International Journal of Agriculture Innovations and Research
Volume 1, Issue 2, ISSN (Online) 2319-1473

Effect of Distribution of Dataset for Model Development using Neural Network Fitting with Emphasis on Yield of Cotton

Dr. T.M.V. Suryanarayana
Water Resources Engineering and Management Institute
Faculty of Technology & Engineering
The Maharaja Sayajirao University of Baroda,
Samiala -391410, Ta. & Dist.: Vadodara, Gujarat, India
tmvkiran@yhoo.com

Dr. Falguni Parekh
Water Resources Engineering and Management Institute
Faculty of Technology & Engineering
The Maharaja Sayajirao University of Baroda,
Samiala -391410, Ta. & Dist.: Vadodara, Gujarat, India
fpparekh@gmail.com

***Abstract* – The study was carried out to develop a weather based model on yield of cotton using Neural Network Fitting Tool for various sizes of data set. There is significance effect of various climatological data on yield of cotton. The climatological data for the hot weather season are collected for the period 1981- 2006 and correlated with yield of cotton in Vallabh Vidyanagar using neural network fitting. To form the dataset for model development and Evaluation, three alternatives are considered. In Alternative 1, 70% of data of Maximum and minimum temperatures, sunshine hours, relative humidity and wind velocity are correlated with yield data and 30% data are used for Evaluation of the model. In Alternative 2, 60% of above mentioned data are used to develop the model and 40% data are used to validate the model. In Alternative 3, 80% data are correlated with yield and 20% data are utilized for Evaluation of the model. Then, all the model are re-trained until the best coefficient of correlation is obtained and this corresponding model is considered as the best model and this is further used to validate the remaining dataset. This whole procedure is repeated for three different Alternatives. The results shows that the best model is obtained for alternative 1, in which R for training, testing and validation for model development are 0.92, 0.66 and 0.86 respectively and R for Evaluation is 0.72. The overall R value is 0.5 for this alternative. It can be concluded from the overall study that considering all the five climatological parameters, with use of 70% data for model generation and 30% data for Evaluation, the correlation is achieved as the best. The use of Neural Network fitting is quite helpful in studying the effect of distribution of dataset for model development.**

***Keywords* — Climatological Data, Neural Network Fitting, Coefficient of Correlation, Yield.**

I. INTRODUCTION

Climatic variability is the major factor influencing the agriculture productivity. Global climate change and its impacts on agriculture have becoming an important issue. Agriculture production is highly dependent on climate and it also adversely affected by increasing climatic variability. The aim is to develop the methodology for assessing this component of the total impact of climate variability on agricultural productivity. There is a need to quantify climatic variability and the amount of data to assess its effect on crop yield.

Agrometeorological models are defined as the product of two or more weather factors each representing functioning between yield and weather. These models do not require hypothesis of the plant and environment process. Thus the input requirement is less stringent but the output information is more dependent on the input data. Thus agrometeorological models are a practical tool for the analysis of crop response to weather and estimating the yield.

It always becomes important to decide how much proportion of data should be considered for model development and how much proportion for model evaluation. In this paper, an attempt had been made to check the efficacy of a model by changing the different proportions of data for model development and model evaluation.

Khashei-siuki et al. (2011) studied the ability of Artificial Neural Network (ANN) technology and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) for the prediction of dryland wheat (Triticum aestivum) yield, based on the available daily weather and yearly agricultural data.

Maqsood et al.(2004) presented the applicability of an ensemble of artificial neural networks (ANNs) and learning paradigms for weather forecasting in southern Saskatchewan, Canada.

Parekh and Suryanarayana (2012) carried out the study to determine the predominance of various climatological data on yield of wheat.

Schlenker and Roberts (2006) studied effect of climate Change on yield of crops and showed the importance of non-linear temperature effects on yield.

The present study was undertaken with a view to determine effect of distribution of dataset for model development using neural network fitting with emphasis on yield of cotton in Vallabh Vidhyanagar, Gujarat, India wherein the different weather parameters that are considered will be maximum & minimum temperatures, relative humidity, wind velocity and sunshine hours.

II. STUDY AREA

The entire Gujarat is divided into the various agro-climatic zones. Vallabh Vidyanagar is located in the Anand district and lies in middle Gujarat agro-climatic zone III of Gujarat state. Vallabh Vidyanagar is located at 22□32’ N latitude, 72□54’E longitude at an altitude of 34 m above mean sea level. It is bounded on the north by the Kheda district and south by the Gulf of Khambhat, on the west by Ahmedabad district and, on the east by Vadodara district.

The climate of Vallabh Vidyanagar is semi-arid with fairly dry and hot summer. Winter is fairly cold and sets in, in the month of November and continues till the middle of February. Summer is hot and dry which commences from mid of February and ends by the month of June. May is the hottest month with mean maximum temperature around 40.08°C. The average rainfall is 853 mm.

The soil of the region is popularly known as Goradu soil. It is alluvial in origin. The texture of the soil is sandy loam and black. The soil is deep enough to respond well to manuring and variety of crops of the tropical and sub-tropical regions. The soil is low in organic carbon and nitrogen, medium in available phosphorus and available sulphur. In this area paddy, tur, and ground nut, til are grown in kharif season. In rabi season wheat, gram, and jowar are grown. Especially, in summer season the bajara cotton and ground nut are grown. Tobacco is grown from August and harvested in March. In last few years, there is increase in amount of rainfall which facilitated in agriculture production and various irrigation scheme.

III. Data Collection

The data required for evaluation in this study are collected from India Meteorological Department, Pune and Krishi Bhavan, Gandhinagar.

Long term climatological daily data are collected from IMD (Indian Meteorological Department), Pune for Vallabh Vidyanagar, town of Anand district of Gujarat.

The basic weekly climatological data used comprises of Maximum and minimum temperature(˚C) , Relative humidity (%), Wind speed (Kmph) and Sunshine hours (hour). The yield data of various crops grown in Vallabh Vidyanagar are collected from the Krishi bhavan, Gandhinagar from year 1981-2006.

IV. Methodology

The weekly climatological data viz. temperature, relative humidity, sunshine hours, wind speed etc. are converted into average monthly data and seasonal data. Cotton is cultivated in Hot Weather season.

Some data gaps or missing values are indentified in data. The missing values were found out with the SPSS 11.5 software. The linear trend at a point method is used to find the missing values of the data. To study the impact of climatological data on yield of cotton, Neural fitting tool of Artificial Neural Network (ANN) of MATLAB is used.

ANN was first introduced as a mathematical aid and were inspired by the neural structure of the brain. An input layer, which is used to present data to the network. An output layer, which is used to produce an appropriate response to the given input; and one or more intermediate layers, which are used to act as a collection of feature detectors. The ability of a neural network to process information is obtained through a learning process, which is the adaptation of link weights so that the network can produce an approximate output. In general, the learning process of an ANN will reward a correct response of the system to an input by increasing the strength of the current matrix of nodal weights.

There are several features in ANN that distinguish it from the empirical models. First, neural networks have flexible nonlinear function mapping capability which can approximate any continuous measurable function with arbitrarily desired accuracy, whereas most of the commonly used empirical model, do not have this property. Second, being non-parametric and data-driven neural networks impose few prior assumptions on the underlying process from which data are generated. Because of these properties, neural networks are less susceptible to model misspecification than most parametric nonlinear methods.

An ANN can be defined as data processing system consisting large number of simple highly interconnected processing elements (PEs or artificial neurons) in architecture analogous to cerebral cortex of brain. An ANN consists of input, hidden and output layers and each layer includes an array of artificial neurons. A typical neural network is fully connected, which means that there is a connection between each of the neurons in any given layer with each of the neuron in next layer. An artificial neuron is a model whose components are analogous to the components of actual neuron in next layer. An artificial neuron is a model whose components are analogous to the components of actual neuron. The array of input parameters is stored in the input layer and each input variable is represented by a neuron. Each of these inputs is modified by a weight (sometimes called synaptic weight) whose function is analogous to that of the synaptic junction in a biological neuron. The neuron (processing element) consists of two parts. The first part simply aggregates the weighted inputs resulting in a quantity 1: the second part is essentially a nonlinear filter, usually cooed the transfer function or activation function. The activation function squashes or limits the values of the output of an artificial neuron to values between two asymptotes. The sigmoid function is the most commonly used activation function. It is a continuous function that varies gradually between two asymptotic values typically 0 and 1 or -1 and +1.

Neural Fitting Tool

In fitting problems, you want a neural network to map between a data set of numeric inputs and a set of numeric targets. Examples of this type of problem include estimating house prices from such input variables as tax rate, pupil/teacher ratio in local schools and crime rate (house_dataset); estimating engine emission levels based on measurements of fuel consum and speed (engine_dataset); or predicting a patient's bodyfat level based on body measurements (bodyfat_dataset).

The Neural Network Fitting Tool will help you select data, create and train a network, and evaluate its performance using mean square error and regression

analysis.

A two-layer feed-forward network with sigmoid hidden neurons and linear output neurons (newfit), can fit multi-dimensional mapping problems arbitrarily well, given consistent data and enough neurons in its hidden layer.

The network will be trained with Levenberg-Marquardt backpropagation algorithm. (trainlm), unless there is not enough memory, in which case scaled conjugate gradient backpropagation (trainscg) will be used. During past couple of years, the Levenberg-Marquardt (LM), a second order optimization technique is extensively employed in evapotranspiration modeling using neural networks.

In the present study, the climatological data for the Hot weather season are collected for the period 1981-2006 and correlated with yield of cotton. To form the dataset for model generation and Evaluation, three alternatives are considered. In Alternative 1, 70% of data of Maximum and minimum temperatures, sunshine hours, relative humidity and wind velocity are correlated with yield data and 30% data are used for Evaluation of the model. The first part, i.e. 70% of the Dataset, is further divided into 70% for Training, 15% for Validation and 15% for Testing. In Alternative 2, 60% of above mentioned data are used to develop the model and 40% data are used to validate the model. In Alternative 3, 80% data are correlated with yield and 20% data are utilized for Evaluation of the model. For these datasets, correlations of input and output are observed using neural fitting tool. Then, the model is re-trained until the best coefficient of correlation is obtained and this corresponding model is considered as the best model and this is further used to validate the remaining 30% of the Dataset.

This whole procedure is repeated for three different Alternatives.

Coefficient of Correlation, R

Measure of the "goodness of fit" is the coefficient of correlation, r. To explain the meaning of this measure, one has to define the standard deviation, which quantifies the spread of the data around the mean:

$$s_t = \sum_{i=1}^{n} \left(\bar{o} - o_i \right)^2 \quad (1)$$

Where St is the standard deviation, oi is the observed data points

$$\bar{o} = \frac{1}{n} \sum_{i-1}^{n} o_i \quad (2)$$

The quantity St considers the spread around a constant line (the mean) as opposed to the spread around the regression model. This is the uncertainty of the dependent variable prior to regression. One also defines the deviation from the fitting curve as

$$s_r = \sum_{i=1}^{n} \left(o_i - p_i \right)^2 \quad (3)$$

Where Sr is the deviation from the fitting curve, pi is the predicted data points.

$$R = \sqrt{\frac{s_t - s_r}{s_t}} \quad (4)$$

Note the similarity of this expression to the standard error of the estimate; this quantity likewise measures the spread of the points around the fitting function.

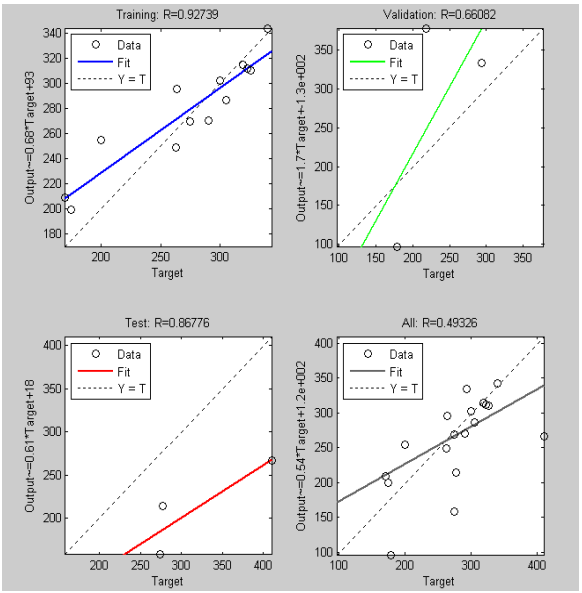
Where, R is defined as the coefficient of correlation.As the regression model starts improving describing the data, the correlation coefficient approaches unity. For a perfect fit, the standard error of the estimate will approach Sr = 0 and the correlation coefficient will approach R = 1.

V. Results And Analysis

The Coefficient of correlation, R for Alternatives 1, 2 and 3 are given in Table 1.

Table 1 Coefficient of correlation, R for Alternatives 1, 2 and 3

Alternative	Model				
	Development				Evaluation
	Training	Validation	Testing	All	
1	0.92	0.66	0.86	0.50	0.72
2	0.99	1.00	1.00	0.89	0.30
3	0.86	0.92	0.98	0.54	0.12



International Journal of Agriculture Innovations and Research
Volume 1, Issue 2, ISSN (Online) 2319-1473

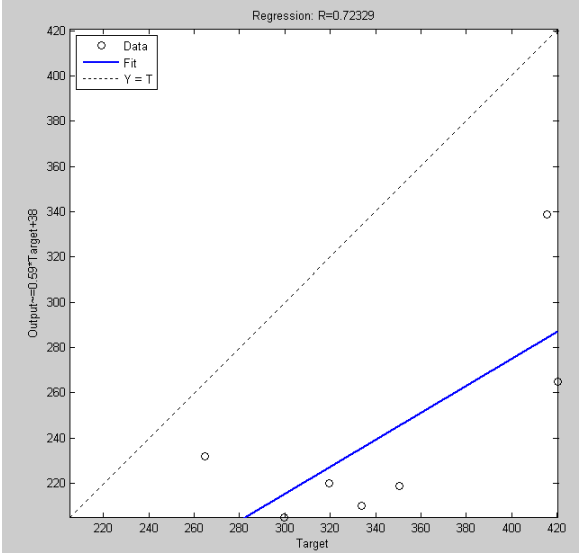
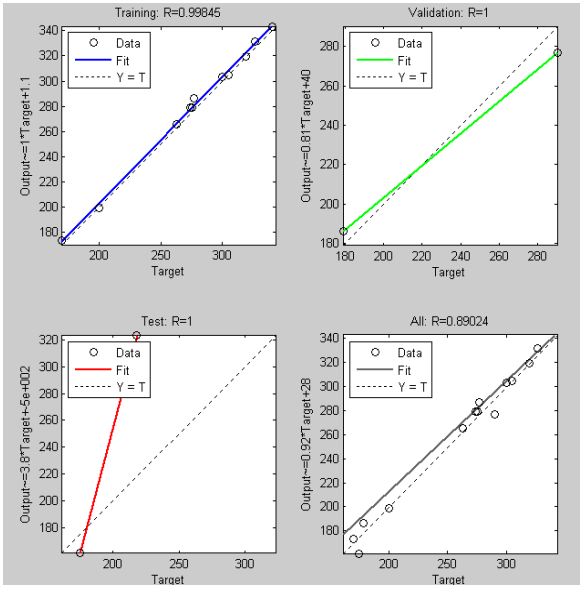


Fig.1. Correlation coefficient, R for Alternative 1

The predefined correlation coefficient, R is found for each stage (training, validation, testing) for each model and these values also were found for over all data in addition to an additional results for a randomly selected data for additional testing. Values of Correlation coefficient, R are plotted for (training, validation, testing) for each model and are given in Fig. 1, 2 and 3 for Alternatives 1, 2 and 3 respectively.



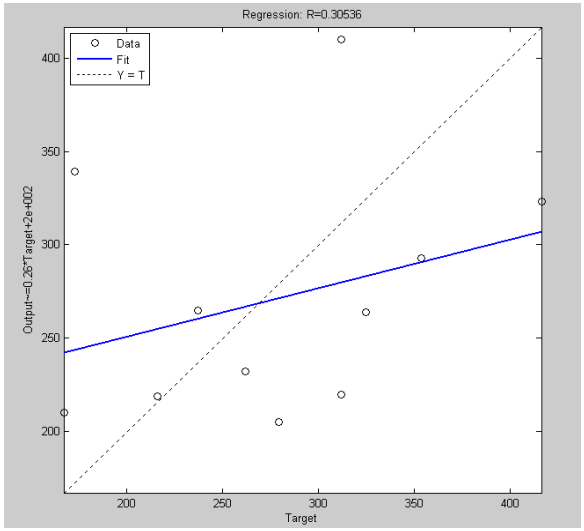
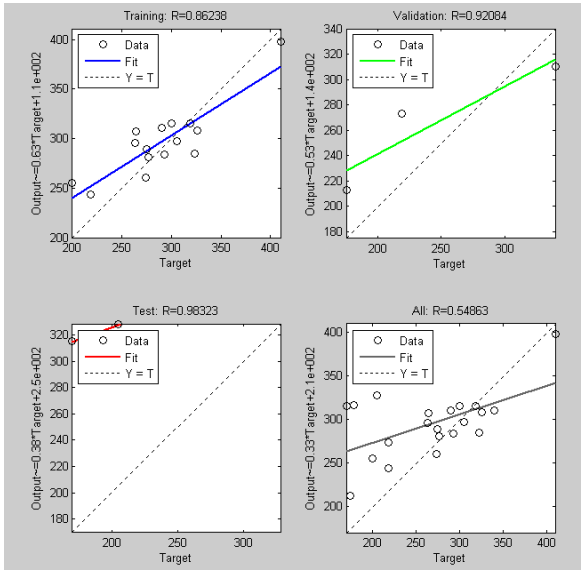


Fig.2. Correlation coefficient, R for Alternative 2



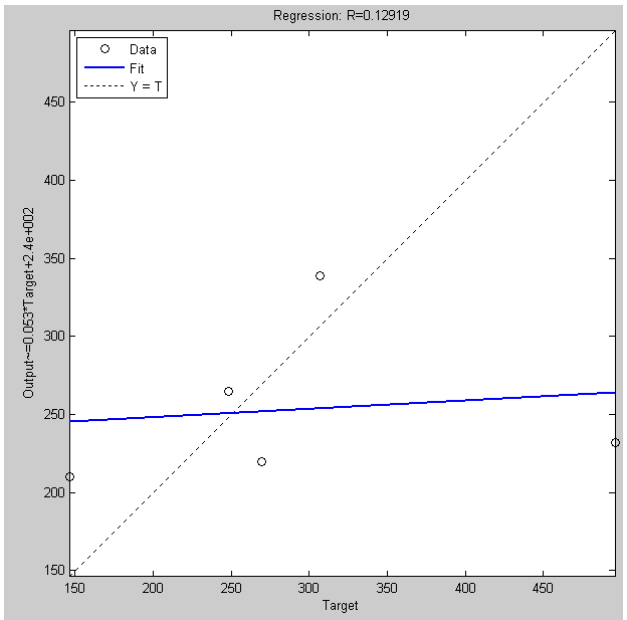


Fig.3. Correlation coefficient, R for Alternative 3

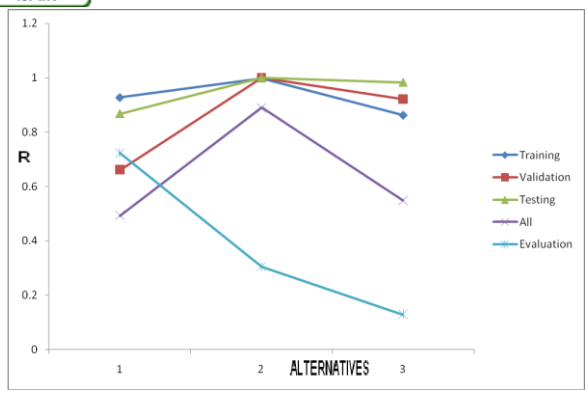


Fig.4. Correlation coefficient, R for Alternatives 1, 2, 3

Looking to the Table 1 and Figs. 1, 2, 3 and 4, as per Alternative 1, i.e. considering correlation between maximum and minimum temperatures & Sunshine Hours & relative Humidity & Wind velocity and Yield, the co-efficient of correlation for training and testing are better compared to validation for 70%, whereas in 30% dataset, R comes out to be 0.72, which is very good. The overall R is also 0.5.

As per Alternative 2, the co-efficient of correlation for training, testing and validation are 0.99, 1.0 and 1.0 respectively which is best for 60%, whereas in 40% dataset, R comes out to be 0.30, which is comparatively very low. Here, one can observe that 10% variation in data set affects the R value significantly, which is about 40% lesser.

As per Alternative 3, the co- efficient of correlation for training, testing and validation are 0.86, 0.92 and 0.98, for 80%, whereas in 20% dataset, R comes out to be 0.12, which is very low. Here, one can observe that 10% variation in data set affects the R value significantly, which is about 60% lesser.

VI. Conclusions

R varies from 0.5 to 0.92 for alternative 1, whereas for alternative 2, is 0.3 to 1.0 and for alternative 3, it ranges from 0.12 to 0.98. In general, a better model is always judged to have a minimum R of 0.50, which reveals that model developed for Alternative 1 is better than other two models. It can be concluded from the overall study that considering all the five climatological parameters, with use of 70% data for model generation and 30% data for Evaluation, the correlation is achieved as the best. The use of Neural Network fitting is quite helpful in studying the effect of distribution of dataset for model development.

References

[1] Imran Maqsood, Muhammad Riaz Khan and Ajith Abraham (2004). ‘‘An ensemble of neural networks for weather forecasting’’, Neural Comput & Applic (2004) 13: 112-122.

[2] Khashei-Siuki,A., Kouchakzadeh,M. and Ghahraman,B. (2011). ‘‘Predicting Dryland Wheat Yield from Meteorological Data Using Expert System, Khorasan Province, Iran’’, J. Agr. Sci. Tech. Vol. 13: 627-640.

[3] Parekh, F.P. and Suryanarayana, T.M.V. (2012). “Impact of climatological Parameters on yield of wheat using neural network fitting.”, International Journal of Modern Engineering Research, Vol.2,Issue.5,Sep-Oct., pp-3534-3537.

[4] Schlenker, Wolfram and Michael, (2006). “Estimating the impact of climate change on crop yields:
The importance of non-linear temperature effects” Paper No.: 0607-01 Department of Economics, Columbia University, New York.

Author’s Profile

Dr. T.M.V. Suryanarayana

is born in Visakhapatnam on 11th February, 1979 and completed B.E.(Civil-IWM) in May 2001, M.E.(Civil) in Water Resources Engineering in November 2002 and Ph.D. in Civil Engineering in May 2007 from The M.S. University of Baroda, Vadodara, Gujarat, India.

He is serving as Assistant Professor and recognized Ph.D. Guide in Water Resources Engineering and Management Institute, Faculty of Technology and Engineering, The M. S. University of Baroda. He has 9 years of teaching and Research Experience. His areas of research include Operations Research, Hydrologic Modeling, Conjunctive Use, Hydraulics of Sediment Transport, Soil and Water Conservation, Reservoir Operation, Soft Computing Techniques, Climate Change.

Dr. Suryanarayana is Secretary and Treasurer, Gujarat Chapter of Association of Hydrologists of India and Joint Secretary, Indian Society of Geomatics_Vadodara Chapter. He is Collaborator of Columbia Water Centre, New Delhi Division. He is an Editorial Board Member in Seven International and National Journals, some of them include International Journal of Water Resources and Environmental Engineering, International Journal of Civil Engineering, A Water and Enviro-Modelling Group Journals, Tanstellar Journals and CSC Journals and also Member of International Association of Hydrological Sciences (IAHS, France), International Association of Engineering and Management Education. He is Life Member of various Professional Bodies like Association of Agrometeorologists, Association of Hydrologists of India, Indian Society of Hydraulics, Indian Water Resources Society, Indian Society for Geomatics. He is Nodal Principal Investigator for an AICTE Sponsored NCP-Research Project worth Rs. 40.00 Lakhs. He is Collaborator of Columbia Water Centre, New Delhi Division. He has been Invited as speaker on various programmes related to Awareness on Water Management and one of the consultants in carrying out various Technical Projects assigned by Government of Gujarat/Private firms to WREMI. He has obtained Best Paper Award at National conference on emerging vistas of technology in 21st Century, Organized by Gujarat Technological University, Ahmedabad in Collaboration with The Indian Journal of Technical Education & Indian Society for Technical Education Supported By Parul lnstitute of Engineering & Technology and Ahmedabad Management Association, Ahmedabad, Gujarat. He has under his credit 45 Research Papers published in various International / National Journals / Seminars / Conferences / Symposiums.

Dr. F. P. Parekh

born in Vadodara on 24th May 1970 and has completed B.E. (Civil-IWM) in August 1991, M.E. (Civil) in Irrigation Water Management in July 1998 and Ph.D. in Civil Engineering from The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India.

She is serving as Associate Professor in Water Resources Engineering and Management Institute, Faculty of Technology and Engineering, The M. S. University of Baroda. She has 16 years of research experience and 14 years of teaching experience. Her areas of research include Reservoir Operation, Soft computing techniques, Micro Irrigation , Benchmarking of Irrigation Projects, Climate Change and its Impact on Water Resources, Rain Water Harvesting, and Low cost Micro Irrigation Systems.

Dr. Parekh is Life Member of various professional bodies like Indian Society of Hydraulics, Indian Water Resources Society, Association of Hydrologists of India and Association of Agrometeorologists. She is chairman of Board of studies of Water Resources Engineering and Management Institute and member of Faculty Board of Technology &

Engineering. She is also joint secretary of Gujarat Chapter of Association of Hydrologists of India. She Worked as Principal Investigator for Research Project on Evolution of design criteria for Low cost micro irrigation system and its response on yield of crop funded by Gujarat Council on Science and Technology. She has been invited as speaker by various organizations and one of the consultants in carrying out various Technical Projects assigned by Government of Gujarat/Private firms to WREMI and completed five institutional consultancy projects. She is recipient of "Prof. S.C. Puranik Young Scientist Award" for the award winning paper in 2004 by Association of Hydrologist of India. She has published 29 Research Papers in various International/ National Journals/ Seminars / Conferences.