Forecasting Summer Rainfall Over India Using Genetic Algorithm

C.M.Kishtawal, Sujit Basu, Falguni Patadia and P.K. Thapliyal

Meteorology and Oceanography Group

Space Applications Centre

Ahmedabad-380015, INDIA

Abstract. In this study we have assessed the feasibility of a nonlinear technique based on genetic algorithm for the prediction of summer rainfall over India. The genetic algorithm finds the equations that best describe the temporal variations of the seasonal rainfall over India, and therefore enables the forecasting of the future rainfall. The forecast equation developed in the present study uses the monthly mean rainfall during June, July, and August for past years over five rainfall homogeneous zones of India to predict the seasonal rainfall (JJA combined) over the Indian landmass.

Introduction

Modeling natural phenomena has been a standard practice in atmospheric sciences. Traditionally, modeling a dynamical system requires one to derive the equations of motion from first principles, to measure initial conditions and, finally, to integrate the equations of motion forward in time. Alternatively, when such an approach is not feasible due to some reasons, e.g., the model may be far from perfect, and initial conditions may

be erroneous or even the required computing resources are not available, empirical laws governing the physical processes can be obtained by model-fitting approaches based on the observed variability of the system evolution. Nowadays, it is known that not all random-looking behavior is the product of complicated physics but it may result from the chaotic nature of a nonlinear and deterministic dynamics involving few degrees of freedom. In such cases, it is possible to exploit this determinism to make short-term forecasts that are more accurate than those obtained employing a linear stochastic model. These forecasts are carried out by deterministic models directly built from observations of the system evolution. A genetic algorithm is programmed to approximate the equation, in symbolic form, that describes the time series [*Szpiro*, 1997, *Alvarez et al.,* 2001]. The genetic algorithm considers an initial population of potential solutions, which is subjected to an evolutionary process, by selecting those equations (individuals) that best fit the data. The strongest strings choose a mate for reproduction whereas the weaker strings become extinct. The newly generated population is subjected to mutations that change fractions of information. The evolutionary steps are repeated with the new generation. The process ends after a number of generations *a priori* determined by the user.

The works of [*Takens* ,1981], [*Casdagli* ,1989], and many others have established the methodology for nonlinear modeling of chaotic time series. Explicitly, Takens' theorem [*Takens*, 1981] establishes that given a deterministic time series $\{x(t_k)\}$, $t_k=k\Delta t$, $k=1\ldots N$, there exists a smooth map P satisfying:

$$x(t) = P[x(t-\Delta t), x(t-2\Delta t), \ldots\ldots x(t-m\Delta t)] \qquad (1)$$

where *m* is called the embedding dimension obtained from a state-space reconstruction of the time series [*Abarbanel et al.*, 1993] and Δt is the sampling time interval, 1 year in our case. A genetic algorithm basically tries to obtain the function P(.) in Eq. (1) that best represents the amplitude function of a chaotic time series, which can then be used to predict the future state of the system. Generally the evolution of a natural dynamical system is not restricted to a single variable, and a nonlinear interaction among several variables is quite common. Such situation demands the use of multivariate or vectorial time series to obtain the fittest model that can explain a process. The model of connection between different variables, e.g. x, y, and z can be written as:

$$x(t)=\acute{P}[y(t-\Delta t),y(t-2\Delta t),\ldots y(t-m\Delta t)\ldots..z(t-\Delta t),z(t-2\Delta t)\ldots z(t-m\Delta t).] \quad (2)$$

where $m+1 \leq t \leq T$, T being the length of vector time series. The procedural details of genetic algorithm can be found in *Szpiro* [1997], and *Alvarez et al.*, [2001]. A brief description of genetic algorithm is as follows. First, for an amplitude function x(t), a set of candidate equations for $\acute{P}$(.) is randomly generated. An equation is stored in the computer as a set of characters that define the independent variables, y(t-Δt), y(t-2Δt)…z(t-Δt), z(t-2Δt), etc. in Eq. (2), and four elementary arithmetic operators (+,-,x, and /). A criterion that measures how well the equation strings perform on a training set of the data is its fitness to the data, defined as sum of the squared differences between data and forecast from the equation string. The strongest individuals (equations with best fits) are then selected to exchange parts of the character strings between them (reproduction and crossover) while individuals less fitted to the data are discarded. Finally, a small percentage of the equation strings' most basic elements, single operators and variables, are mutated at random. The process is repeated a large number of times to

improve the fitness of the evolving population of equations. The fitness strength of the best scoring equation is defined as:

$$R^2 = 1 - [\Delta^2 / \sum (x_o - \langle x_o \rangle)^2] \qquad (3)$$

where $\Delta^2 = \sum (x_c - x_o)^2$, x_c is parameter value estimated by the best scoring equation, x_o is the corresponding "true" value, $\langle x_o \rangle$ is the mean of the "true" values of x.

Szpiro [1997] has shown the robustness of genetic algorithm to forecast the behavior of one-dimensional chaotic dynamical system. Later, *Alvarez et al.* [2000] applied the genetic algorithm to real physical systems and used this algorithm for the prediction of space-time variability of the sea surface temperature (SST) in the Alboran Sea.

In the present study we have attempted to predict summer monsoon (June to August) rainfall over the Indian landmass. The amount and the variability of monsoon rainfall have profound economic and social impacts for India and other countries of East Asia. This fact has led the researchers to develop means to predict the monsoon rainfall with sufficient lead-time.

Krishnakumar et al. [1995] developed a multiple regression equation for predicting all India summer monsoon rainfall (AISMR) from a comprehensive set of 19 predictors for the data period 1951-80. The multiple correlation coefficient was 0.893. They tested the performance of regression model during the training (1951-80) and verification (1981-94) periods. In the independent verification, eight years (1981-82,1984,1987-90,1992) showed good agreement (within ± 5%) between the observed and estimated values. Six years (1983,1985-86,1991,1993-94) showed large differences. Incidentally, a majority of regression models also failed to predict the AISMR accurately in these six years. Auto-regressive integrated moving average (ARIMA) models were also used to

predict the AISMR as well as the monsoon rainfall over north-west India and peninsular India. These were reported to have shown marginally better-forecast skills over the multiple regression model [*Thapliyal*, 1990]. However, the auto-correlations in AISMR during the period 1871-1900 are statistically insignificant. In view of this, the applicability of ARIMA models for monsoon rainfall forecasting is doubtful.

Basu and Andharia [1992] analyzed the AISMR time series during 1871-1989 using the method of deterministic chaos and found a strange attractor implying the existence of a prediction function. The reported correlation between predicted and observed rainfall was 0.69, however the RMS error was a bit large (5.84 cm).

Goswami and Srividya [1996] studied various neural network configurations for prediction of AISMR and found that a composite network (CN) produced the best result. They trained various networks on a 51-year dataset (rainfall of 1870 – 1920) and validated their performances on a 16-year (1926 – 1941) data set. The average error corresponding to each configuration varied from 4.6% to 7.3% and the maximum error varied from 13.6% to 16.7% (the least error is for the CN). However, a look at the Fig. 3 of their paper will convince the reader that the prediction for most of the individual years is far from satisfactory.

In the present work, we have tried to formulate the time evolution of summer rainfall over India using the genetic algorithm. The analytical expression thus obtained is used for the prediction.

Data

For the present study we have used the monthly mean rainfall data set over India prepared by Indian Institute of Tropical Meteorology (IITM) [*Parthasarathy et al*., 1995].

This data set contains monthly rainfall observations at 29 subdivisions of India for the years 1871-2002. Based upon the intensity and variability of rainfall, the geographical region of India has been divided into 6 homogeneous zones (Fig. 1), each containing 3 to 10 subdivisions. An area weighted averaging of sub divisional rainfalls was carried out to obtain the monthly rainfall over five homogeneous zones (excluding North-West Himalayan zone). The hilly regions consisting of four meteorological subdivisions of India, which are parallel to Himalayan mountain range, (shown as unfilled region in Fig. 1) have not been considered in view of the meager rain-gauge network and low areal representation of a rain-gauge in a hilly area. Thus, the contiguous area having network of 306 stations over 29 meteorological subdivisions measures about 2,880,000 sq.km, which is about 90 percent of the total area of India. In the final data set the predictor set consists of time series of June, July, and August rainfall over 5 homogeneous zones (total 15 independent time series, each containing 132 points), while the predicted variable is the all-India rainfall during summer season (cumulative rainfall for June-to-August (JJA hereafter)) over the combined area of 5 homogeneous zones). It was expected that the use of multivariate data in the predictor set would allow the genetic algorithm to capture the complex behavior of space-time variability of rainfall over the region of study. Fig. 1 shows the region for which the prediction of JJA rainfall is attempted. The mean JJA rainfall from 132 years over the region of study is 67.77 cm, while its standard deviation is 5.84 cm (8.61 %). Maximum rainfall occurred in 1956 when JJA rainfall was 16.75% above mean value, while the minimum rainfall took place in 1877 when it was 33.59% below the mean value (Fig. 2).

Results

Genetic algorithm (GA) was applied to find the equation that best fits the rainfall data in one part of the dataset, the training set, ranging from 1871 to 1992 (122 years). The predictability skill of the solution equation was then validated with data ranging from 1993 to 2003. In order to understand the dependence of performance of the employed algorithm on lag (m), we generated several prediction equations for different values of m using two different approaches, viz., univariate (forecasting all-India summer (JJA) rainfall in terms of its past values) as well as multivariate (forecasting all-India summer (JJA) rainfall in terms of past values of rainfalls over 5 homogeneous zones during the months of June, July, and August). The evolution process that consisted of 1000 generations for each case was initiated with identical set of initial random population of 800 equations, and the final equation was allowed to contain sufficiently large number of terms. The performance of the algorithm for each case was evaluated using the statistical criteria of standard error (SE $=[\sum(R_{fitted}-R_{actual})^2/N]^{1/2}$, where R_{fitted} and R_{actual} are the predicted and observed rainfall and N is length of the training data set) and fitness strength defined by Eq. (3). Fig. 3 shows the results of above analysis. For smaller values of m the resulting equations have relatively large standard error and small fitness strength. The performance of the algorithm improves with increasing m but after m=52 the performance starts declining. Standard errors of univariate and multivariate approaches are comparable but in terms of fitness strength the multivariate approach outperforms the univariate approach (Fig. 3-b). For these reasons we selected the multivariate approach with m = 51. Consequently, the prediction equation contains rainfall data at very long time lags (~ 50 years) for univariate as well as for multivariate

simulations. Meteorological systems including Indian monsoon cannot be expected to have the memory of such a large period, and hence the source of such lags appears to be some external forcing. *Mehta and Lau* [1997] analyzed decadal and multi-decadal variability in Indian monsoon rainfall and linked it to solar cycles. *Agnihotri and Dutta* [2003] concluded that the solar activity cycle of 60±10 years period have a significant and coherent influence on the monsoon rainfall variability of ~ 55 years period.

As mentioned in section-1, GA begins the evolution process with a set of randomly selected candidate equations, and the more and more fit candidates emerge in successive generations through crossover and mutation. Unless there exists an equation that perfectly fits the time series under consideration, the final outcome of GA process may depend upon the choice of initial set of candidate equations. In other words, it is possible that there may be more than one equation that may fit the time series with identical strength index. In such situation, one has to adopt criteria that can help in selection of the best equation. In the present study, we tried to pick an equation in such a way that the difference between the fitted and the actual values of the time series does not exceed a threshold at any instance, in addition to the fact that such equation should achieve the maximum possible fitness strength.

It is to be noted that due to the limitation on the population of candidate equations, or the number of operators allowed in an equation, it is difficult to obtain the fittest expression at the end of the initial evolution process. In our case also, the fitness value (R^2) was just 0.56 for training data set and 0.53 for the validation data set after the first run of the genetic algorithm consisting of 10,000 iterations. Following *Szpiro* [1997], genetic algorithm was once again applied to the series of residuals. The latter is defined

as the series of values ε_t that remain after the results of fittest equation have been subtracted from the original series:

$$\varepsilon_t = x_t - F^* (y(t-\Delta t),(y-2\Delta t).....(y-m\Delta t), z(t-\Delta t),z(t-2\Delta t)....z(t-m\Delta t)), m+1 \leq t \leq T \quad (4)$$

Here T is the total length of the data set, and $F^*(.)$ is the top scoring equation at the end of the first evolution process. We call this the second evolution process. In the first evolution process, the maximum value of the lag (m) was 51 years. Thus only 71 years of training data (1922 – 1992) was available for the residual time series, see Eq. (4). The value of m in top ranking equation that best fitted the residual time series was 20 years. The final equation that combines the top ranking equations from both the above evolution processes thus holds to the rainfall data from 1943-1992 (50 years), the part of the data used during the training, and a validation part from 1993-2003 that was not used during any of the evolution processes. The fitness of the final equation (R^2) was 0.705 for the training data set and was 0.644 for the validation data set. For the period 1943-2003, the natural variability or the standard deviation of the all India summer rainfall was 6.16 cm (~ 9% of the mean), while the root mean square difference between the fitted and actual rainfall was just 2.86 cm, which is less than half the natural rainfall variability in this period. Fig. 4 shows a scatter diagram of the actual and predicted rainfalls for the period 1943-2003. We have marked the data for the training period with one symbol and that for the validation period with another symbol. GA prediction for past two years is worth mentioning. The year 2002 was an abnormally poor monsoon year when the JJA rainfall over the region of study was 16.53% below normal. The prediction for 2002 by GA model was 11.03% below normal. GA model shows a general tendency to slightly overestimate the rainfall in drought years. The GA model prediction for 2003 was

10.96% above normal (75.1 cm), while the actual rainfall was 9.89% above normal (74.4 cm) as per the observations by India Meteorological Department [IMD, 2003]. The form of the final GA model is presented in Appendix-A.

Conclusion

An empirical technique has been developed using genetic algorithm that allows the prediction of summer monsoon (June to August) rainfall over a large part of India. The present empirical model uses the monthly mean rainfall during June, July, and August for past years over five rainfall homogeneous zones of India to predict the seasonal rainfall (JJA combined) over the Indian landmass. The major advantage of using genetic algorithm versus other nonlinear forecasting techniques such as neural networks is that an explicit analytical expression for the dynamic evolution of the rainfall time series is obtained. At present we have used only the monsoon rainfall during past years at selected homogeneous zones as the predictor set. Our future plan is to use more predictors; particularly the boundary variables such as SST, snow cover and soil moisture during present and past years to develop an equation with better fitness and lesser variability.

Acknowledgements: We pay our sincere thanks to Dr. Pranav S. Desai, Dr. M.S.Narayanan and Dr M Rajeevan for their valuable suggestions regarding this study. We are also indebted to Dr. Alvarez who generously provided us the computer code of the genetic algorithm used by us.

Appendix-A: Analytical expression of rainfall amplitude function.

The final analytical expression obtained after two processes of evolution is

EVOL1=71.87+((x05(t-28)/(x02(t-37)-43.95))/((9.4+0.48*x13(t-46)*x14(t-30)-2.1*(x13(t-46)+x14(t-30)))/(x03(t-39)*(x06(t-51)-x10(t-23)))))

RESID=(9.4+0.48*x14(t-1)*x14(t-1)-4.2*x14(t-1))/(x10(t-12)*(57.29-x01(t-1)-x12(t-19)+0.69*x14(t-2)-x02(t-14)))-x06(t-12)/x02(t-11)

Rainfall (t) = EVOL1 + RESID

t = Year for which the prediction of JJA rainfall (unit = centimeter) is made.

x01(t-1) = June rainfall at Zone-1 for year (t-1)

x02(t-1) = July rainfall at Zone-1 for year (t-1)

x03(t-1) = August rainfall at Zone-1 for year (t-1)

x04(t-1) = June rainfall at Zone-2 for year (t-1)

x05(t-1) = July rainfall at Zone-2 for year (t-1)

x06(t-1) = August rainfall at Zone-2 for year (t-1)

x13(t-1) = June rainfall at Zone-5 for year (t-1)

x14(t-1) = July rainfall at Zone-5 for year (t-1)

x15(t-1) = August rainfall at Zone-5 for year (t-1)

Similar notations hold good for other time steps.

References

Abarbanel, H.D.I., R. Brown, J. Sidorowich, and L.S.Tsimring, *Rev. Mod. Phys*., *65*, 1331-1392, 1993.

Agnihotri, R., and K. Dutta, Centennial scale variations in monsoon rainfall (Indian, east equatorial and Chinese monsoons): Manifestations of solar variability, *Current Science, 85*, 459-463, 2003.

Alvarez A., C. Lopez, M. Riera, E. Hernandez-Garcia, and J. Tintore, Forecasting the SST space-time variability of the Alboran Sea with genetic algorithms, *Geophys. Res. Lett. 27*, 2709–2712, 2000.

Alvarez A., A.Orfila, and J. Tintore , DARWIN: An evolutionary program for nonlinear modeling of chaotic time series, *Comp. Phys. Comm*., *136, 334-349*, 2001.

Basu S., and H.I.Andharia, The chaotic time series of Indian monsoon rainfall and its prediction, *Proc. Indian Acad. Sci. (Earth Planet. Sci.)*, *101*, 27-34, 1992

Casdagli M., Nonlinear prediction of chaotic time series, *Physica.D.35*, 335–356, 1989.

Goswami P., and Srividya, A novel neural network design for long range prediction of rainfall pattern. *Current Science, 70*, 6, 447-457, 1996

IMD, All India Weekly Weather Report-2003.

Krishnakumar K., M.K.Soman, and K. Rupakumar, *Weather, 50*, 12, 449-467, 1995.

Mehta V.M., and K. -M. Lau, Influence of solar irradiance on the Indian monsoon-ENSO relationship at decadal-multidecadal time scales. *Geophys. Res. Lett*. , 24, 159-162, 1997

Parthasarathy B., A.A. Munot, and D.R. Kothawale, *Research Report* No. RR-065, *IITM*, Pune, 1995

Szpiro G.G., Forecasting chaotic time series with genetic algorithms, *Phys. Rev. E 55* 2557–2568, 1997.

Takens, F., in *Dynamical systems and turbulence*, edited by D. Rand and L.S. Young, Springer-Verlag, Berlin, 1981.

Thapliyal V., *Mausam, 41*, 339-346, 1990

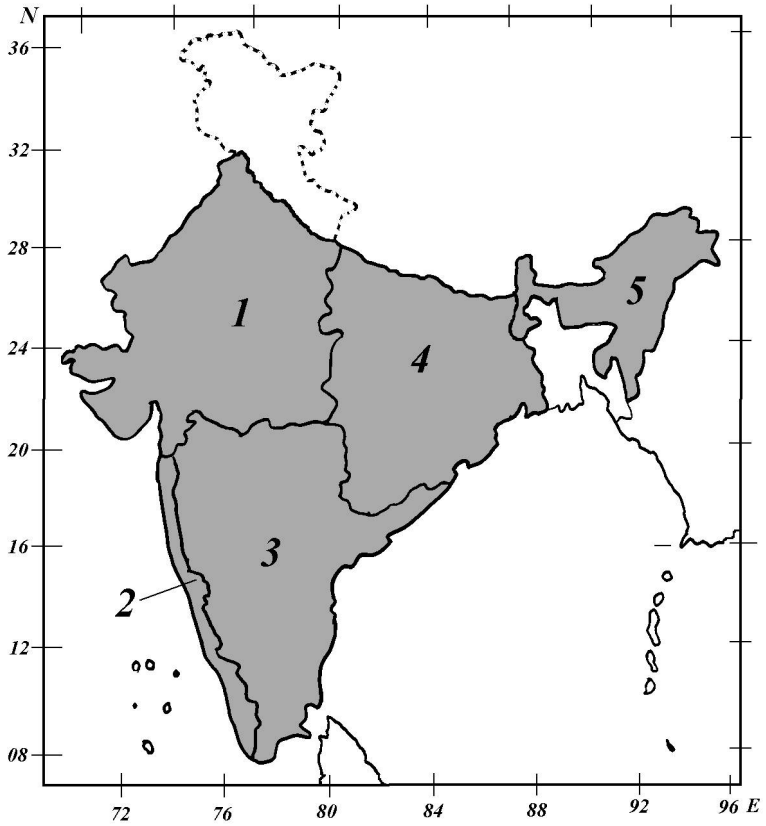

Fig. 1 : Shaded region shows the part of India for which the June-July-August rainfall has been attempted in the present study. Numerals indicate the indices of the rainfall homogeneous zones as they appear in final prediction equation.

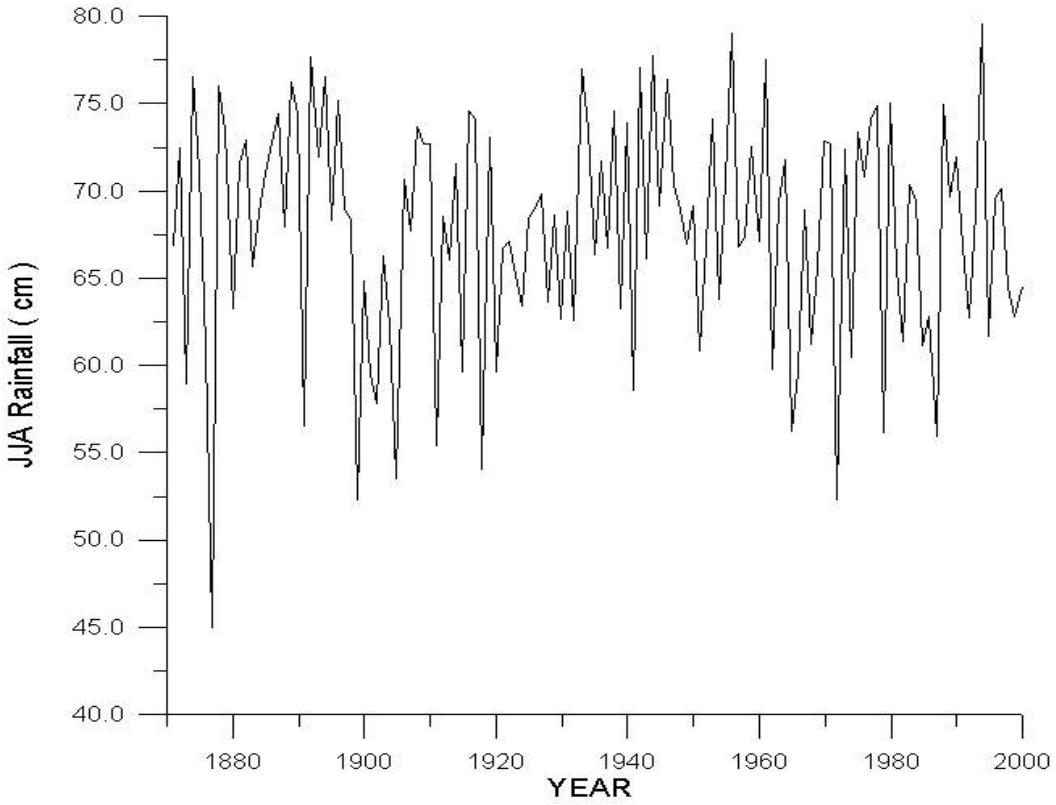

Fig. 2 : June-July-August rainfall for the years 1871-2000 over the region shown in Fig. 1

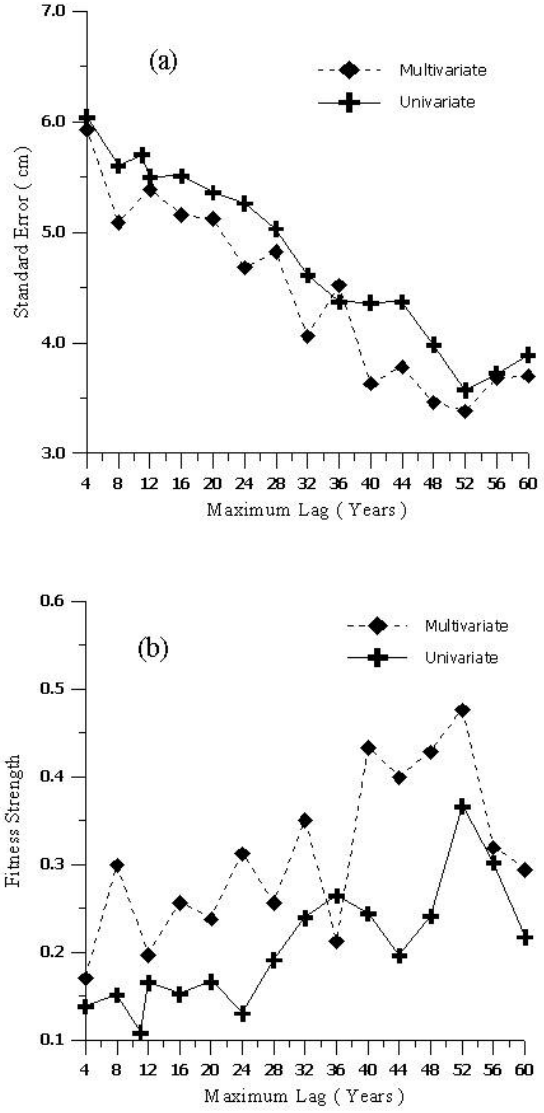

Fig. 3 : (a) Standard error and (b) fitness index for equations with different values of *m*, at the end of 1000 generations. Each equation was sufficiently long (38 terms).

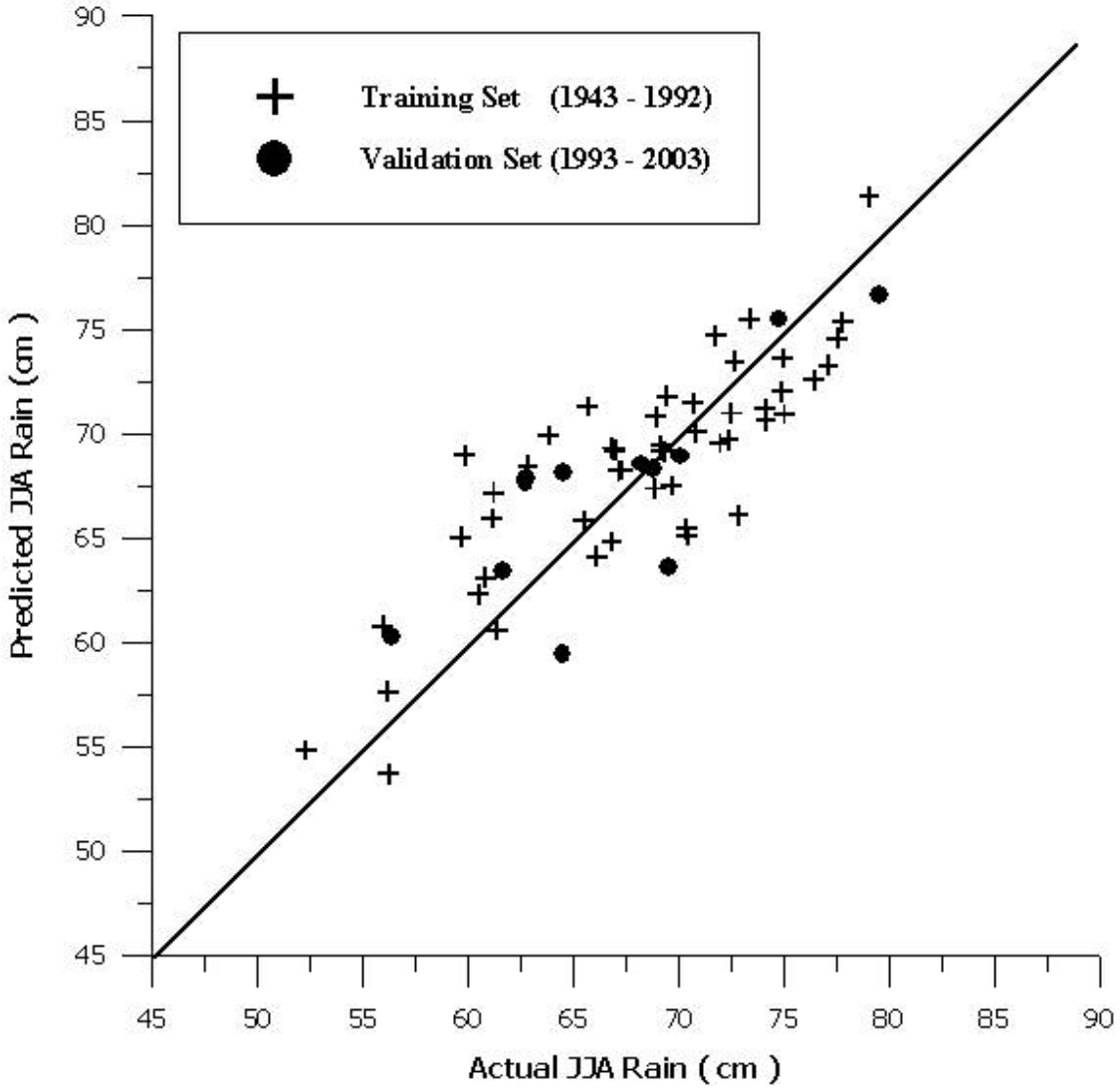

Fig. 4 : A comparison of observed and predicted rainfall for the years 1943-2003. Line of perfect fit (x=y) is also shown.