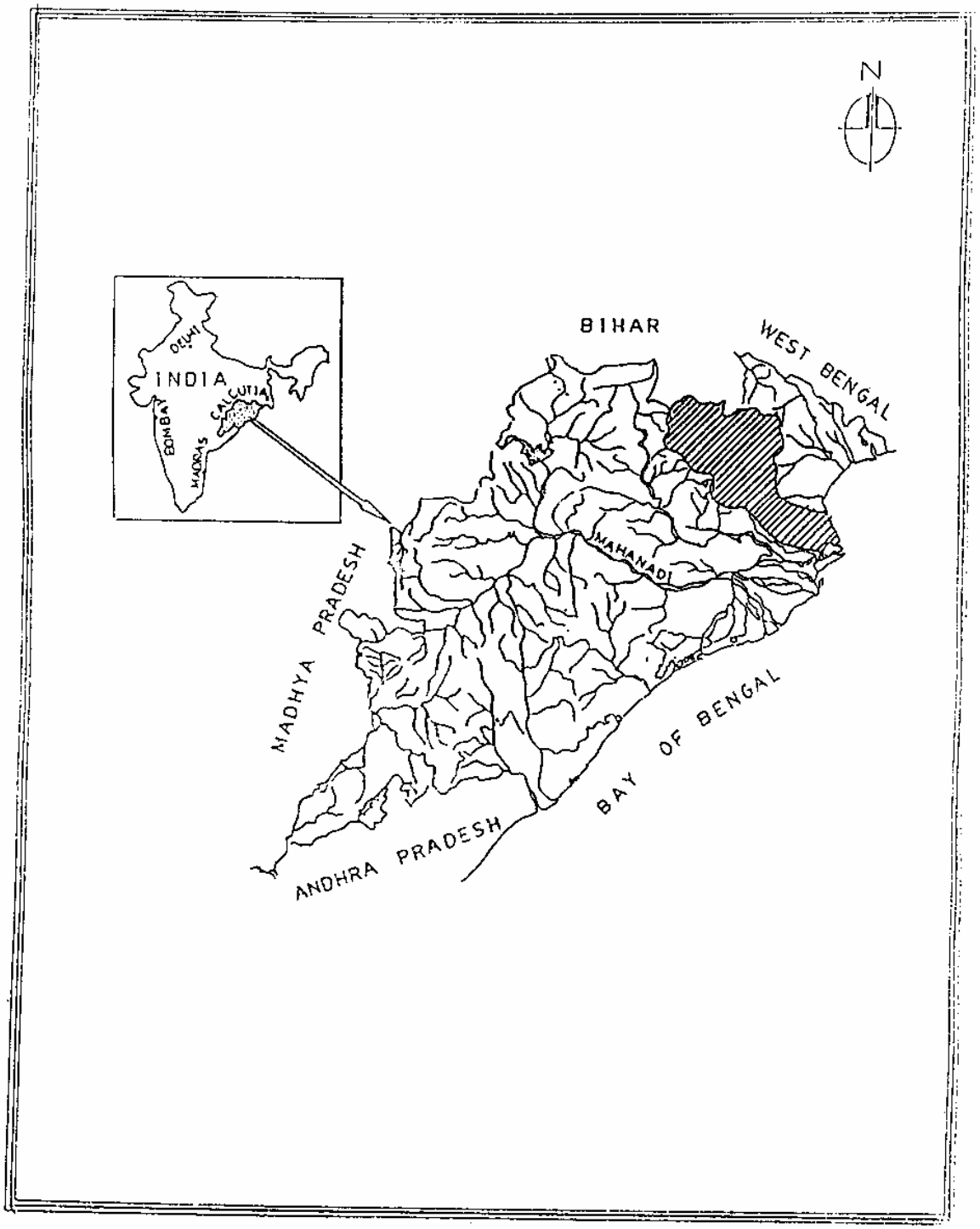


Fig 2.1 Index map of Baitarani river basin

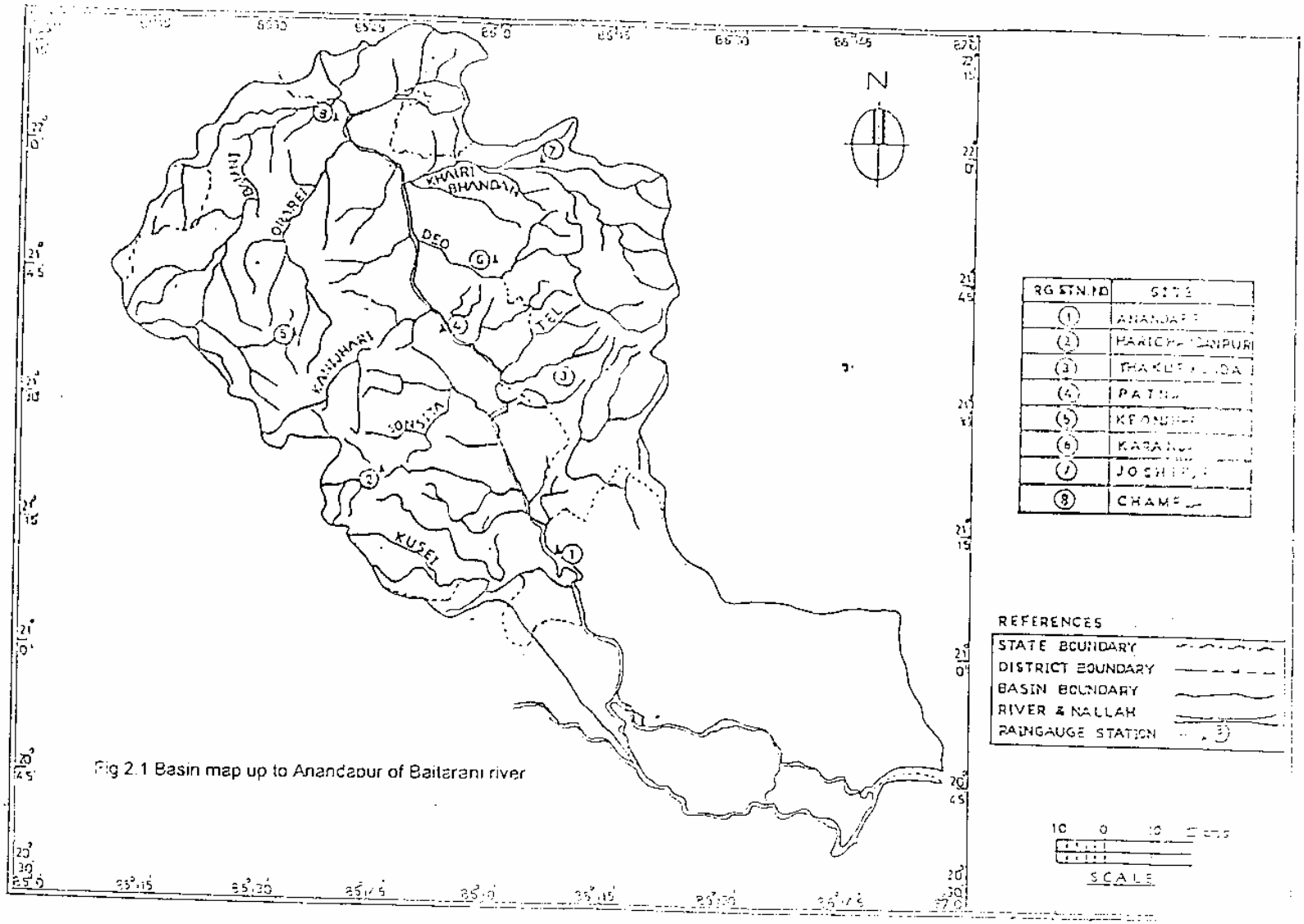


Fig 2.1 Basin map up to Anandapur of Baitarani river

Keonjhar district, situated in the middle of the basin. One IMD station is functioning at Keonjhar from which all detail information pertaining to climate and other meteorological data can be obtained. The maximum-recorded temperature of Keonjhar district in summer days is 48.5 degree centigrade and minimum in winter days is 6 degree centigrade.

The rainfall received in the basin is mainly from South West Monsoon and lasts from June to October. About 80% of annual precipitation occurs during these months. The annual rainfall is of the tune of 1595 mm. and average rainfall is 1187 mm.

The average monthly humidity data are available at IMD station Keonjhar as well as at Cuttack and Balasore. It is seen that the relative humidity is minimum in the months of April and May and maximum in the months of August and September. The maximum and minimum humidity are of the order of 83.08% and 39.63% respectively, on average.

The maximum cloud cover is observed in the months of June and July whereas minimum is in December and January.

Geology

The geological features in and around the Upper Baitarani are of two main series – the iron ore series and the younger Kolhan series. The iron ore series are represented by mica, hornblende, schist, bornblende, gneiss, phylile, Chert and Jasper which along with Singhbhum granite constitute the surrounding Country rock. The Kolhan series comprises mainly flat bedded Kolhan, sand stone and Conglomerate. The sandstone usually forms the flat-topped hills over the peneplained granite terrain in this area. The generalized geological set up for whole of south Singhbhum and Keonjhar district, is
(i) New Dolerite (ii) Kolhan series (iii) Singhbhum granite (iv) Iron ore series.

Socio-economic status

The population of the basin as per the 1991 Census was 31,05,926 out of which the rural population was 28,09,671. As such the rural population comes to 90.5% of total population. The density of population in the basin as per the 1991 census is 218 per sq.km as against state average of 202 per sq.km.

The river Baitarani flows mostly through the Keonjhar district of Orissa, where the pace of development was absolutely slow in the past. The people of the basin mostly depend on agriculture, which is subjected to vagaries of nature, in form of drought due to erratic, uneven rainfall caused mostly by depression in Bay of Bengal during the monsoon season. There are only a few industries in this basin, though the basin has abundance of mineral resources.

Most of the people, who depend on agriculture, have low per capita agricultural income. So the people are economically poor and backward.

Keonjhar district forms about 60% of the Baitarani basin. The basin population works out to be 31,05,926 (1991 Census). The urban population is 2,96,255 and comes to 9.5% of the total population of the basin. There are 11 towns in the basin namely, Joda, Champua, Barbil, Anandpur, Karanjia, Bhadrak, Chandbali, Jajpur, Jajpur Road, Basudevpur.

Food and agriculture

Agriculture is the primary occupation of the inhabitants of the basin. However due to lack of irrigation facilities, farmers are forced to depend on the uncertainties of rainfall as such agricultural production is much below average production of the State. However after construction of Salandi Irrigation System in Lower sub-basin as well as Kanjhari and Remal Project in middle sub-basin more areas have come under irrigation. Further the ongoing projects like Kanupur, Deo etc. will increase the irrigation potential. Though the basin is rich in minerals, industrial development has not taken place due to lack of infrastructure facilities as well as lack of investment by government and other private agencies.

Land use pattern

Out of 831 thousand hectares of geographical area of Keonjhar district, 307 thousand hectares come under net area sown and 249 thousand hectares come under forest coverage. The net area sown and forest area are 37% and 30% of total area. The above figure of Keonjhar district gives an idea of the land coverage of the basin as it cover 60% area of the basin.

Irrigation

In early days there was very little irrigation in this basin. The trend of irrigation has undergone changes since 1978-79 due to construction of medium and minor irrigation projects. At present the total area under irrigation in Kharif of the basin is reported to be 134458 ha. The technique of modern agriculture systems with application of chemical fertilizers, pesticides and use of tested seeds has not been widely practiced in this basin, as such coupled with lack of irrigation facilities the yield of food grains is coming much less than the average.

Industries

Despite the available rich mineral wealth, the industrial development in Baitarani basin is very slow. Though the basin is rich in minerals, industrial development has not taken place at a faster rate due to lack of infrastructure development and lack of investment by Government and private agencies. The important industries in the basin are, TISCO Ferro Manganese

plant, Joda, Ipitata Sponge Iron Plant, Joda, Orissa Sponge Iron Plant Palaspanga, Kalinga Iron works, Badbil, Tata Iron and Steel Company, Brahamnipal, Electrochem Orissa Ltd., Joda and Ferro Alloys Corporation Lts., Randia. In addition to the above there is a number of small scale industries in the basin.

Data Availability

The present study intends to develop a rainfall-runoff relationship for the basin from the available historical data records, so as to develop effective water management policies to meet the demand from all sectors.

Rainfall

The Baitarani basin has 15 rain gauge stations in and round it. These are concentrated mostly in upper and middle portions of the basin. There are breaks in continuity of data of some of the stations. After checking the data of all the stations for period of availability, four rain gauge stations were considered for the study. The location map of all the rain gauges is depicted in figure 2.3.

Stream flow

Central Water Commission (CWC) maintains one gauge-discharge (G-D) site at Anandpur, intercepting a catchment area (C.A.) of 8570 sq.km. Daily values stream flow data at this G-D site are available from 1971. The mode of observation is current meter. The G-D site at Biridi intercepts a CA of 10125 sq.km. The Department of Water Resources, Govt. of orissa maintain it. The Runoff data are available from 1964, but the reliability of observed data is not good. The Department of Water Resources maintain the G-D site at Basudevpur. The site intercepts a CA of 1525 sq.km. The mode of observation is by float and the quality of data is not reliable in this case too (OWPS, 1994). It has been seen at the time of preparation of updated yield series of the Kanupur project that the observed data were inconsistent. The Department of Water Resources maintain the G-D site at Tondo too. The site intercepts an area of 6708 sq.km. But the data were found to be inconsistent at the time of preparation of detailed project report for Bhimkund Irrigation project, as reported by Orissa Water Planning Organization, Bhubaneswar.

Data preparation

The data of Anandpur G-D site maintained by CWC are reliable and of good quality. The data for the period 1972-1994 were employed in the study. The study has been restricted to monsoon season (June to October) alone, since the interest was to forecast the flood flows in the basin using the developed model. The entire available data has been used for standardizing the records and is described in detail in the next chapter.

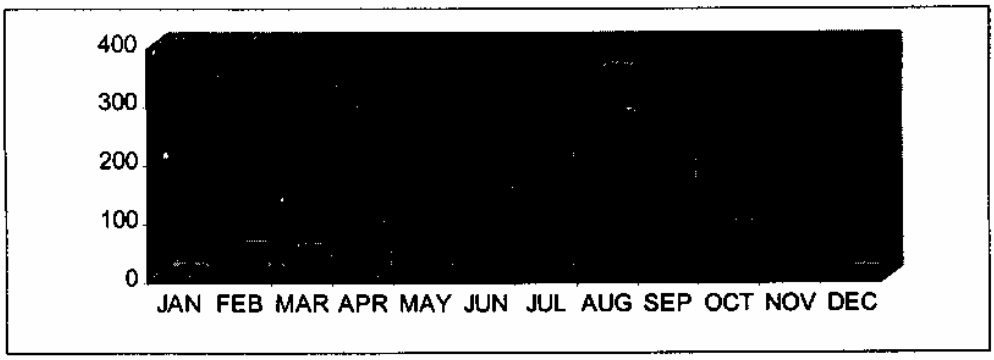


Fig 2.3 Mean monthly rainfall derived from 24 years data at Anandpur

The rainfall data used in the study were for the period of 1980-1982. This constraint for sticking to three years of data has arisen due to non-availability of rainfall data for the corresponding period in which flow data were also available. However, rainfall data were available for more than 10 years in single rain gauge station at Anandpur. Since a true areal representation of rainfall is preferred in any rainfall-runoff modeling, only a short duration data was employed in the study. The mean monthly values of rainfall, derived from the Anandpur station data are presented in Fig 2.3.

A preliminary analysis was conducted to assess the consistency of the data used in the study, on an average, by developing runoff coefficient values for the Baitarani basin. The analysis resulted in consistent values of runoff coefficient for the years 1980 to 1982 (25 to 30%) on an annual basis. The runoff coefficient evaluated for the monsoon season was of the order of 35 to 48%. The high value for the coefficient was obtained during flood events. This analysis confirms the data consistency.

Table 2.1 The rainfall stations and their Thiessen weights

Station Name	Thiessen weight (%)
Champua	32.72
Karanjia	32.14
Thakurmunda	20.95
Anandpur	14.19

Thiessen polygons have been drawn (Fig 2.2) for the four stations considered in this study to compute the weights. The influencing stations and their corresponding Thiessen weights are

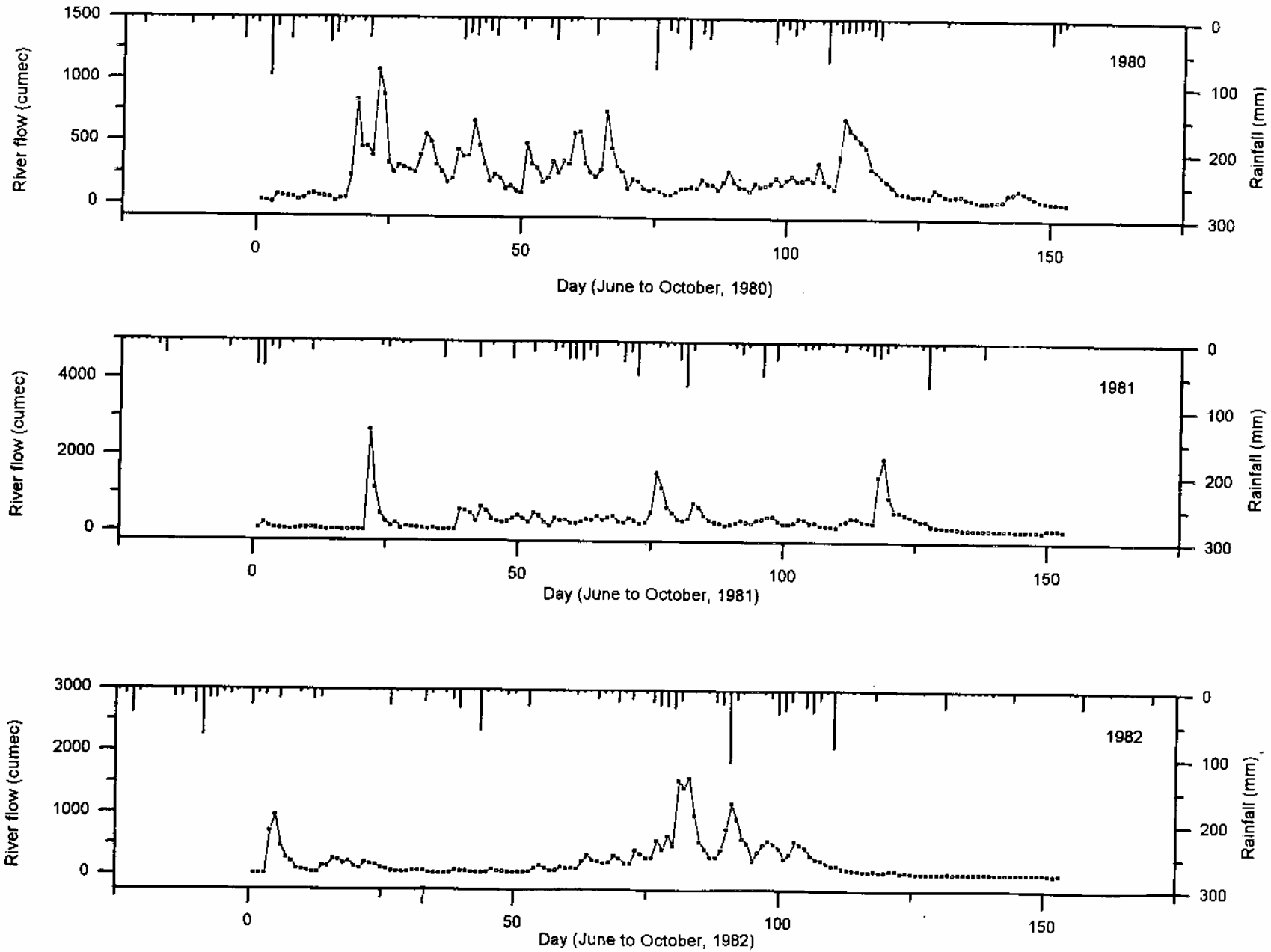


Fig 2.4 Weighted daily rainfall and recorded flow at Anandpur gauge discharge site

presented in Table 2.1. Using the above weights, and corresponding daily rainfall data of the stations, the average areal precipitation for the Anandpur catchment has been estimated. The weighted rainfall and recorded runoff (stream flow) data for the three years 1980-1982 is presented in Fig 2.4.

Chapter 3
Model Development

The need to predict river flow after heavy rains is important for public safety, environmental issues, and water management. One of the most accepted and widely used techniques for predicting future events is time series analysis. Time series refers to observations of a variable that occur in a time sequence. Time series techniques can be used to analyze the statistical behavior of a series of experimental or observed data over time, where these data have significant correlation induced by sampling of adjacent time points. Such models do not attempt to represent the non-linear dynamics inherent in the process, and therefore may not always perform well.

During heavy rain periods, some of the forms of hydrological balance (evaporation, infiltration and storage variations) can be neglected since they give no relevant contribution to the river flow rate in the short period. In contrast, accurate information on rainfall and on the state of the basin must be available. The rainfall gives a measure of the amount of water gathered by the basin and represents the perturbations experienced by the water system. The state of the basin, which is correlated, albeit directly, to the flow rate, represents the capability of the river systems to respond to rainfall perturbation. However, even in these simplified conditions, the usual approaches prove to be inefficient or too burdensome (Woolhiser, 1996).

In the hydrological context, as in many other fields, ANN are increasingly used as black box simplified models (Bishop, 1994). For hydrological applications, ANN models can take advantage of their capability to reproduce the unknown relationship existing between a set of input variables descriptive of the system, for example rainfall river flow rate (Chakraborthy et. al, 1992)

Artificial neural network

An ANN is an information processing system inspired by the way, the densely interconnected parallel structure of the mammalian brain processes information. ANN's are collection of mathematical models that emulate some of the observed properties of biological neuron systems and draw on the analogies of adaptive biological learning. Neurons in an ANN are arranged in to groups called layers. Each neuron in a layer operates in logical parallelism. Information is transmitted from one layer to others in serial operations (Hecht – Nielson, 1990). A network can be comprised of one to many layers. The basic structure of a network usually consists of three layers: the input layer, where the data are introduced to the network;

the hidden layer or layers, where data are processed; and the output layer, where the results of given input are produced.

Although ANN's have been around since the late 1930's it was not until the mid 1980's that algorithms became sophisticated enough for general application. Today, ANN's are being applied to an increasing number of real world problems of considerable complexity. The advantage of ANN's lies in their resilience against distortions in the input data and their capability of learning. They are often good at solving problems that are too complex for conventional technologies. There are multitudes of different types of ANN. The present study employed two types viz. back propagation neural network and radial basis function network. A brief description of both the structures are outlined below.

Back Propagation Network

Back propagation is the most widely used of the neural network paradigms and has been applied successfully in application studies in a wide range of areas. Several neural network models can be used in pattern recognition (both supervised and unsupervised). For supervised algorithm, the most commonly used ANN is the three layer feed forward network trained using the back propagation algorithm (Rumelhart and MeCleland, 1986; Jones and Hoskins, 1987) which is adopted in the present study.

The back propagation algorithm (BPN) involves a forward propagating step followed by a back propagating step. Both the forward and back propagation steps are done for each pattern presentation during training. The forward propagation step begins with the presentation of an input pattern to the input layer of the network, and continues as activation level calculations (activation level parameter associated with each processing unit is its out put value) propagate forward through the hidden layers. In each successive layer, every processing unit sums its inputs and then applies a transfer function to compute its output. The output layer of units then produces the output of the network. The backward propagation step begins with the comparison of the network's output pattern to the target vector, when the difference or "error" is calculated. The backward propagation step then calculates error values for hidden units and changes for their incoming weights, starting with the output layer and moving backward through the successive hidden layers. In this back propagating step, the network corrects its weights in such a way as to decrease the observed error. A general structure of the BPN is depicted in Fig 3.1.

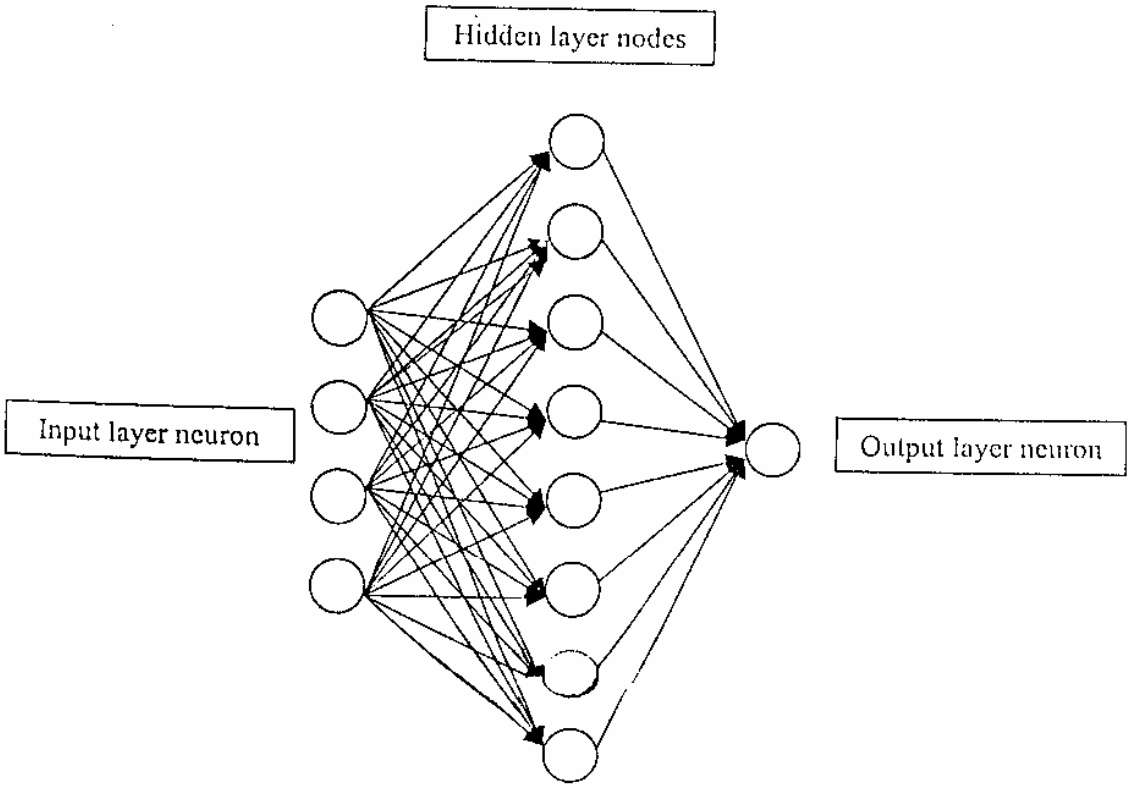
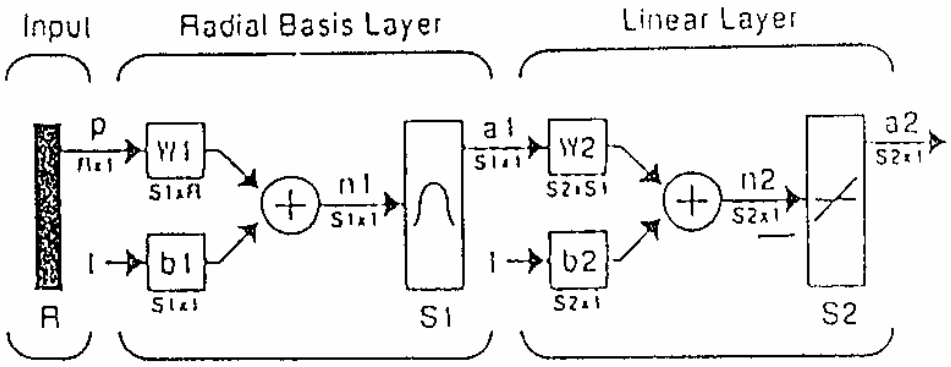


Fig 3.1 General back propagation network structure



Where
R = number of inputs
S1 = number of radial basis neurons
S2 = number of linear neurons

Fig3.2. Radial basis neural network architecture

The back propagation algorithm can be described in three equations. First, weight connections are changed in each learning step (k) with

$$\Delta W^{s}_{ij(k)} = \eta(t)\delta^{s}_{pj}x^{(s-1)}_{i} + m\Delta W^{s}_{ij(k-1)} \qquad (3.1)$$

Second, for output nodes it holds that

$$\delta^{o}_{pj} = (d_j - o_j)f'(I^{s}_{j}) \qquad (3.2)$$

and third for the remaining nodes it holds that

$$\delta^{s}_{pj} = f'_j(I^{s}_{j})\sum_{k}\delta^{(s+1)}_{pk}W^{(k+1)}_{jk} \qquad (3.3)$$

where,

$x_j^{(S)}$	=	actual output of node j in layer s;
$W_{jk}^{(S)}$	=	weight of the connection between node j at layer (s-1) and node k at layer s
$\delta_{pj}^{(S)}$	=	measure for the actual error of node j;
$I_{(j)}^{(S)}$	=	weighted sum of the inputs of node j in layer s;
$\eta(t)$	=	time dependent learning rate;
f( )	=	transfer function;
m	=	momentum factor (between 0 and 1) and
d_j and o_j	=	desired and actual activity of node j (for output nodes only)

Radial Basis Function Network

Radial Basis Function networks has a state Gaussian function as the non-linearity for the hidden layer processing elements. The Gaussian function responds only to a small region of the input space where the Gaussian is centered. The key to successful implementation of the network is to find suitable centres for the Gaussian functions. This can be done with supervised learning, but an unsupervised approach usually produces better results. For this reason, the present study employed as hybrid supervised – unsupervised topology for learning.

The most common idea in a hybrid learning procedure is to have one layer that learns in an unsupervised way, followed by one (or more) layers trained by back propagation. The network architecture examined by Moody and Darken (1989) has been employed in the study. The hidden units in the Moody-Darken network are neither linear, nor sigmoidal, instead they have normalized Gaussian activation functions of the form:

$$g_j(\varepsilon\ ) = \frac{\exp[-(\varepsilon - \mu_j)^2/2\sigma\ _j^2]}{\sum_k \exp[-(\varepsilon\ - \mu_k)^2/2\sigma_k^2]} \qquad (3.4)$$

where $g_i(\varepsilon)$ is the input vector itself. The Gaussians are a particular example of radial basis functions. Radial basis networks consist of two layers: a hidden radial basis function layer and an output linear neuron layer. The network architecture is presented in Fig 3.2.

The network functions as follows. Suppose a particular input vector ε^μ lies in the middle of the receptive field for unit j, so $\varepsilon^\mu = \mu_j$. If the overlaps between the receptive fields are ignored, only hidden unit j will be activated, making it the only "winner". One could simply choose the output weights leading from that unit to be $w_{ij} = \xi_i^\varpi$ (for each i), which will produce the target pattern ξ_i^μ at the output assuming linear output units. If another input lies say, between two receptive field centers, then those two hidden units will be appreciably activated and out put will be the weighted average of the corresponding targets. In this way the network is expected to make sensible smooth fit to the desired function.

The unsupervised part of learning is the determination of the receptive field centers μ_j and weights σ_{ij}. Appropriate μ_js can be found by any vector quantisation approach including the usual comflowitive learning algorithm (Hertz et. al., 1991). The σ_{ij}s are usually determined as ad hoc choice, such as mean distance to the first few nearest neighbor m's. The performance of the network is not very sensitive to the precise values of the σ_{ij}s.

Moody and Darken tried their method out on the extrapolation problem for the Mackey-Glass equation and found that the present method, with Gaussian receptive fields, allows one to fit an arbitrary function with just one hidden layer (Hartman, 1990).

The advaniage of the RBF network is that it finds the input to output maps using local approximators. Usually, the supervised segment is simply a linear combination of the approximators. Since linear combines have few weights, these network train extremely fast and require training samples.

Rainfall Runoff Modeling

The steps involved in the identification of a dynamic model of a system are (I) selection of input – output data suitable for calibration and validation, (ii) selection of a model structure

and estimates of its parameters, and (iii) validation of the identified model. This study compares the performance of three kinds of different model structures with respect to their ability to represent the rainfall runoff process. For the bulk of the study only monsoon season (June to October) data were used for calibration as well as validation, the data for the years 1980 and 1981 were used for calibration of ANN models. The models were validated for the year 1982.

ANN Model Identification

The ANN model structure is ideally suited for modeling highly non-linear input – output relationship such as these encountered in the transformation from rainfall to runoff. The main objective of the study was to use an ANN to predict the stream flow from available distributed rainfall and discharge data. Most of the previous work considered rainfall data averaged over the basin scale; this has the advantage of reducing the number of input variables to the network.

As reported by Minns and Hall (1996), rainfall information alone is not sufficient to compute flow rate, since the state of the basin plays an important role in determining flow rate behavior. For this reason, flow data at certain time intervals before the time of predictions have been used as additional input information to the network. In this way information about the state of the basin is introduced. The selection of the number or previous flow data as input to the network was done by statistical analysis as briefed below.

Standardization of time series

A time series may often contain periodic components that tend to repeat over a period of time intervals, due to astronomic cycles. The behavior of time series is known as a periodicity, which means that the statistical characteristics change periodically within the year. In order to develop any model, the periodic component must be removed from the time series. The periodic component can be removed as follows.

$$z_{v,\tau} = \frac{PET_{v\tau} - \mu_\tau}{\sigma_\tau} \qquad (3.5)$$

where $z_{v\tau}$ is the standardized flow; v is the year; τ is the time interval within the year; and μ_τ and σ_τ are respectively the population periodic mean and standard deviation of the flow time series. The standardization method preserves the first two moments (mean, standard deviation) of the historical series. The sample periodic means and standard deviations can be estimated from the observed time series for each day of the growing season, and can be

substituted in equation to obtain the standardized flow series. Salas et.al (1988) mentioned that the sample estimates of means and standard deviations are subjected to larger errors, since they are usually estimated from a relatively small number of year's data, as compared to the population estimates. Also, the use of too many-estimated parameter violates the principle of statistical parsimony in the number of parameters. In the above time series, for example, the number of estimated parameters would be 153 means and standard deviations (one for each day in the monsoon season) plus the model parameters. To reduce the number of estimated parameter and to obtain better estimates of these parameters, they suggested the use of a Fourier Series for the estimation of periodic parameters. In the light of this suggestion, estimates of periodic parameters were obtained in this study by using the Fourier Series. The method is described as follows:

Let u_τ represent a periodical statistical characteristic of the flow series, such as the daily mean or standard deviation. Also assume u_τ is a sample estimate of the unknown population periodic parameter denoted by v_τ. The population periodic parameter estimate, $\hat{v}_\tau$ can be obtained by:

$$\hat{v}_\tau = \bar{u} + \sum_{j=1}^{h}\left[A_j \mathrm{Cos}(2\pi j\tau/\omega) + B_j \mathrm{Sin}(2\pi j\tau/\omega)\right] \qquad \tau = 1,2,....,\omega \tag{3.6}$$

where $\bar{u}$ is the seasonal mean of u_τ, A_j, B_j are the Fourier series coefficients, j is the harmonic, and h is the total number of harmonics, which is equal to $\omega/2$ or $([\omega - 1]/2)$ respectively, depending if ω is even or odd. The mean $\bar{u}$ and the Fourier coefficients A_j and B_j can be determined by:

$$\bar{u} = \frac{1}{\omega}\sum_{\tau=1}^{\omega} u_\tau \tag{3.7}$$

$$A_j = \frac{2}{\omega}\sum_{\tau=1}^{\omega} u_\tau \mathrm{Cos}(2\pi j\tau/\omega) \qquad \text{for } j = 1,\ldots, h \tag{3.8}$$

and

$$B_j = \frac{2}{\omega}\sum_{\tau=1}^{\omega} u_\tau \mathrm{Sin}(2\pi j\tau/\omega) \qquad \text{for } j = 1, \ldots, h \tag{3.9}$$

When $\hat{v}_\tau$ from equation (3.6) is determined considering all the harmonics j=1,.....h) (All the coefficients A_j and B_j), $\hat{v}_\tau$ is exactly the same as u_τ for all the values of τ =1,....,ω. However,

this approach will neither reduce the estimation error nor the number of estimated parameter. To achieve this, smaller number of harmonics h * < h is used, such results are still significant.

For the historical flow series, the daily sample means and standard deviations were calculated for the monsoon season and are shown in Figures 3.4 and 3.5 respectively. Daily mean flow was lower at the beginning of the season as compared to that of the mid of the season and, showed an increasing trend throughout the season. This trend was also found in the historical flow series (Fig 3.3). The daily standard deviation values for the growing season showed a larger variation for the first month as compared to the rest of the season. These sample estimates were utilized to perform the Fourier series analysis as described above to obtain an estimate of the daily means and standard deviations.

As mentioned earlier, the Fourier series fit procedure requires the selection of the number of the significant harmonics. Salas et. al (1988) provide a procedure for selecting the number of significant harmonics by plotting the periodogram. However, this procedure added too many harmonics to the function (Aboitiz et. al., 1986). Thus the selection of the number of significant harmonics was done by visual inspection of the resulting function. The number of selected harmonics, h*, was chosen by plotting the periodic parameters and the Fourier series function for several values of h*, and inspecting these plots. As climatic conditions should not change drastically in the basin from day to day over the season, it can be expected that the population daily mean and standard deviation will be reasonably smooth function over time. Therefore, the value of h* selected was that which produced a smooth function without much fluctuations. The Fourier fit of the daily means and standard deviations for different harmonics considered are presented in Figs 3.6, 3.7 and 3.8.

Table 3.1 Parameters of Fourier series models for daily mean and standard deviation

Parameters	Mean	Standard deviation
Seasonal mean, $\bar{u}$	330.03	382.86
Seasonal variance	203.35	355.37
Fourier coefficients		
A_1	-231.43	-227.54
B_1	6.98	53.15
A_2	-38.96	-87.42
B_2	1.61	-11.62
A_3	-44.11	-114.1
B_3	16.41	-1.46
Overall explained variance (% of total)	67.99	28.59

The Fourier series model with three harmonics fitted well to the periodic mean, except at the beginning and at the end of the season (Fig 3.7). The values of Fourier coefficients for both the parameters are depicted in Table 3.1. The model explained about 67% of the variance in the sample mean series (Table 3.1). In case of the periodic standard deviation, the Fourier

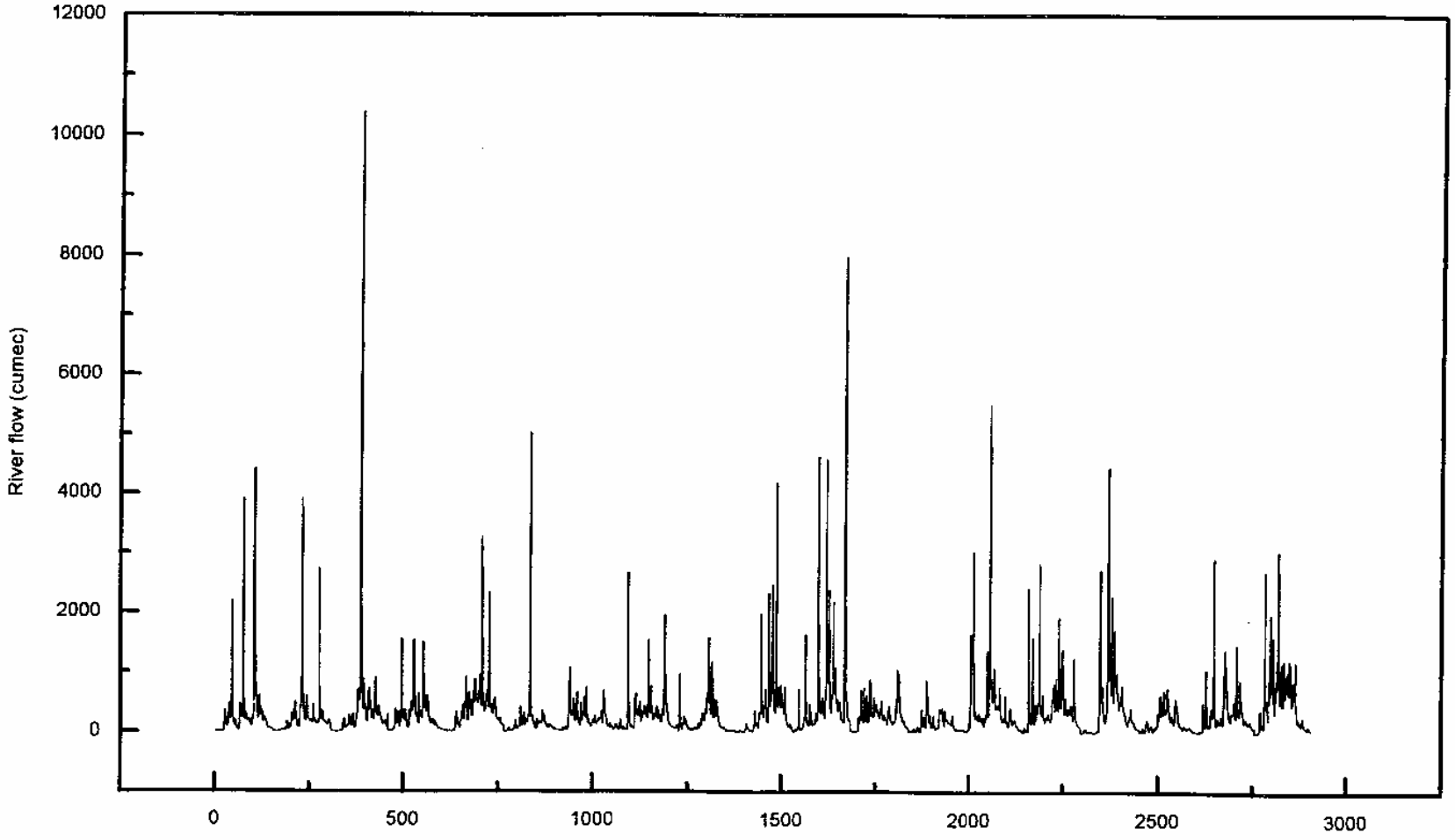


Fig 3.3 Historical flow series for the years 1972-1994

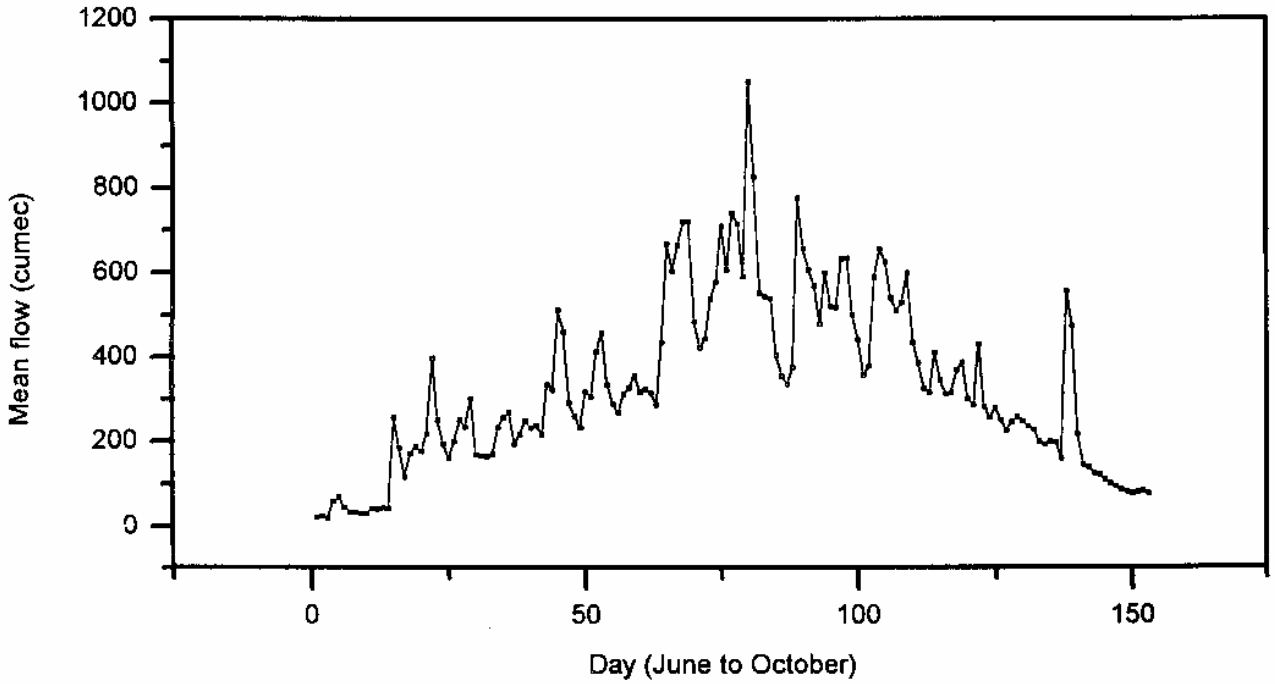


Fig 3.4 Plot of sample mean flow derived from 22 year data (monsoon season)

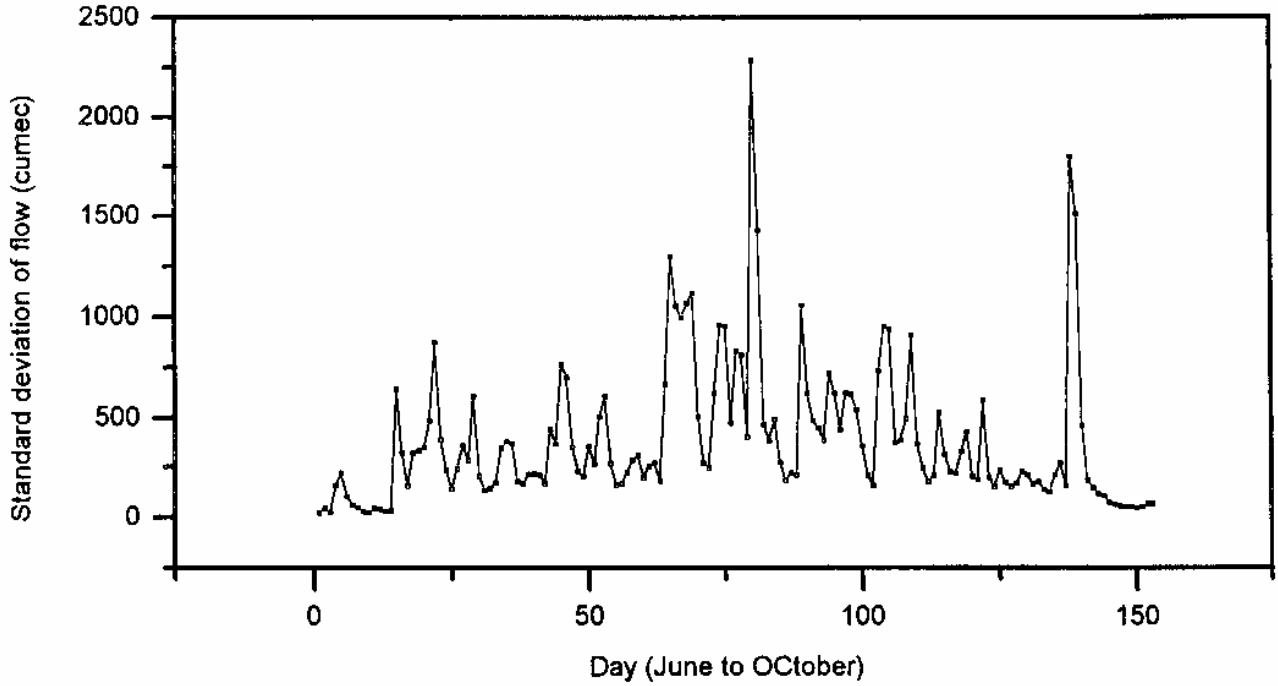
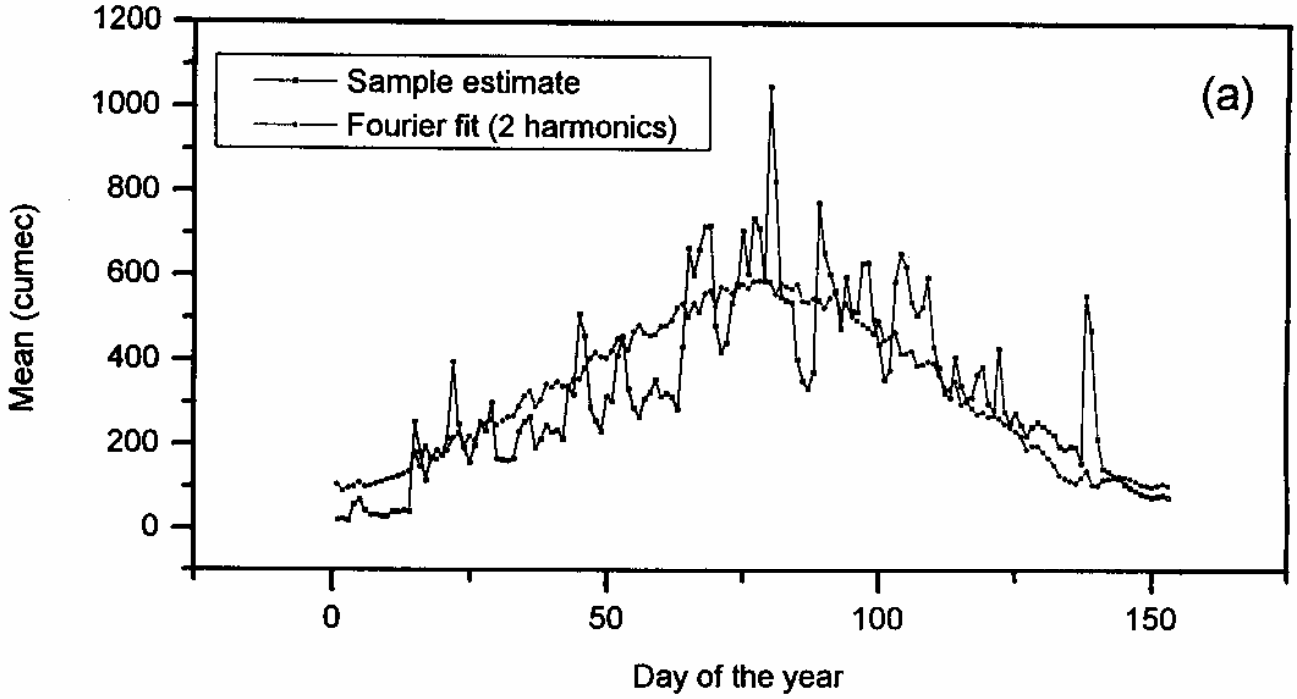


Fig 3.5 Plot of sample standard deviation of flow derived from 22 year data (monsoon season)



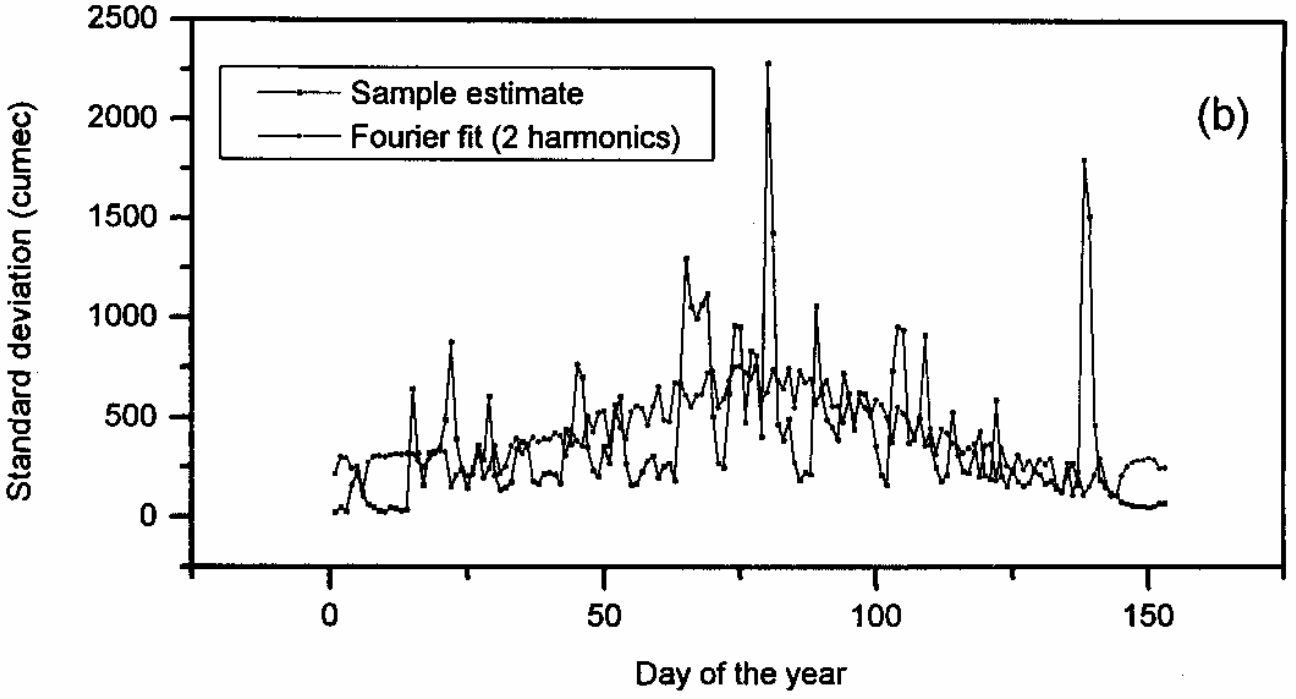
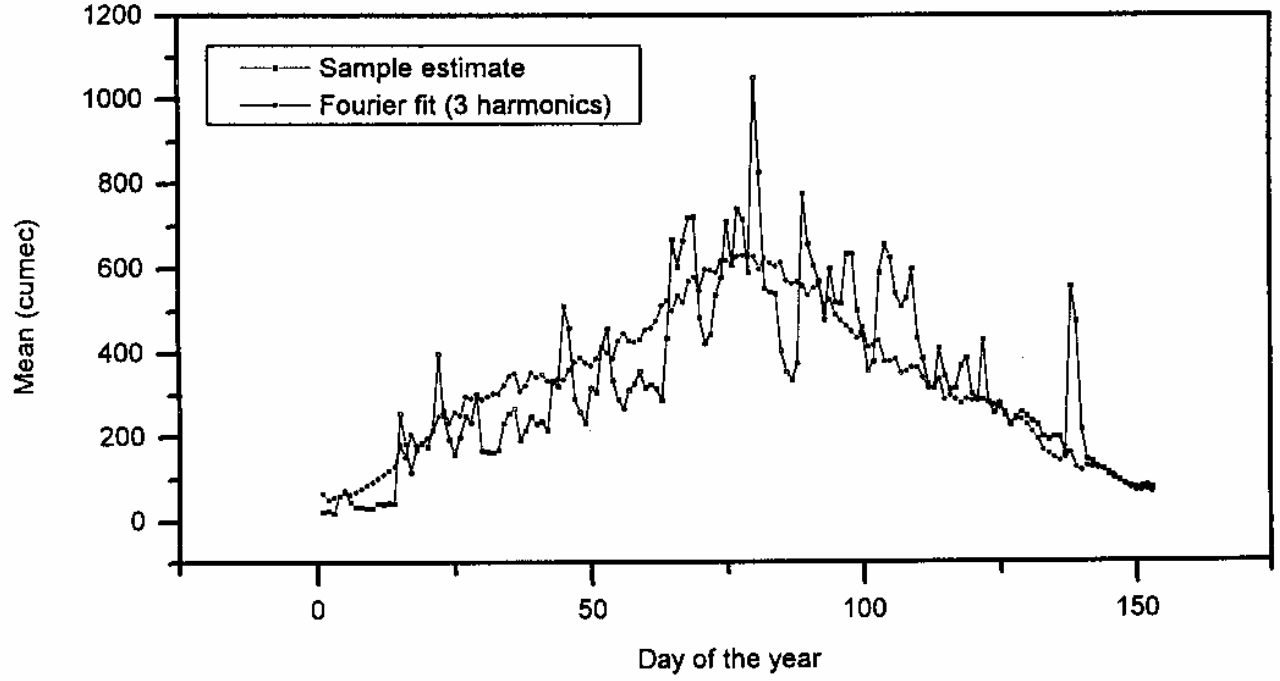


Fig 3.6 Plot of fourier fit with 2 harmonics
(a) for mean flow
(b) for standard deviation



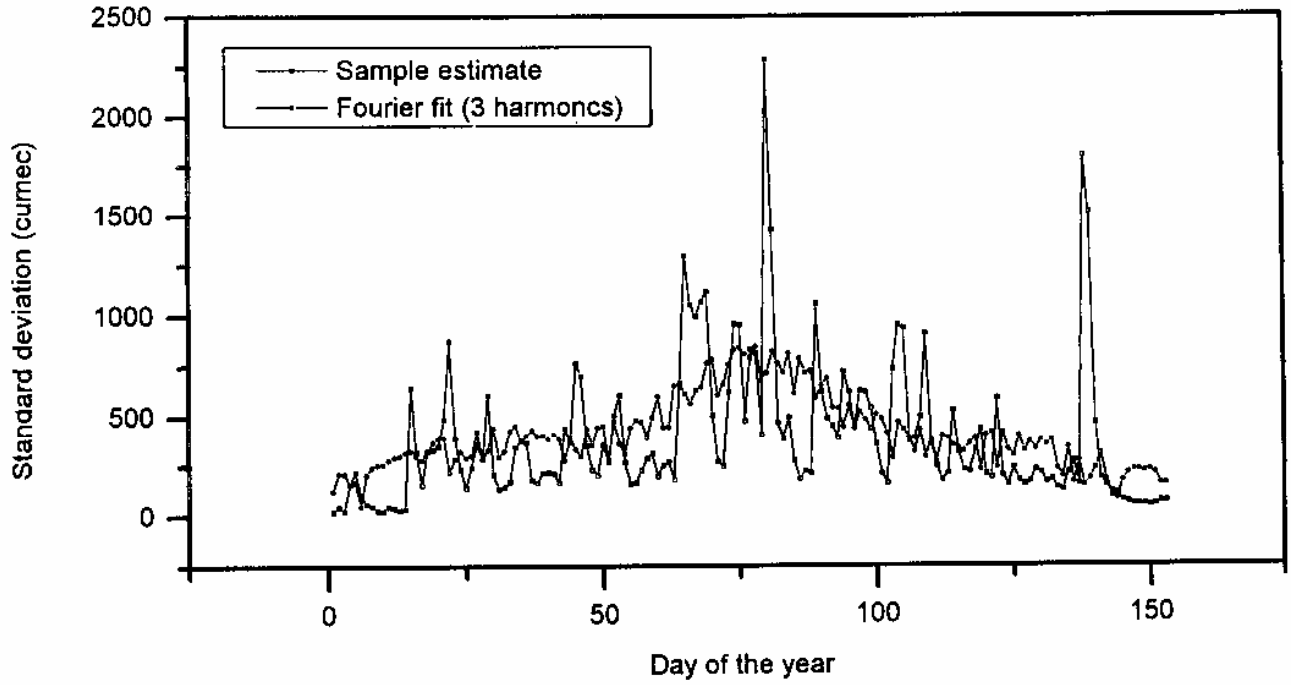
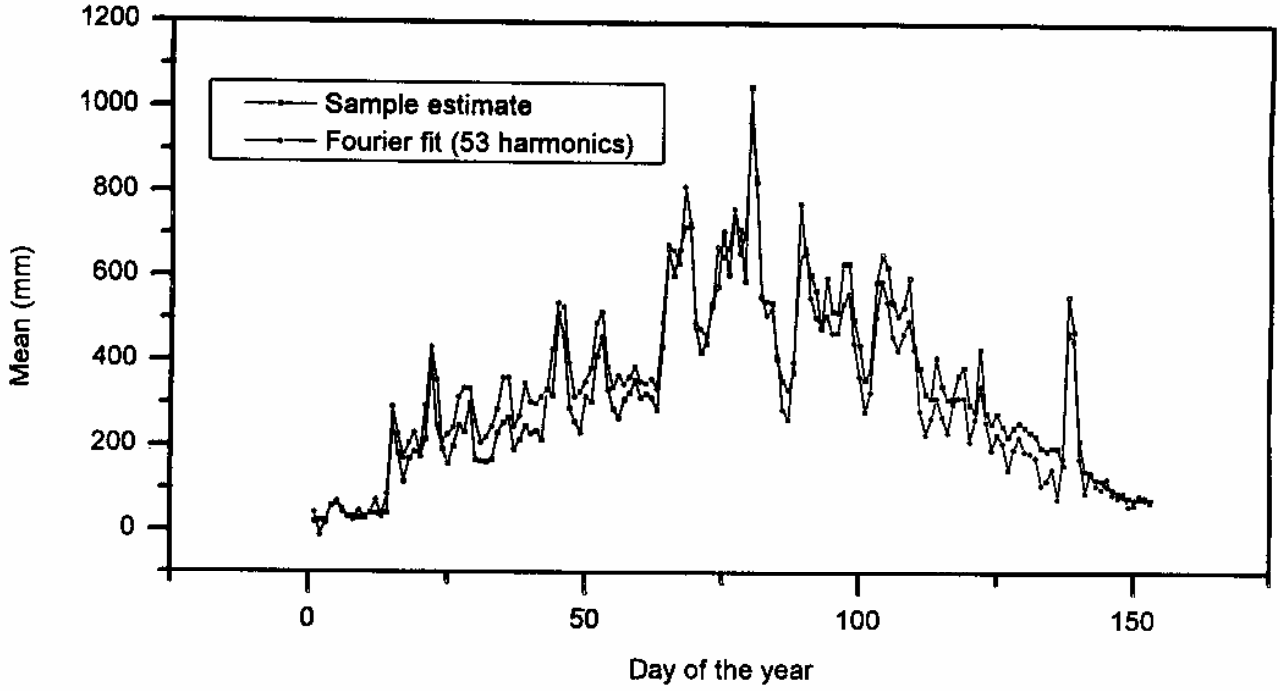


Fig 3.7 Plot of fourier fit with 3 harmonics
(a) for mean flow
(b) for standard deviation



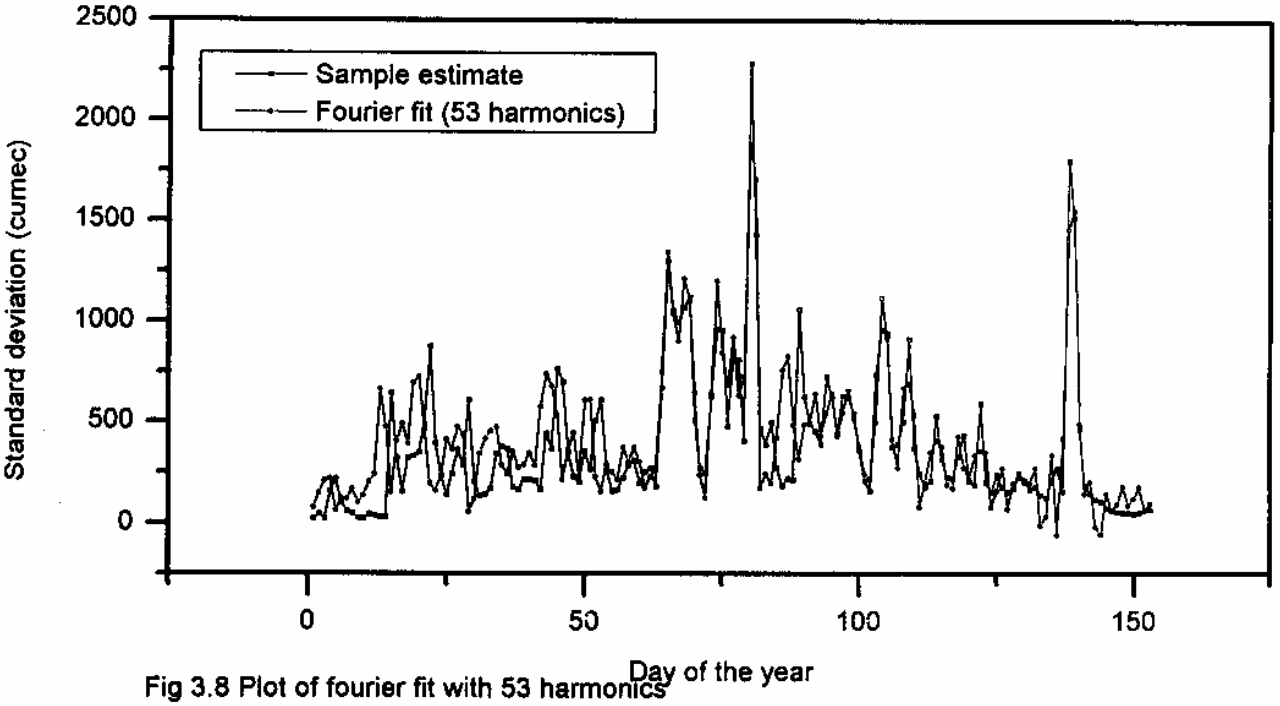


Fig 3.8 Plot of fourier fit with 53 harmonics
(a) for mean flow
(b) for standard deviation

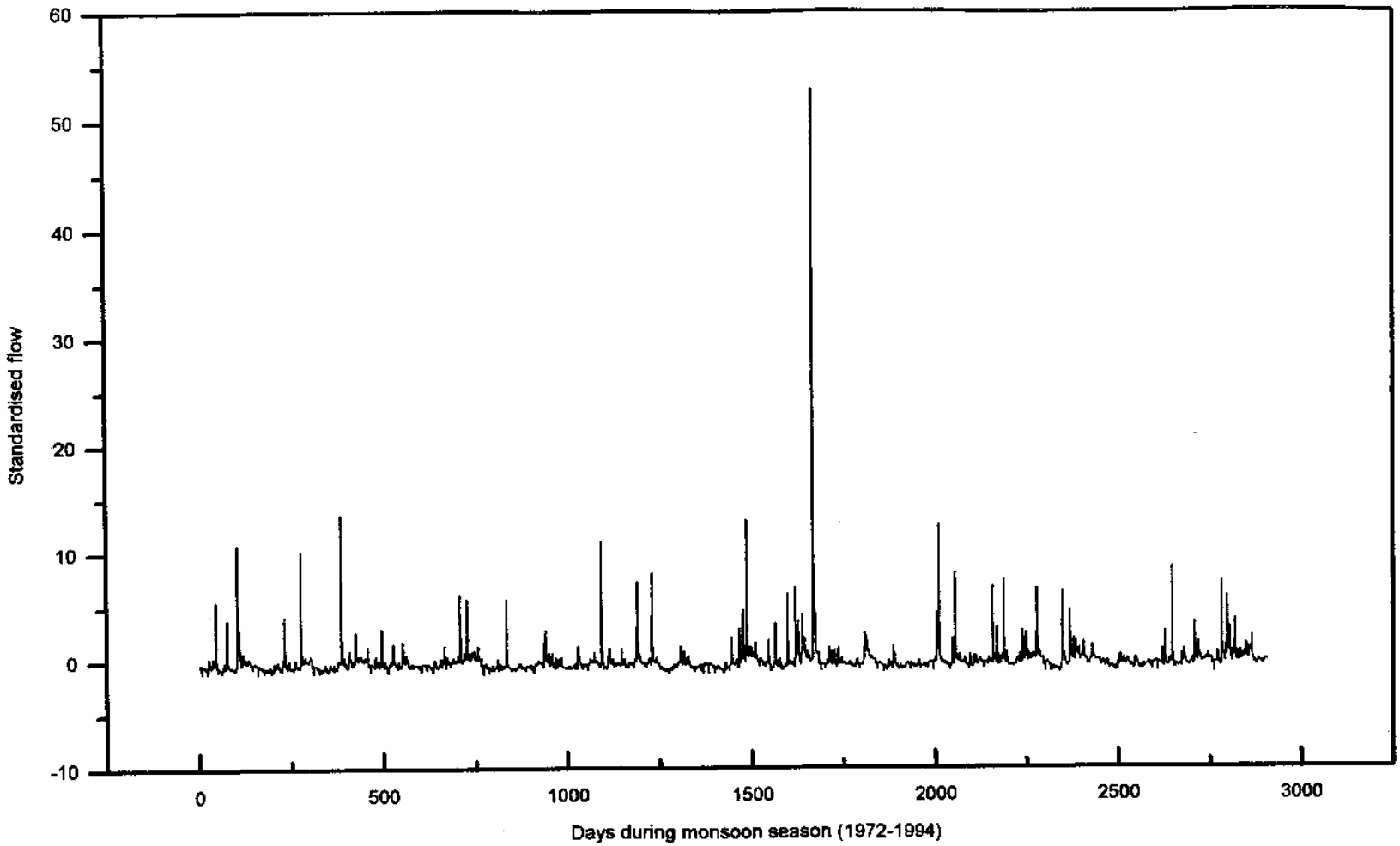


Fig 3.9 Standardized flow series for the years 1972-1994

series fit showed similar fluctuations as that of the periodic standard deviation series, Fig 3.6, but was only able to explain about 28% of the variance, Table 3.1. However, for both the mean and standard deviation, the fitted models resulted in smooth functions, which can be expected with a large sample size.

Estimates of the periodic mean and standard deviation obtained from the fitted Fourier series models were utilized to obtain the standardized flow series using equation 3.1. The standardized flow series is presented in Fig 3.9. The mean and standard deviation of the resulting standardized series were found to be 0.0 and 1.63 cumec, respectively, which are sufficiently close to the theoretical values (0.0 and 1.0 respectively).

Identification of the Input Vector

Identification of number of flow series in the input vector is determined by means of sample auto correlation and partial auto correlation functions. These functions reveal the correlation structure of the time series and, thus, are helpful in determining the underlying stochastic process. The theory is based on the assumption of second order stationarity. The assumption can be explained by letting (z_t, z_{t+h}) be a pair of flow measurements at t and t+h in time separated by a vector h (lag). Each z_t is a realization of the random variable (Z_t, t within the time domain of interest) is called a random function and is said to be second order stationary if:

(i) the expected value $E(Z_t)$ exists and is the same within the time domain:

$$E(Z_t) = m, \tag{3.10}$$

(i) the covariance for each pair of random variables (z_t, z_{t+h}) exists, is the same in time, and depends on h,

$$Cov(h) = E[z_t . z_{t+h}] - m^2 \tag{3.11}$$

Stationarity of the covariance implies stationary of the variance.

Autocorrelation function

The autocorrelation function expresses the degree of dependency among neighboring observations. It is a process of self-comparison expressing the linear correlation between an equally spaced series and the same series at a specified lag.

Let z_0, z_1 , z_2 , ...z_{N-1} be a realization of a stationary stochastic process, then the population autocorrelation function can be defined as the quotient of the population autocovariance, cov (z_t, z_{t+h}) and variance, var(z_t):

$$\rho(h) = \frac{\mathrm{cov}(z_t, z_{t+h})}{\mathrm{var}(z_t)} \qquad (3.12)$$

where z_t is the value of variable at the t^{th} time, and h is the time lag. Since, the series analyzed is just one particular realization (out of an infinite number of realizations) of a stochastic process produced by the underlying probabilistic mechanism, the population autocorrelation function (Eq 3.12) can be estimated using the simple autocorrelation function, r(h):

$$r(h) = \frac{\sum_{t=1}^{N-h}(z_{t+h} - \bar{z})(z_t - \bar{z})}{\sum_{t=1}^{N}(z_t - \bar{z})^2} \qquad -1 \le r(h) \le 1 \qquad (3.13)$$

where $\bar{z}$ is the sample mean. The 95% of confidence band for sample autocorrelation functions given by Anderson and Jenkins, 1970:

$$r(h) = 0 \pm \frac{1.96}{\sqrt{n}} \left\{1 + 2\sum_{j=1}^{q} r_j^2\right\}^{1/2} \qquad h > q \qquad (3.14)$$

where q is the order of the process and n is the number of observation in the series. The autocorrelation function is a diagnostic of the moving average process. These processes do not have any dependence. Therefore, the value of a variable at a given time can be estimated from a purely random series using the weighted sum of the values at previous time steps.

Partial autocorrelation function

The partial autocorrelation function is another way of representing the time dependence structure of a series or of a given model. It is useful for diagnosing the order of autoregressive processes. The autoregressive process has a time previous time steps. Therefore, the idea of autocorrelation, which measures the correlation of variables separated by assigned lags, can be extended to that of the correlation where dependence on the intermediate terms has been removed. Mathematically, it can be defined as:

$$\phi_k(k) = \mathrm{corr}(z_t, z_{t-k} / z_{t-1}, \ldots, z_{t-k+1}) \qquad (3.15)$$

and is the correlation between z_t and z_{t-k} excluding the effects of $z_{t-1}, z_{t-2}, \ldots, z_{t-k+1}$. In this equation k is the distance or time lag measured between the measured quantities. In general, for an autoregressive process of order k, the partial autocorrelation coefficient, $\phi_k(k)$, is a measure of the linear association between ρ_j and ρ_{j-k} (auto correlation function at lag j and j-k) for $j \leq k$. It is the k^{th} autoregressive coefficient and $\phi_k(k)$, for k =1,2,... , is the partial autocorrelation function. Lag j autocorrelation for an Auto Regressive [AR(k)] process can be written as:

$$\rho_j = \phi_1(k)\rho_{j-1} + \phi_2(k)\rho_{j-2} + \ldots + \phi_k(k)\rho_{j-k}; \qquad j = 1,2,\ldots\ldots,k \tag{3.16}$$

where $\phi_j(k)$ is the j^{th} auto regressive coefficient of the AR(k) model. Eq 3.16 constitutes the set of linear equations, which can be written in terms of sample partial auto correlation functions $\phi_k(k)$, as:

$$\begin{bmatrix} 1 & r_1 & \cdots & r_{k-1} \\ r_1 & 1 & \cdots & r_{k-2} \\ \vdots & \vdots & & \vdots \\ r_{k-1} & r_{k-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_1(k) \\ \phi_2(k) \\ \vdots \\ \phi_k(k) \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix} \tag{3.17}$$

Thus, the sample partial auto correlation function can be obtained by solving Eq 3.17. Bartlett (1946) gave the 95% confidence band for the sample partial auto correlation function as,

$$\phi_k(k) = 0 \pm \frac{1.96}{\sqrt{n}} \tag{3.18}$$

where n is the number of observations in a series.

The auto correlation function (ACF) and the corresponding 95% confidence bands from lag 0 to lag 16, (0 to 16 days) were estimated for the standardized flow series using Eq 3.13 and 3.14 respectively. The results are shown in Fig 3.10. Lag 0 auto correlation is always is unity as it is correlation of the variable with itself. However, as the lag increases the correlation between the variable and the same variable at specified lag decreases, i.e., covariance decays. The auto correlation function showed significant correlation, at 95% confidence level, up to lag 7 (7 day), and thereafter, fell below the confident band. The gradual decaying pattern of auto correlation exhibits the presence of a dominant auto regressive process.

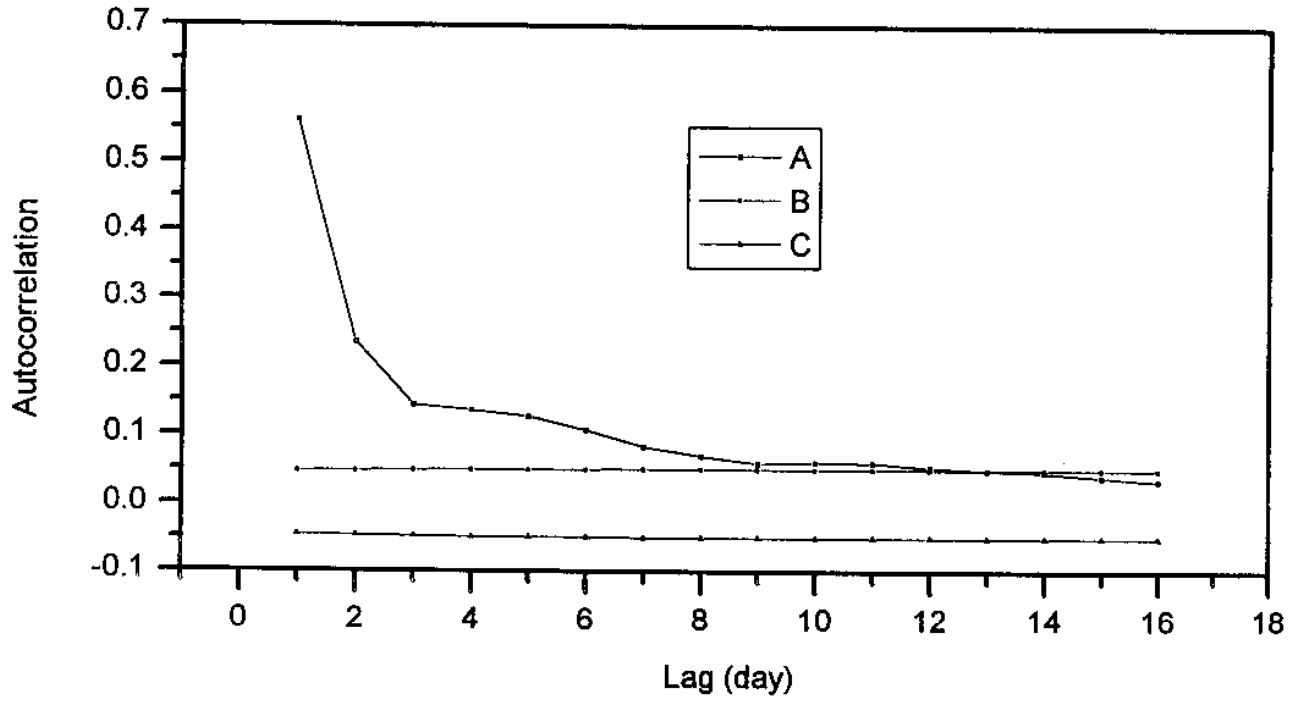


Fig 3.10 Autocorrelation function of standardized flow series

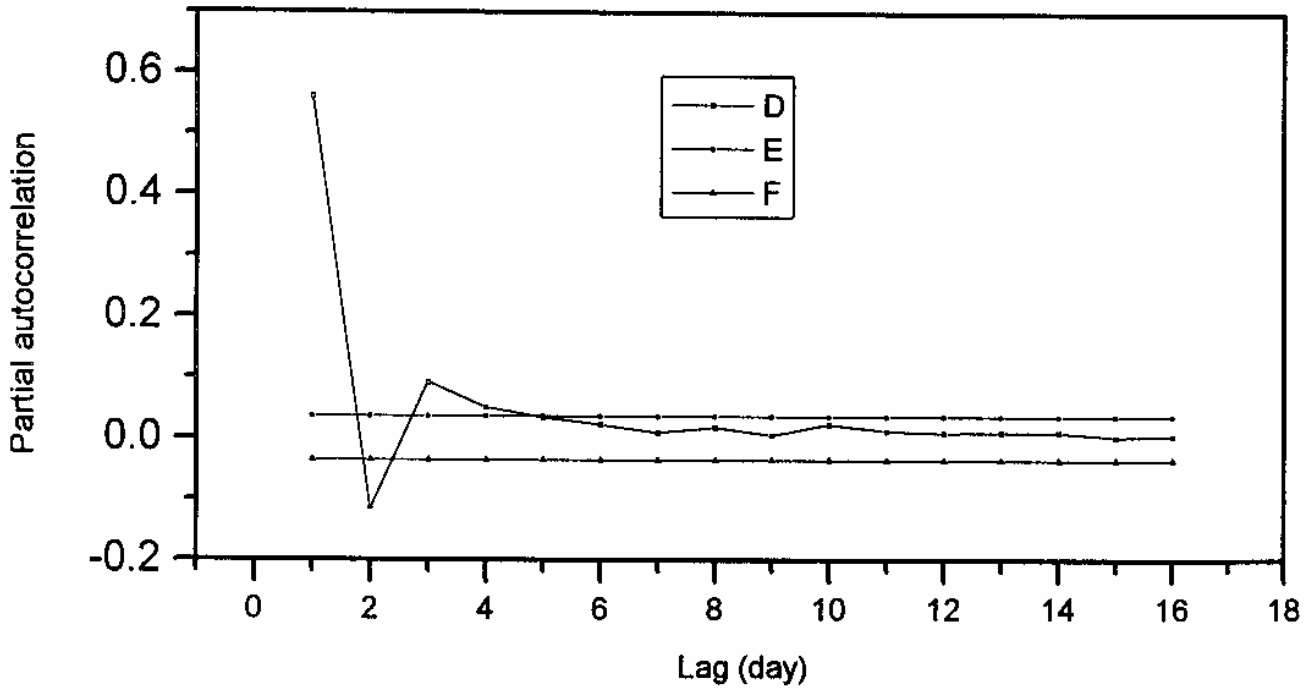


Fig 3.11 Partial autocorrelation function of standardized flow series

Similarly the partial autocorrelation function (PACF) and corresponding 95% confidence were estimated for lag 0 to lag 16 using eq 3.17 and 3.18 respectively. These are shown in Fig 3.11. The PACF showed significant correlation at lag 4 (4 day) and thereafter fell below the confidence band. The rapid decaying pattern of the PACF confirms the dominance of auto regressive process, relative to the moving average process. The above analysis of auto and partial correlation coefficients suggested incorporating flow values with 3 days lag in the input vector to the network.

Number of rainfall patterns in the input vector

The number of previous day's rainfall which influence the flow rate to be predicted was determined in a trial and error manner. The procedure that was used to identify the number of rainfall patterns as input to the network is summarized below.

A sample model was selected by representing stream flow at the present time, 't' as a function of precipitation at (t-1) and stream flow at t-1, t-2, t-3 as desired from the statistical analysis. The model can be represented as

$$Q(t) = f(P_{t-1}, Q_{t-1}, Q_{t-2}, Q_{t-3}) \qquad (3.19)$$

Various ANN configurations were trained and tested using the model. The numbers of neurons in the hidden layer of BPN were varied from one to as many as 25 during training. Among the network trained, the best-fit network was selected based on the goodness of fit statistics of training and testing. The precipitation at time (t-2) was added as an additional input variable to the above model. Hence the model becomes

$$Q(t) = f(P_{t-1}, P_{t-2}, Q_{t-1}, Q_{t-2}, Q_{t-3}) \qquad (3.20)$$

The goodness of fit statistics for the present model were computed for training and testing procedure and compared with those for the best fit model at the previous step. If the goodness of fit statistics of the present model were significantly different from the previous model, then the precipitation at time (t-3) was added as another input to the present model. This procedure was repeated by adding precipitation at previous time periods as input variables until there was no significant change in model training and testing accuracy based on the goodness of fit statistics.

Goodness of fit statistics

For each model, fit to the training and testing data was done using popular residual statistics: the root mean square (RMSE), the Akaike information criterion (Akaike, 1974) and the

Bayesian information criteria (Rissanen, 1978). The AIC and BIC are computed using the following equation (Shumway, 1988).

$$AIC = \ln(RMSE) + \frac{2n}{N} \tag{3.21}$$

$$BIC = \ln(RMSE) + \frac{n\ln(N)}{N} \tag{3.22}$$

Notice that while RMSE statistics are expected to progressively improve as more parameters are added to the model, the AIC and BIC penalize the model for having more parameters and therefore tend to result in more parsimonious models.

Another index used to evaluate the goodness of fit of the model was the efficiency of the model defined by Nash and Sutcliffe (1970),

$$\text{Efficiency} = 1.0 - \frac{\sum(Q_o - Q_c)^2}{\sum(Q_o - \overline{Q})^2} \tag{3.23}$$

where, Q_o, Q_c are the observed and computed values of flow and $\overline{Q}$ is the mean flow over the period.

The model efficiency can be used to evaluate the capability of the model in predicting the next day river flow, different from mean value, which is assumed to be the prediction, however available, in the worst case. The results of the above analysis are presented in the next chapter.

Chapter 4
Results and Discussion

The main objective of the study was to develop a rainfall runoff model for the Baitarani river basin, Orissa, which would be able to forecast the stream flow using historic time series data of rainfall and runoff. A critical examination of the research work in this field suggested that ANN algorithms were capable of modelling the rainfall-runoff relationship due to its ability to generalize the patterns in noisy and ambiguous input data without a priori knowledge of probability distributions. Therefore an ANN approach has been employed in the study, as described in detail in chapter 3. The details of the basin and its characteristics are presented in chapter 2. The present chapter deals with the result pertaining to the work as the performance of the models considered, inter comparison of their performance, and relative performance of the best-fit model over existing models used in the basin.

Identification of the input vector to the network

To create any rainfall-runoff model by system theoretic approach, such as ANN, it is required to determine from the available historical sequences of rainfall and runoff data, the choice of how many and which delayed runoff patterns and rainfall patterns affect the next output. This is one of the complexities which make the forecast more difficult than the simple straight regression analysis. Conducting auto and partial correlation analysis of the river flow and determining the lags that have significant effect on the next flow can contain this complication. In the present study, this analysis was carried out, and is described in the chapter 3. However, the number of previous rainfall pattern that are having significant effect on the next day flow has been identified in a trial and error procedure, as described in the methodology (chapter 3).

The study considered two ANN structures based on the algorithm with which it learns the patterns. The two algorithms considered are back error propagation (usually called as back propagation network or BPN) and radial basis function network. The details of functioning of these two algorithms are described in chapter 3. While determining the input vector to the network, both these algorithms were considered. The trial started with presenting input vector to the networks considered, which consisted of one previous day rainfall value and estimating the goodness of fit statistics. The trial was continued with adding one more day (rainfall at lag of two days) in the input vector. The performance of the new input vectors was examined based on the statistical indices.

An abstract of the effect of rainfall lags in input vector is presented in Table 1. The table contains the goodness of fit statistics of three candidate models, which are described in detail below. These candidate models were selected based on their performance as examined by the goodness of fit statistics.

Table 4.1 Goodness of fit statistics for effect of number of previous day rainfall on input vector

		RMSE			Efficiency		
		Training		Validation	Training		Validation
		1980	1981	1982	1980	1981	1982
BPN with 6	P(t-1)	1.82E-02	2.78E-02	6.76E-02	81.63	91.96	30.38
Neurons	P(t-1, ..t-5)	1.25E-02	1.21E-02	4.26E-02	91.38	98.47	72.35
Radial Basis	P(t-1)	1.87E-03	0.00	6.52E-02	99.81	99.26	35.32
Network	P(t-1, ..t-5)	5.76E-03	6.06E-03	2.20E-02	98.16	99.62	92.64
BPN with 12	P(t-1)	1.72E-02	2.66E-02	5.73E-02	83.53	92.65	50.04
Neurons	P(t-1, ..t-5)	9.59E-03	9.11E-03	4.86E-02	94.90	99.14	64.08

Note: P(t-1) corresponds to input vector with one previous day rainfall
P(t-1,...t-5) corresponds to input vector with 5 previous day rainfall

The RMSE error has improved when the number of rainfall patterns in the input vector to the network increased from one to five, as can be seen from Table 4.1. Though this is the case with all models considered (during training as well as validation), there is a slight deterioration in the case of radial basis network during training. However, the RMSE value has been improved during validation. This may be due to the nature of the radial basis function, that the network reproduces the training pattern with least error always. The lower value of the RMSE in the case with only one previous day rainfall pattern in the input vector may be due to less number of variables in the pattern. The first case considered 4 variables in the pattern, while the second considered eight.

A similar trend is observed in the efficiency of the model too (Table 4.1). The table presents the calculated efficiency based on the normal values. All the models performed to a satisfactory level during training, as expected with any ANN architecture, but showed vide variations in the performance during validation. The radial basis network improved the efficiency from 35.32% to 92.15% with change in the input pattern during validation. The other models too performed in a similar manner during validation. These results lead to the conclusion that the number of previous rainfall data has a significant effect in the model performance. The experiment resulted in the conclusion that an input vector with 1,2,3,4 and 5 day lags can produce the river flow patterns in a satisfactory manner.

Back propagation network

Identifying the number of neurons in the hidden layer is another complicated task in finalizing the network architecture. This is commonly done by trial and error evaluation. The number of

hidden layers in the network was fixed to one as reported by Sridhar (1996). He, after experimenting with various combinations of ANN structure, reported that increasing the number of hidden layer to two or more have no significant effect in the performance of the network. The experiment starts with only one neuron in the hidden layer and increasing it by one every time after training the network and computing the goodness of fit statistics. The abstract of the goodness of fit statistics, on the effect of number of neurons in the hidden layer is presented in Table 4.2. The final selection of the number of neurons is made based on the statistical indices considered.

Table 4.2 Goodness of fit statistics for the effect of number of neurons in the hidden layer

	RMSE			Efficiency		
Number of neurons in the hidden layer	Training		Validation	Training		Validation
	1980	1981	1982	1980	1981	1982
6	1.82E-02	2.78E-02	6.76E-02	81.63	91.96	30.38
12	1.72E-02	2.66E-02	5.73E-02	83.53	92.65	50.04
25	1.35E-02	1.70E-02	6.28E-02	89.97	96.99	39.94

From Table 4.2, it can be observed that the RMSE error gets reduced as the number neurons in the hidden layer increases. During the experiment, it was observed that the improvement in RMSE value was not significant after 12 number of neurons in the hidden layer. The efficiency of all architecture was similar during training, but the performance got worsened during validation. The AIC and BIC values were considered for selecting the best-fit model from the trail runs. These values were computed using equation 3.21 and 3.22 respectively. Based on these criteria, a model with minimum value of AIC and BIC are to be selected. The AIC and BIC values for the models are presented in Table 4.3.

Table 4.3 The AIC and BIC values for selecting best fit model with different number of neurons in the hidden layer

	AIC			BIC		
Number of neurons in the hidden layer	Training		validation	Training		Validation
	1980	1981	1982	1980	1981	1982
6	-3.53638	-3.12642	-2.23612	-2.83075	-2.43318	-1.54288
12	-3.1883	-2.77881	-2.00981	-1.87785	-1.49137	-0.72237
25	-2.56376	-2.37541	-1.06825	0.05713	0.199467	1.50663

In the present study, two candidate models were selected based on their performance in representing the R-R process. The trial and error procedure resulted in selecting two candidate BPN models, one with 6 number of neurons in the hidden layer and the other with 12 number neurons (these models are represented as BPN 6N and BPN 12 N respectively in this report).

Both these models were trained and validated. The recorded daily values of rainfall and stream flow for the years 1980 and 1981 were used for training. The data for the year 1982 was used for validating the model. The input data was normalized using the following function (Romesburg, 1984):

$$Z_{ij} = \frac{x_{ij} - C_{minj}}{C_{maxj} - C_{minj}} \qquad (4.1)$$

where C_{maxj} and C_{minj} are the maximum and minimum of j^{th} variable in all observations. The main reason for standardizing the data matrix is that the variables are usually measured in different units. By standardizing the variables and recasting them in dimensionless units, the arbitrary effect of similarity between objects are removed.

The resulted hydrograph from both the models are presented in Fig 4.1 and 4.2 respectively for BPN 6N and BPN 12N respectively. A visual inspection of the observed and computed hydrograph supports the capability of ANN algorithm to represent the rainfall runoff transformation. However, the effectiveness of each model is to be understood through statistical analysis of the resulted hydrograph, and is described later in this chapter.

Radial basis function network

The trial and error evaluation of the network performance resulted in one candidate model with similar input vector as to that for BPN model (the model is represented by RBF in this report). The relative performance of the input vector variations can be examined, as extracted in Table 4.1. The optimum network structure was achieved by experimenting with various network parameters of the algorithm. This model was also trained using the data for the year 1980 and 1981, and validated for the year 1982.

The resulting hydrograph during training and validation are plotted in Fig 4.3. A simple visual analysis reveals that the model was able to perform satisfactorily, and is further justified by statistical analysis.

Inter-comparison of the candidate models

The performance of the three identified models for training as well as validation periods are critically examined using various statistical indices and are reported in Table 3.

The RMSE statistic measures the residual variance; the optimal value is 0.0. All the three models tend to have very small RMSE during training. The value of RMSE is found slightly

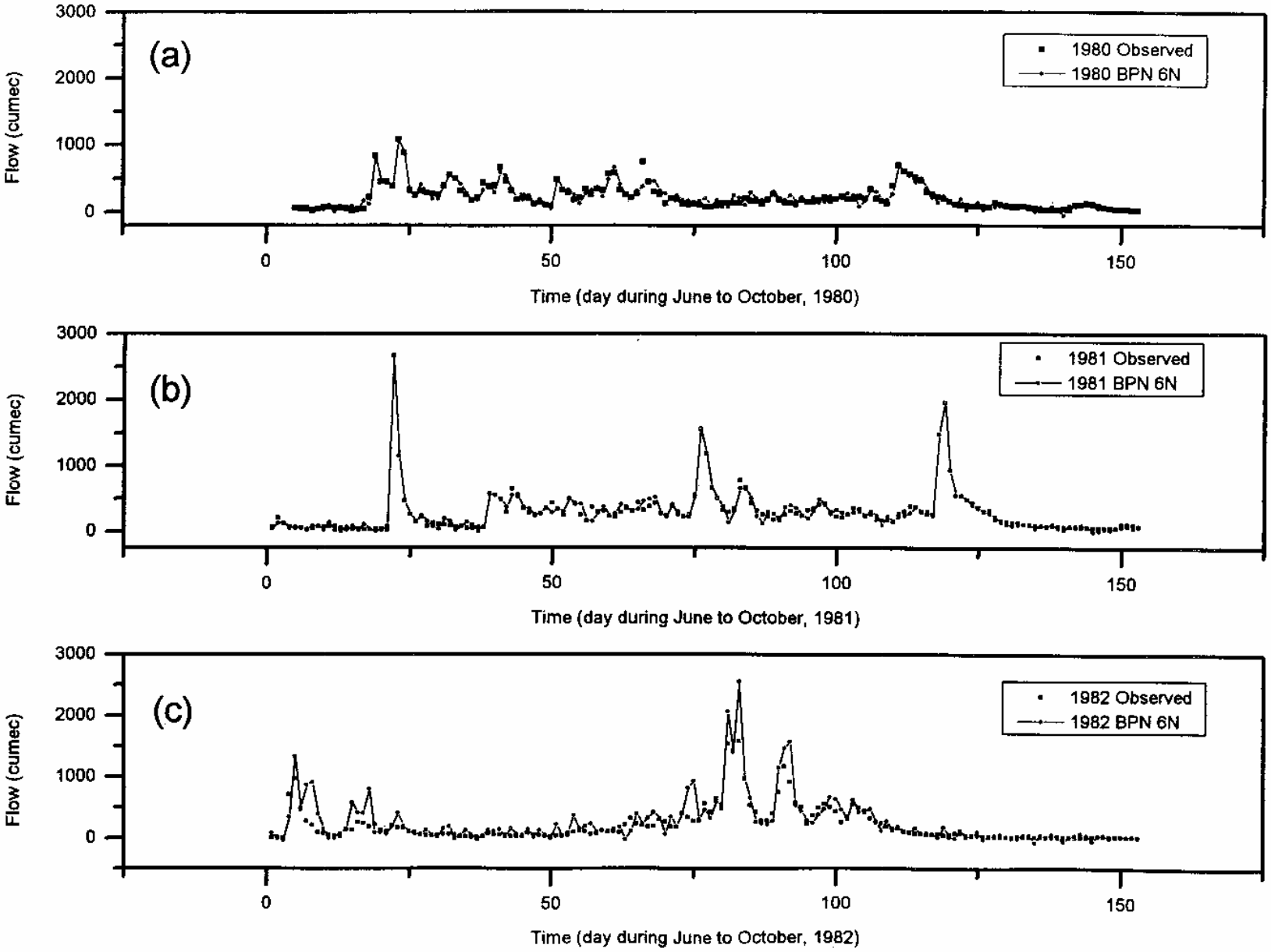


Fig 4.1 Observed and computed hydrographs using BPN 6N: (a) & (b) Calibration (c) validation

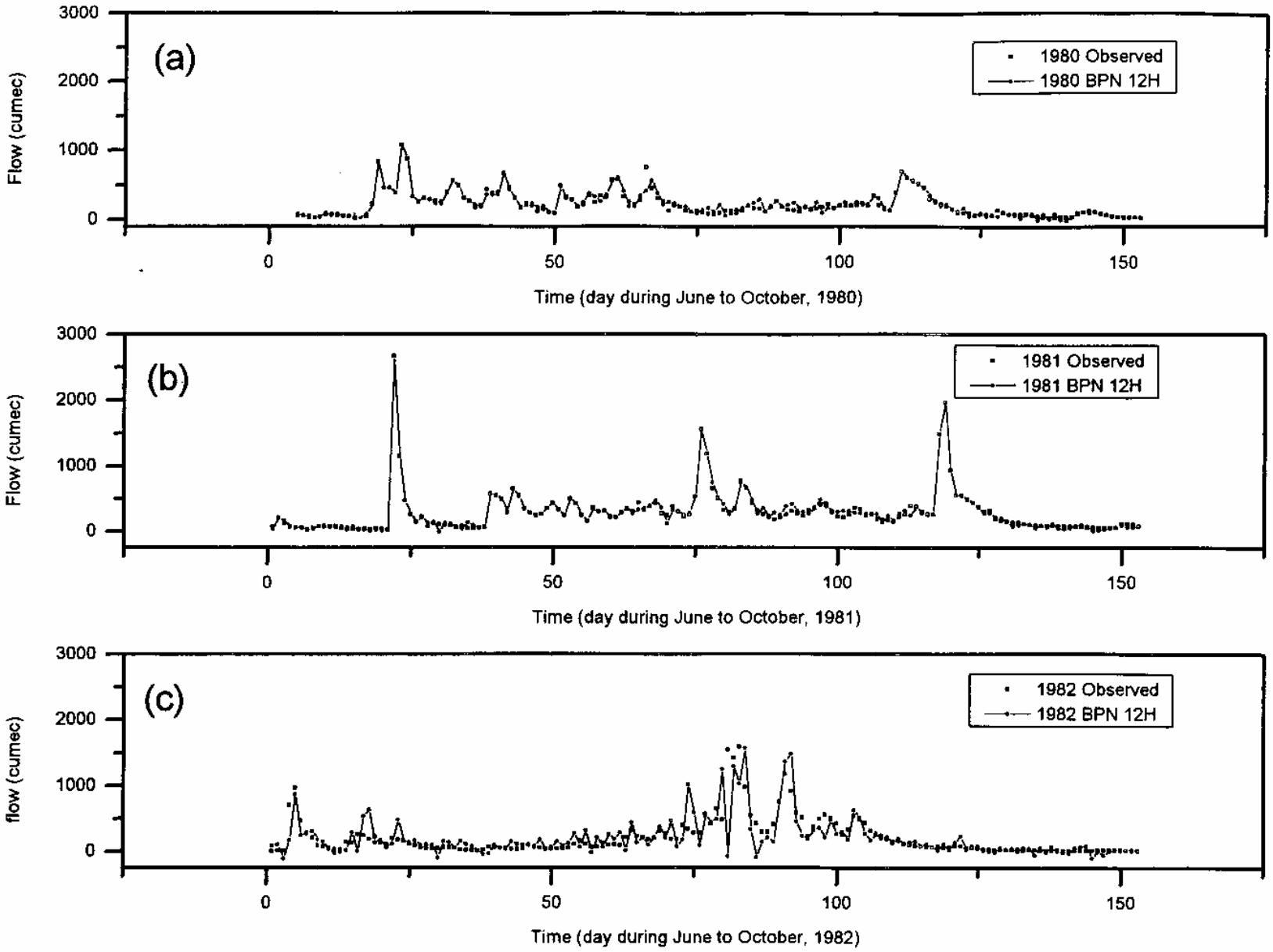


Fig 4.2 Observed and computed hydrographs using BPN 12N: (a) & (b) Calibration (c) validation

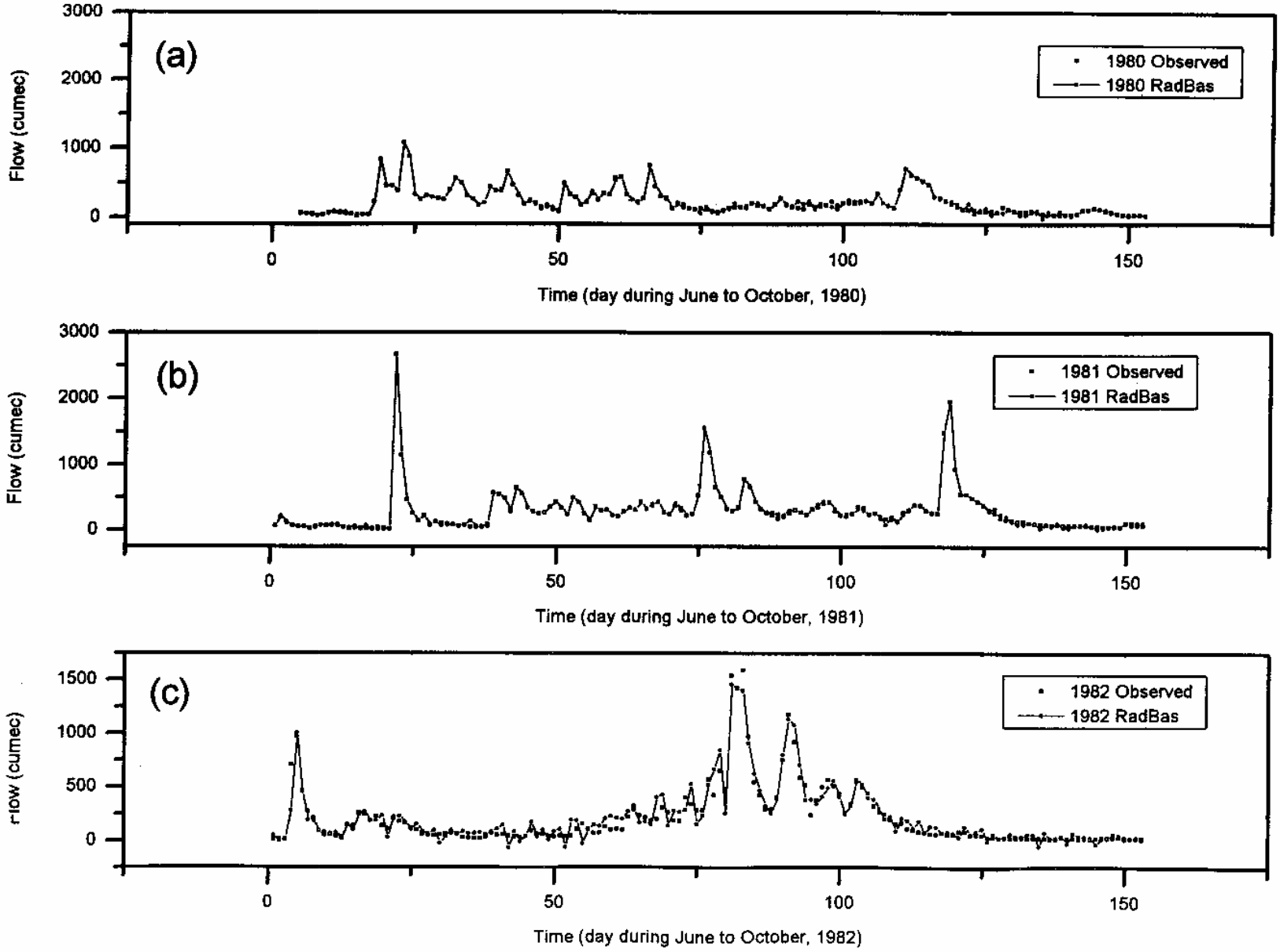


Fig 4.3 Observed and computed hydrographs using radial basis network: (a) & (b) Calibration (c) validation

deteriorating during validation, for all the models. However, it is worth noting that the errors are fairly small. For RBF network, the RMSE during validation is about 0.02 normalized units, which corresponds to 79 cumecs. The distribution of error over different patterns was also calculated, using the index N_{RMSE}. The values of N_{RMSE} indicated that, more than 80 percentage of the patterns was reproduced with error less than RMSE. The BPN 6N has worst RMSE during calibration as well as validation. On average, RBF model performed best as measured by this statistic.

Table 4.4 Goodness of fit statistics for the candidate models

	Correlation			RMSE		
	Training		Validation	Training		Validation
	1980	1981	1982	1980	1981	1982
BPN 6N	0.94	0.99	0.90	1.25E-02	1.21E-02	4.26E-01
RBF	0.99	0.99	0.96	5.76E-02	6.06E-03	2.20E-02
BPN 12N	0.96	0.99	0.72	9.59E-02	9.11E-03	4.86E-02

	% Error in Maximum flood			Efficiency		
	Training		Validation	Training		Validation
	1980	1981	1982	1980	1981	1982
BPN 6N	0.00	0.20	-61.05	88.04	97.11	56.64
RBF	0.00	0.00	5.14	98.17	99.57	92.15
BPN 12N	-0.03	0.36	1.74	92.72	98.46	41.71

The percentage error in maximum flow (%MF) measured the percent error in matching the maximum flow of the data record; 0.0 is the best, positive values indicates overestimation, and negative values indicate underestimation. During calibration, all the three models match the peak flow very well, but during validation the performance deteriorates in every case. The worst deterioration is for BPN 6N model (0.0 to –61.06 %), while RBF model slightly over estimated the peak flow during training (5%). However, the performance of the BPN 12N was the best among three during validation, and could be employed for flood estimation compared to other models.

The correlation statistic measured the linear correlation between the observed and simulated flows; the optimal value is 1.0. The correlation coefficient (CORR) is worse (smaller) during validation than during training for all the models, as was expected. The RBF model showed consistent correlation throughout the training and testing.

The efficiency of the model as defined by Nash-Sutcliffe criteria (equation 3.23) is a measure of the performance of the model in predicting the output values. A value of 90% and above indicates very satisfactory performance, a value in the range of 80-90% indicates fairly good

performance, and a value below 80% indicates an unsatisfactory fit (Shamseldin, 1997). According to this statistic, all the model predictions were extremely good during training. However, both the BPN models failed miserably during validation (nearly 50% efficiency). The RBF maintained the efficiency during validation too.

The results obtained from all the models during training and testing are plotted in a Fig. 4.4 in a dispersion diagram. The statistical index %MF, considered only the peak flow in the season, while the scatter diagram considered all the high flows and a better evaluation could be made based on this diagram. Reduced scatter confirms that good results have been obtained. However, the prediction of high flows (during training as well as testing) was not to the mark in all the models, though RBF was better among the three. The performance of the BPN 6N during the year 1980 was not satisfactory.

All the above analysis shows that a radial basis function network was able to model the rainfall-runoff process in the Baitarani river basin in a reasonably accurate manner. However, in forecasting of flood peaks BPN 12N is superior to RBF.

Comparison of best fit model with other models in use

From the reported analysis, the RBF model was selected as the best fit model among all the three candidate models considered, and its performance was evaluated with that of the other models in use in the basin.

In one of the pioneering studies in the basin, the Orissa Water Planning Organization (OWPO) of the Government of Orissa has conducted a regression analysis on 20 years of monthly rainfall-runoff data. The reported regression models were employed in the present study to compare the relative performance of RBF model and regression model. The OWPO has recently employed the "Sacramento rainfall-runoff model" in the Baitarani river basin for water resources planning (the details of these two studies can be obtained from OWPO, 1993).

In the present study, the results as reported by OWPO has been taken directly, without any cross checks for validity, so as to compare the relative performance of these models with RBF model developed in this study. The reported results from OWPO were in the form of monthly flows in mm. To make a direct comparison, the RBF simulated daily flows have been summed up over the months and transformed into similar units. The relative performance of these models is depicted in Fig 4.5. The figure shows values of monthly flows for the year 1980 to 1982. The figure clearly indicates the failure of Sacramento model (note that no analysis was carried out by the authors of this report to check the accuracy of the reported values).

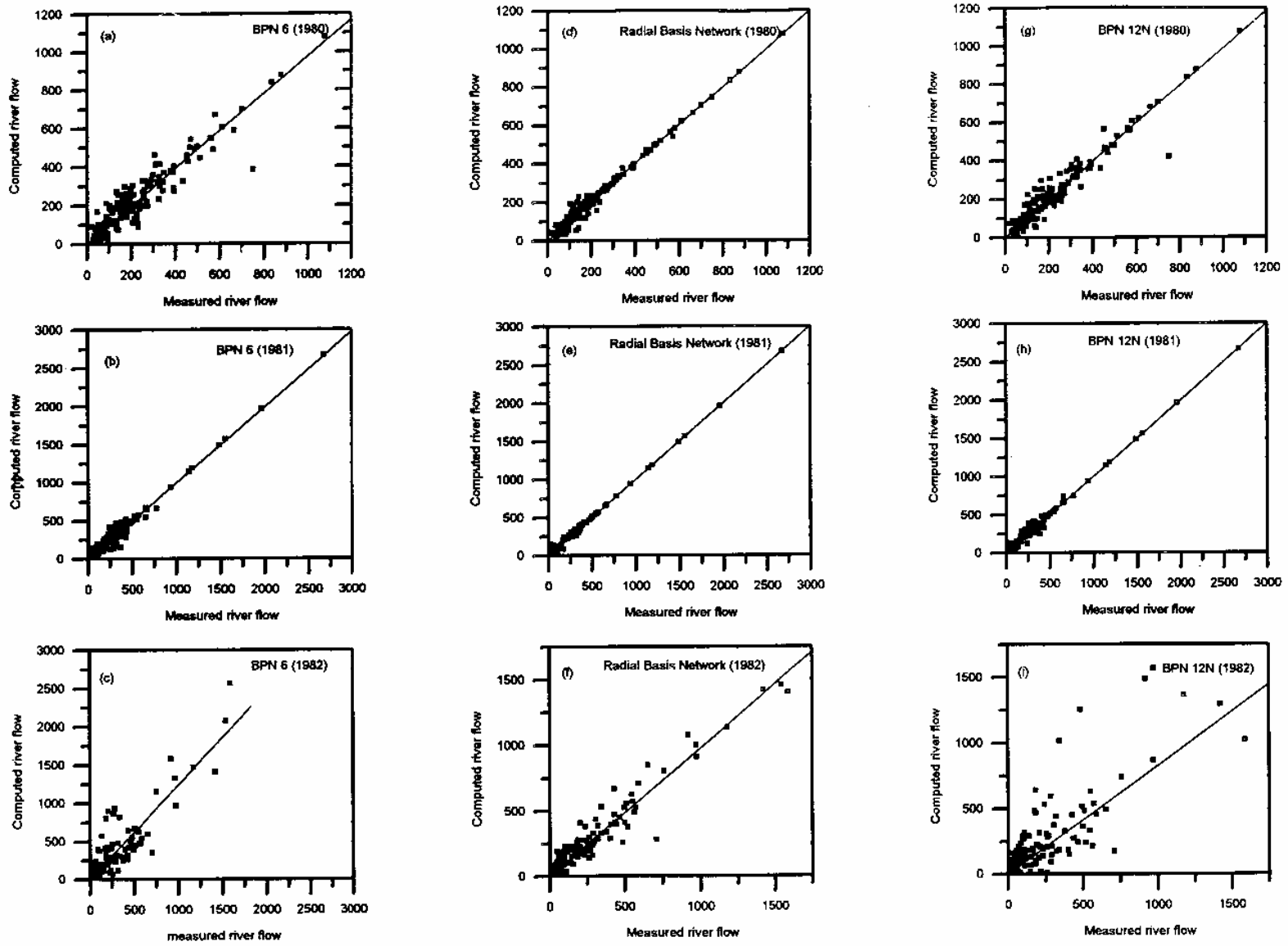
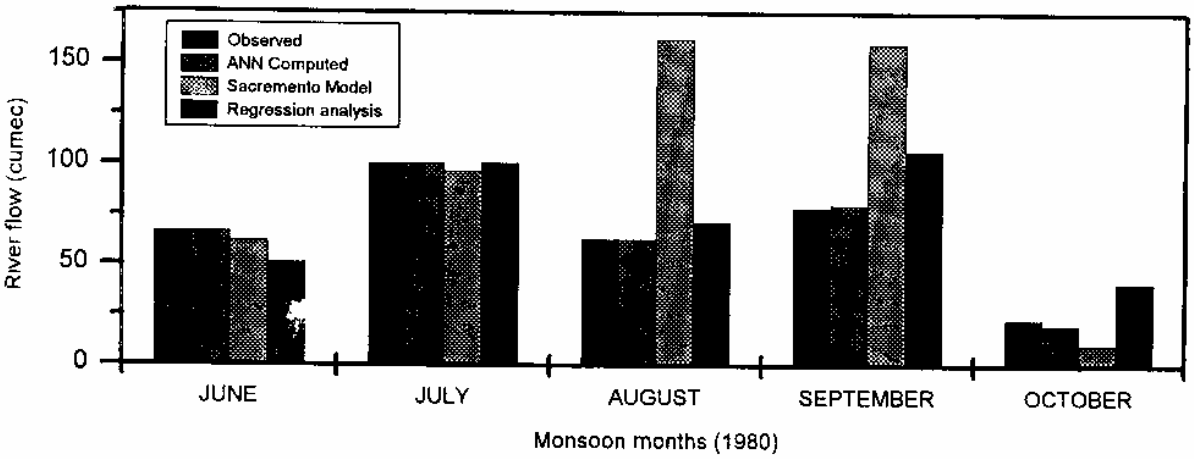
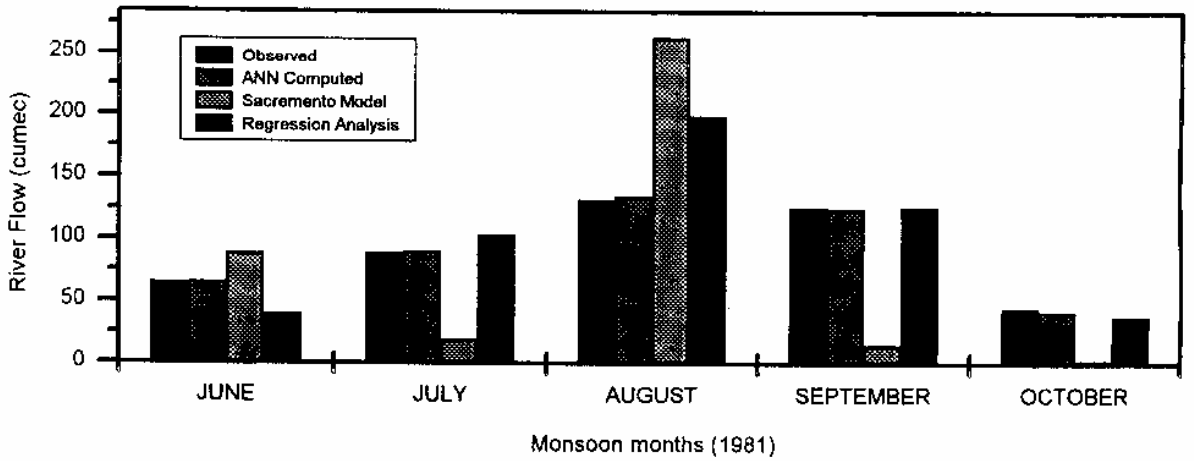


Fig 4.4 Scatter plot of the computed and observed flows: BPN 6N (a,b,c); Radial Basis Netwrok (d,e,f); BPN 12N (g,h,i)) for years 1980 & 1981 (training), 1982 (validation) respectively





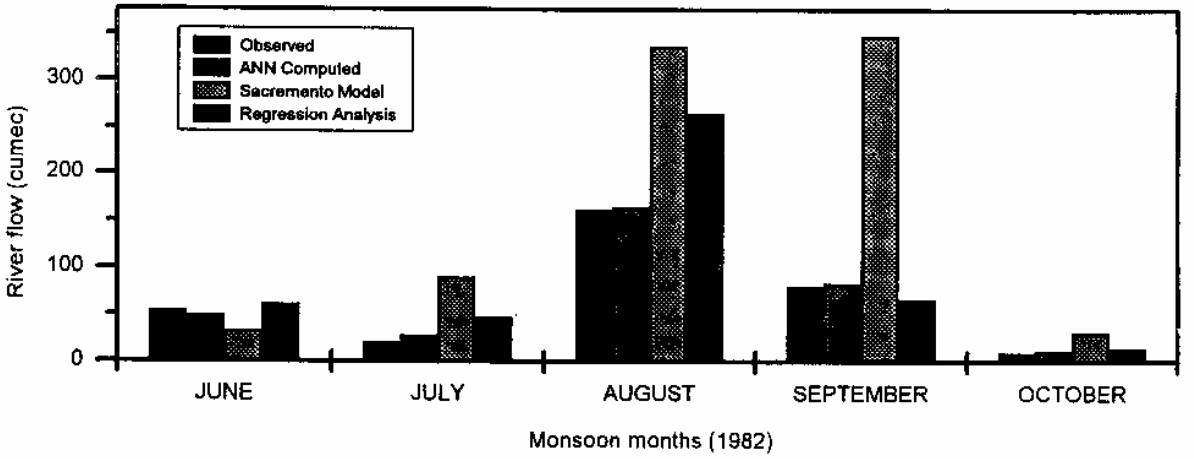


Fig 4.5 Comparison of ANN model performance with others

However, the fitted regression equations performed better than the Sacramento model. In all the years, the relative performance of the RBF model was superior to others. Hence it can be concluded that use of an RBF model for the water resources management in the basin may result in better utilization of the resources.

Chapter 5

Summary and Conclusions

A research study has been conducted to develop a rainfall-runoff model for the Baitarani river basin, Orissa. A detailed review of the research work in the area of interest revealed that the approach of neural computations was very effective in developing the required model, due to its various advantages. Accordingly, three candidate models based on ANN architecture were developed for the study area, to represent rainfall-runoff transformation.

All the three model's architecture was determined based on a trial and error procedure and examining various goodness of fit statistics. An auto correlation and partial auto correlation analysis of the standardised daily flow series suggested that the flow at time 't' was highly correlated to previous three days flows viz (Q_{t-1},Q_{t-2}, Q_{t-3}). These parameters were included in the input vector of the network, apart from 5 days rainfall series prior to the day, at which the flow was to be predicted. The number of rainfall patterns in the input vector was finalised by trial and error procedure.

Statistical analysis was done on the performance of each model in estimating the runoff. The study revealed that an ANN with radial basis function algorithm was able to model the R-R transformation more accurately than a back propagation network. However, for estimation of peak flows a BPN with 12 neurons in the hidden layer was found efficient. The results from the RBF model were compared with the results of existing models such as Sacramento model and regression equations developed, and the performance of the RBF model was found superior to others.

Hence it was concluded that the Radial Basis Function network model developed for rainfall-runoff process in the Baitarani river basin might be employed for water resources planning. While such a model is not intended as a substitute for a physically based model, it can provide a viable alternative when the hydrologic application requires that an accurate forecast of stream flow be provided using only the available input and output time series data, and with relatively little conceptual understanding of the hydrologic dynamics of the particular basin under investigation.

Aboitiz, M., J. W. Labadie, and D. F. Heerman, (1986). *Stochastic soil moisture estimation and forecasting for irrigation fields.* Water resources research, 22(2):180-190.

Akaike, H. (1974). *A new look at the statistical model identification.* IEEE transactions of automation and control, AC-19:716-723.

Bishop, C. M., (1994). *Neural networks and their applications.* Rev. Science Instruments., 65:1803-1832.

Box, G. E. P. and G. M. Jenkins, (1970). *Time series analysis, forecasting and control.*, San Fransisco, Holden day.

Chakraborty, K., K. Mehootra, C. K. Mohan, and S. Ranka. (1992). *Forecasting of the behavior of multivariate time series using neural networks*, Neural networks, 5:961-970.

Hartman, E. J., J. D. Keeler, and J. M. Kowalski (1990). *Layered neural networks with gaussian hidden units as universal approximations.* Neural computations, 2: 210-215.

Hecht-Neilson, R. (1990). *Neurocomputing.* Addison-Wesley Publishing Company, Reading, Mass.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the theory of neural computations.* Addison-Wesley publishing company, New York, pp: 246-250.

Minns, A. W., and M. J. Hall, (1996). *Artificial neural networks as rainfall runoff models*, Journal of hydrological sciences, 41:399-417.

Moody, J. and C. Darken (1989). *Fast learning in networks of locally tuned processing units.* Neural computation, 1: 281-294.

Orissa Water Planning Organization (1998). *Report on basin planning: Baitarani Basin (Second spiral study).* Govt. of Orissa, Bhubaneswar.

Rissanen, J. (1978). *Modeling by short data description*, Automation, 14:465-471.

Romesburg, H. C. (1984). *Cluster analysis for researchers.* Lifetime learning publications, Belmont., California.

Rumelhart, D. E., and McClelland, J., eds. (1986). *Parallel distributed prccessing,* Vol. 1, MIT Press, Cambridge.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, (1988). *Applied modeling of hydrologic time series*, Littleton, Colarado, Water Resources Publications.

Shemseldin, A. Y. (1997). *Application of a neural netwrok technique to rainfall runoff modeling,* Journal of Hydrology, 199:272-294.

Shumway, R. H. (1988). *Applies statistical time series modeling.* Eaglewood Cliffs, New Joursey, Prentice Hall.

Sridhar, S. P. (1996). *Synthesis of artificial neural networks using genetic algorithms.* M. Tech thesis submitted to Indian Institute of technology, New Delhi, India.

Woolshier, D. A. (1996). *Search for physically based rainfall runoff model – a hydrological El Dorado*?, Journal of hydrologic engineering, 122: 122-129.

STUDY GROUP

K. P. SUDHEER, Scientist 'B'
P. C. Nayak, Scientist 'B'
D. Mohan Rangan, Technician Gr. II