Estimation of Fractal Dimension of a Noisy Time Series

Muhammad Saleem Khan
Department of Mathematics
University of Karachi
Karachi, Pakistan

Tanveer Ahmed Siddiqui
Department of Mathematics
University of Karachi
Karachi, Pakistan

ABSTRACT
Estimation of the fractal dimension by using correlation dimension of precipitation time series play a fundamental role in the development of dynamic models of meteorological phenomena. As we know that the fractal dimension provides bounds for the number of independent variables necessary to model the system. We computed the correlation dimensions by Takens algorithm, Grassberger and Procaccia algorithm and by R/S method which gives the lower bound. In this paper, the fractal dimension by the method of correlation dimension of 20-years monsoon daily rainfall time series from June to September of Lahore region is estimated. The simulation of our time series is also considered which is based on wavelet fractional Brownian motion (wfBm) as a model that exhibits the self-similarity.

General Terms
Fractals, Time series analysis

Keywords
Fractal dimension, correlation dimension, lower bound, simulation, wavelet fractional Brownian motion.

1. INTRODUCTION
We analyzed twenty years monsoon daily rainfall time series from June to September of Lahore region, taken by Pakistan Metrological department from 1986 to 2005. The time series plot is given in figure 1. There is a small negative trend in the time series as shown in the figure; hence for analyzing the time series, we removed the trend by using first difference method.

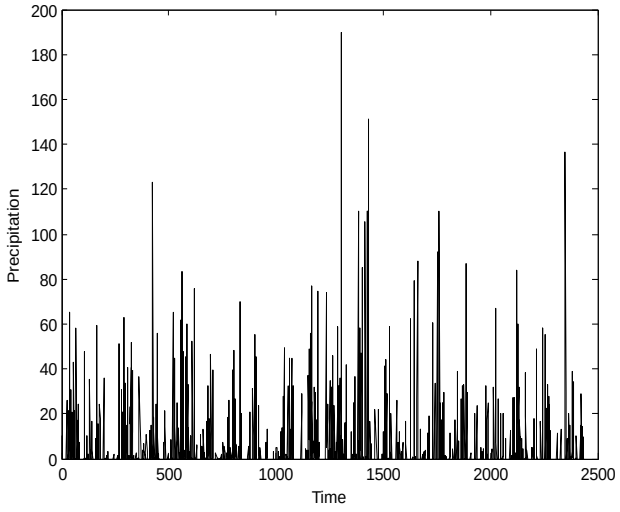

Fig. 1: Time series plot of Lahore precipitation data (1986 – 2005)

Generally, the nonlinear time series is analyzed by its phase space portrait. The phase portrait of a dynamical system can be reconstructed from the observation of a single variable by the method of delays as proposed by [1]. According to Takens, almost all d-dimensional sub-manifolds could be embedded in a (m=2d+1) dimensional space while preserving geometrical invariants. The observational time series X(1),X(2),….X(N) is represented by the set of vectors $X(t) = [X(t), X(t+\tau), \ldots X(t+(m-l)\tau)]$ (where t=1,2...N-[m-1] τ) in the reconstructed phase space.
The time delay τ is usually chosen by autocorrelation function or average mutual information (AMI). Time delay is estimated where the value of the autocorrelation function is close to zero, thus minimizing the statistical dependence among the coordinates of the vectors while the AMI is the standard way to calculate time delay τ. In practice, one does not know a priori the dimension of the dynamical system, and the embedding dimension, which is necessary for the phase space reconstruction. So, the dimensional estimate is computed for increasing embedding dimensions until the dimensional estimate stabilizes.

2. MATERIAL AND METHODS

2.1 Phase Space Reconstruction
One approach of investigating nonlinear behavior is by embedding a time series. A point in such a space is then associated with a single state of the system, which is fully defined by a set of m dynamical variables. When an experimental dynamical system is investigated, these m physical quantities should be all measured, at least in principle, to have a complete description of the state of the system under study. The next step is therefore to reconstruct a phase space from this scalar time series. The reconstructed trajectory can have the same features as the original trajectory in the phase space.

A pioneering paper by [2], points out two ways of reconstructing a phase space, namely, by using time delay or time derivative coordinates. Another kind of coordinates, namely principal components, may also be used. [3] demonstrated that the relationships between delays, derivatives and principal components consist of rotation and rescaling. Consequently, from [3]'s point of view, statements about the nature of the equivalence between the original and the reconstructed phase portraits would not depend on the coordinate system.

It is sometimes wanted to obtain a model able to reproduce the trajectory in the reconstructed phase space. Phase space is an abstract mathematical space spanned by the dynamical variables of the system. A point in this phase space can represent the state of the dynamical system at a given instant in time. If there are n dynamical variables then the state at a given time can be represented by a point in the Euclidean space $\Re^n$. As the dynamical variables change their values in time, the representative point traces out a path in the phase space—a continuous curve in the case of a continuous

dynamical system and a sequence of points in the case of a discrete dynamical system.

Those dynamical systems, which are dissipative and exhibiting chaotic behavior generally, display strange attractors in the phase space [4, 5]. Any time series generated by a nonlinear process can be considered as the projection on the real axis of a higher-dimensional geometrical object that describes the behavior of the system under study [6]. The most common method used for phase space reconstruction of this object relies on the so-called Delay Embedding Theorem [1, 7]. This theorem states that a series of scalar measurements X(t) can be used in order to define the orbits describing the evolution of the states of the system in an m-dimensional Euclidean space. The orbits will then consist of points X (t) with coordinates

$$\boldsymbol{X}(t) = [X(t), X(t+\tau), ...X(t+(m-1)\tau)] \quad (1)$$

Where τ is the delay time and the dimension m of the vector is known as the embedding dimension. A new time series of the state space vectors X (1), X (2)… …… X (n) is generated from Eq. (1). Each vector X (i) describes a point in an m dimensional phase space. Thus, the sequence of these vectors defines a trajectory in time as shown in Fig. 2.

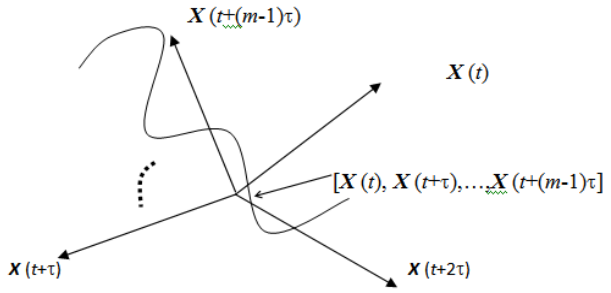

Fig. 2: The diagram of evolution in m-dimensional phase space.

Geometrically, the entire set of these points forms a pattern, termed an attractor, in the phase space. According to [1], if d is the dimension of the original attractor, it is sufficient that the embedding phase space dimension m must be greater than or equal to 2d+1. However, in reconstructing an attractor from a time series of unknown dynamics, the dimensionality of the attractor is unknown.

2.2 Average mutual information (AMI)

We apply Average Mutual Information method for determining the delay time. This method defines how the measurements X (t) at time t are connected in an information theoretic fashion to measurements X(t+τ) at time (t+τ) [8]. The average mutual information is given mathematically as:

$$I(\tau) = \sum_{X(i),X(i+\tau)} P(X(i), X(i+\tau)) \log_2 \left[\frac{P(X(i), X(i+\tau))}{P(X(i))\, P(X(i+\tau))} \right] \quad (2)$$

Where *i* is the total number of samples. P(X(i)) and P(X (i+τ)) are individual probabilities for the measurements of X(i) and X(i+τ). P(X (i), X (i+τ)) is the joint probability density for measurements P (X(i)) and P (X(i+τ)). The appropriate time delay τ is defined as the first minimum of the average mutual information I(τ). Then the values of X(i) and X(i+τ) are independent enough of each other to be useful as coordinates in a time delay vector but no so independent as to have no connection with each other at all.

Now we apply AMI technique to our climate data for the calculation of time delay using TISEAN software [9]. We plot AMI of the daily summer rainfall of Lahore after removing trend as in fig. 3. As the first minima occurs at τ = 4, we inferred that delay time for our climate data is 4 days.

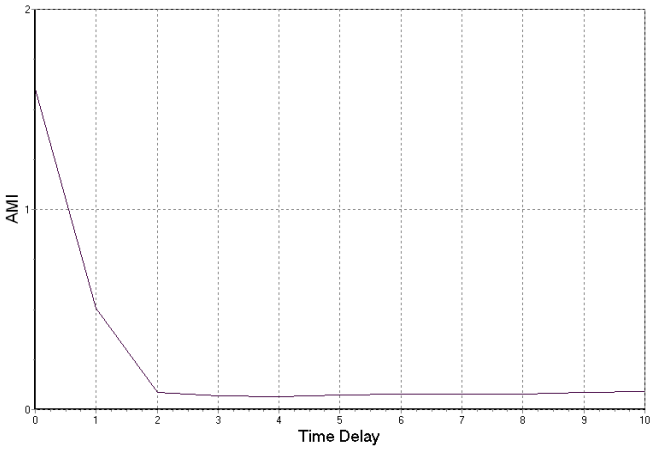

Fig. 3: Average mutual information plot of trend removed time series.

2.3 Embedding dimension

To estimate a best possible value of embedding dimension m is to check for the closed false neighbors (FNN) in the trajectory of phase space at different value of m [10]. Kennel et al., [10] developed an algorithm that estimates the sufficient dimension for phase space reconstruction. It is known as the false nearest neighbor method. The false nearest neighbor algorithm identifies points within a nonlinear time series that looks to correlate, or relate, at a certain point in time. By increasing the embedding dimension, m, it is possible to detect "false neighbors" within the vectors because once the attractors unfold; the vectors close in dimension m, move a significant distance apart in the next state. This indicates that the attractors of the system have not been accurately identified then embedding dimension is increased by one or both the vectors and its neighbor by increasing the appropriate value of the data.

FNN elucidate mathematically, as for each point X(i) in the time series look for its nearest neighbor X(j) in an m-dimensional space. Calculate first the distance $|X(i) - X(j)|$. Then, iterate both points and compute

$$R(i,m) = \frac{\left\| X\ (i, m+1) - X\ (j, m+1) \right\|}{\left\| X\ (i, m) - X\ (j, m) \right\|} \quad (3)$$

If R (i, m) exceeds a given heuristic threshold say R(t), this point is marked as having a false nearest neighbor. The criterion that the embedding dimension is high enough is that the fraction of points for which R(i, m) > R(t) is zero, or at least sufficiently small [9]. In order to apply this method, we

select a possible value of the distance $\|X(i) - X(j)\|$ in TISEAN software to the present time series. In our case the embedding dimension is 18 of original noisy time series because the first zero occurs at 18 as shown in the following figure 4.

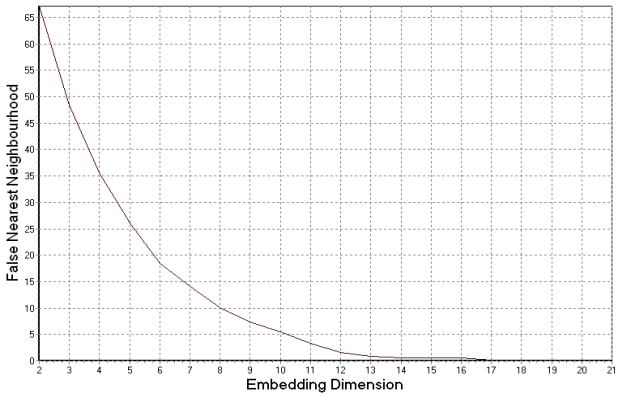

Fig.4: False Nearest Neighbor-hood of trend removed time series.

2.4 Effect of Noise on Correlation Dimension Estimation

Apart from determining the sufficient embedding dimension, the false nearest-neighbor method is also used as an indicator of the amount of noise in our data. In above graph, false-neighbors are being converge slowly in the range m= 7 to 18. Hence, there is high probability for random noise, which is responsible to spread the data. Therefore, it needs to be filter. As a stochastic process, noise should have infinite degrees of freedom and therefore it should show no tendency to unfold at any specific dimension. Thus we were able to eliminate events that showed a slow rating at minimum embedding dimension.

Moving average and low-pass filter are commonly used methods for noise reduction. In the present study, however, we use a nonlinear locally projective noise reduction scheme specifically developed for chaotic data as proposed by [11]. This algorithm gives the hypotheses that output data of the system is produced by a low dimension dynamical system but due to random noise, it exhibits a high dimensional output. If we suppose that the amplitude of the noise is sufficiently small, we can expect to find the data distributed closely around this manifold. The idea of the projective nonlinear noise reduction scheme is to identify the manifold and to project the data onto it. This algorithm is used in GHKSS TISEAN's command. Trend removed time series was embedded in eighteen dimension space (m18) and a delay time four (d4). Manifold dimension is set at three (q3) during noise reduction with three iteration (i3).

Time delay and embedding dimension are again calculated for noise reduced data obtained from GHKSS's command of TISEAN. Following figures 5 and 6 exhibit a delay time is 3 and the embedding dimension is 9 and figure 7 shows the phase space of noise reduced time series with the time delay 3 in the embedding dimension of 3. It shows the trajectory of vectors at different times. The trajectories are exhibiting the random pattern in the time series and could be unfold in high dimension.

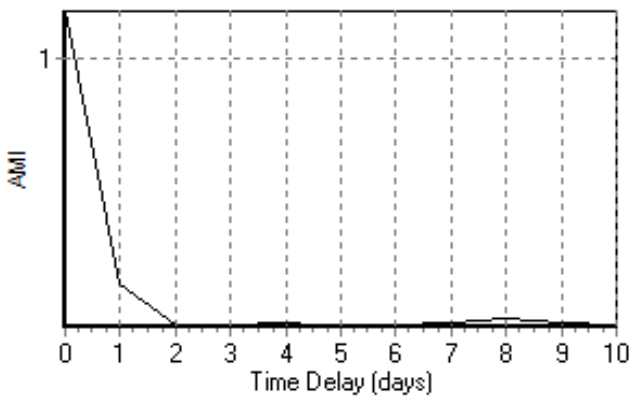

Fig.5 AMI plot of noise reduced data.

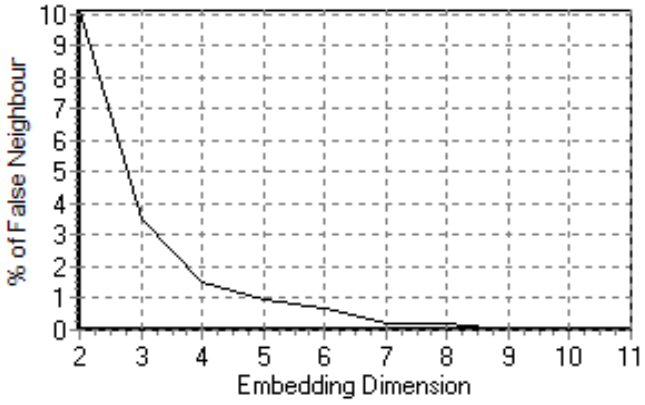

Fig. 6: Percentage of false nearest-neighbor plot.

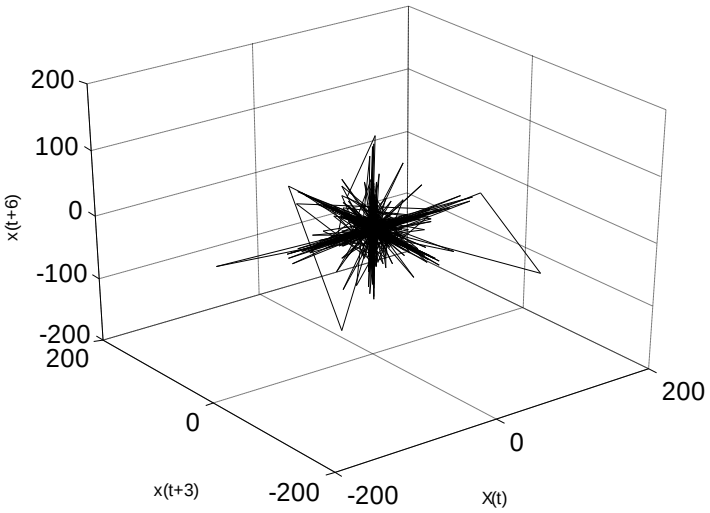

Fig. 7: 3-D representation of phase space of noise reduced data with time delay 3.

2.5 Space-Time separation plot

Any physical quantity can never be known exactly so it is appropriate to provide a confidence interval of the estimated quantity or in other words to identify the errors in estimation. The primary source of errors is Geometrical errors, Dynamical errors and Statistical errors. Dynamical errors are caused by certain properties of the trajectories that are related with dynamics e.g. autocorrelation effects, finite correlation dimension by certain types of filtered noise.

Space time separation plot (STSP) was introduced by [12]. The algorithm of STSP is based on to detect the behavior of r-neighboring points of the current point with respect to time and distance apart at different constant probability of time.
As a remedy, apply Theiler's correction (window) [13] in the calculation of correlation sum C(r) by using space-time separation plot (STSP). It is also used to determine the stationarity of the time series data. This helps identifying temporal correlations inside the time series and is relevant to estimate a reasonable delay time, and, more importantly, the Theiler-window in correlation dimension. In others words, it shows how large the temporal distance between points should be so that we can assume that they form independent samples according to the invariant measure.

The following space-time separation plot of the noise reduced rainfall data is given in fig. 8 and it shows stationarity in the data as lines of constant probability are horizontal. The low temporal variability is dominant in this system and is reflected in fig. 8. The temporal correlation is evident within the first 10 time steps where the lines increase consistently. To be safe, we have chosen the minimum correlation time to be 10. After 10 time steps the lines of each constant probability shows the stationary pattern in space-time separation plot.

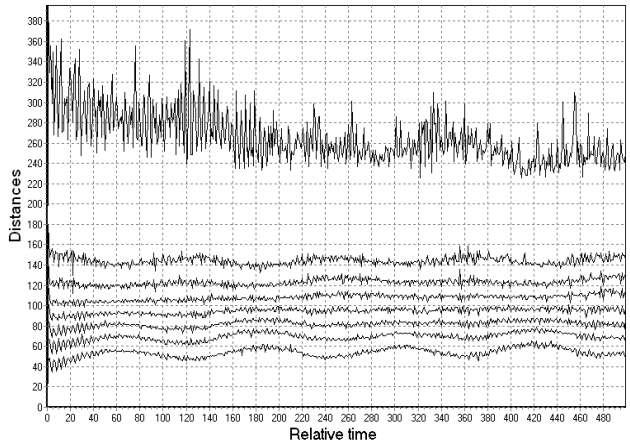

Fig 8: Space-time separation plot of noise-reduced data. Probability densities are 1/8 to 1 with increments of 1/8 from bottom to top.

2.6 Fractal Dimension Analysis

2.6.1 Correlation Dimension

In section 2.1, we constructed a phase space of the attractor of our time series using time delays theorem and estimated embedding dimension m. That value of m could be possible value of the variables in order to make a model of the system. Correlation dimension provides the dimensionality of the dynamical system occupied by the random points and a lower bound to the fractal dimension. Thus it satisfies the following inequality

$$d_c \le d_f \tag{4}$$

There are many ways to define the fractal dimension, but one of the numerically simplest and most widely used is the correlation dimension by [4, 5].The correlation dimension is obtained by considering the cumulative correlation function, defined by [4] as

$$C(r) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{\substack{k,j=1 \\ k\neq j}}^{N} \Theta(r - \|X(k) - X(j)\|) \tag{5}$$

Where r is finite length scale and Θ is the Heaviside function, such that Θ (x) = 0 if x ≤ 0 and Θ (x) = 1 if x > 0. The idea is to construct a correlation sum function C(r) that is the probability that two arbitrary points on the orbit are closer together than r. The Euclidean norm utilized in (5) is defined as

$$\|X(k) - X(j)\| = \sqrt{\sum_{i=1}^{m} (x_{k+(i-1)\tau} - x_{j+(i-1)\tau})^2} \tag{6}$$

The cumulative correlation function is related to the correlation dimension by the power law

$$C(r) \sim r^{d_c}, \qquad where \quad r \to 0 \tag{7}$$

This results in the correlation dimension dc being defined as

$$d_c = \lim_{r\to 0} \frac{\ln C(r)}{\ln r} \tag{8}$$

For the correlation dimension, C(r) verses r is plotted on a log-log graph, and the gradient of the region of the graph, which exhibits scaling behavior, yields an estimate of the correlation dimension. This is usually done by calculating the separation between every pair of N data points and sorting them into bins of width dr proportional to r. The correlation dimension in the limit

$$dr \to 0, \ as \ \ r \to 0, \ N \to \infty$$

is given by

$$d_c = \frac{d \log[C(r)]}{d \log(r)} \tag{9}$$

We used TISEAN software to evaluate correlation dimension of noise reduced data setting time delay 3, embedding dimension 20, Thieler window 10, minimal length scale 0.4 in d2 command. Figure 9 shows the plot slopes of correlation sum against logarithm of r. The linear part of the graph is estimation of correlation dimension, which has the value ≈1.54.

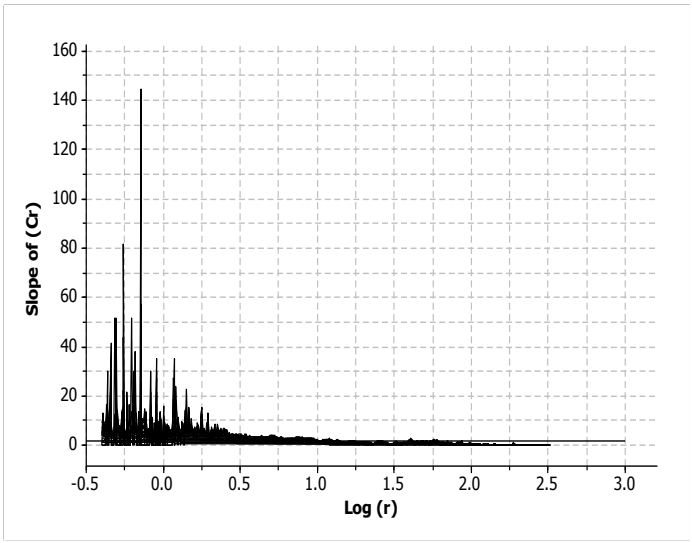

Fig 9: Slopes of correlation sum C(r) by Grossberger and Procaccia algorithm.

2.6.2 Takens Estimator

Another algorithm to calculate correlation dimension is to find Takens estimator. It was suggested by [14] and [12]. It requires maximal cutoff distance R on which all pair wise distances larger than R are discarded and for correlation dimension, a maximum likelihood estimator D_T is calculated by

$$D_T(r) = \frac{C(r)}{\int_0^r \frac{C(r')}{r'} dr'} \qquad (10)$$

D_T is an alternative to the usual local slope. D_T can be obtained from TISEAN command c2t by processing on r and C(r) obtained either from c2naive or c2 file of d2 output file.

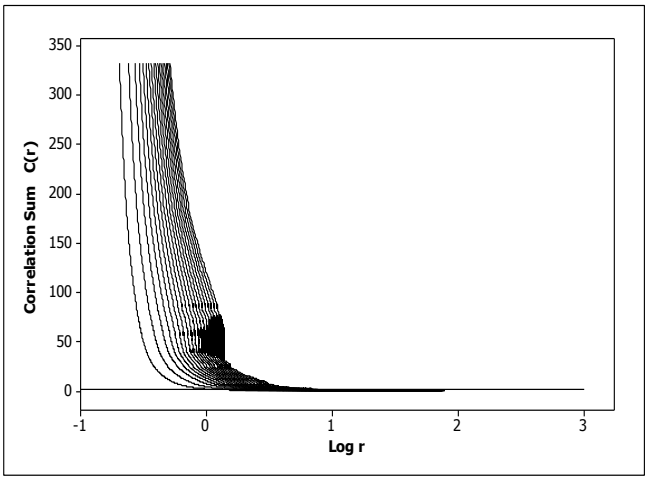

Fig 10: Correlation dimension by Taken method

We used c2t on c2-file obtained from d2 command and obtained D_T with r. Figure 10 shows the plot of D_T against log(r). The value of the correlation dimension by Taken estimator is the value of the linear portion in the graph which has the value ≈1.95.

2.6.3 Rescaled Range (R/S) Analysis

Harold Edwin Hurst in 1951 developed this algorithm in his Nile's study [15]. R/S analysis of a time series is used to find future patterns that might be repeated. In this analysis there are two main inputs, rang and standard deviation of the data. A derivative of this mathematical result is known as a Hurst exponent (H); if a trend actually exists in the data, this Hurst exponent can extrapolate a future value or average for the data point. As we know that many natural phenomena previously suppose to be governed by random Gaussian process shows long term statistical dependence. H (0<H<1) describes measure of degree of correlation.

The relation between Hurst exponent (H) and correlation dimension (D) is D = 2 – H.
The Hurst exponent (H) of our noise reduced daily rainfall time series (2440 points) is 0.06 i.e. D=1.94.

2.7 Wavelet Fractional Brownian motion (wfBm)

Fractional Brownian motion (fBm) is a continuous and self similar in distribution time stochastic process depending on the Hurst parameter H, proposed by Mendelbort and Van Ness [16]. It consists in fractional integration of a white Gaussian process and is therefore a generalization of Brownian motion because it produces a deep connection with concept of self-similarity and fractal [17]. Hence, the variance of the increments of fBm is given by

$$Var[\, fBm(t) - fBm(s)] = \alpha |t-s|^{2H} \qquad (11)$$

Where α is the positive constant. [17] developed an effective wavelet-based algorithm of fBm called wfBm. Because, the formulation of the fBm process as a fractional integral has to be started for the white noise process, the idea of the algorithm is to build a bi-orthogonal wavelet depending on a given orthogonal one and adapted to the parameter H. Then the generated sample path is obtained by the reconstruction using the new wavelet starting from wavelet decomposition at a given level and exact coefficients are independent random Gaussian realizations and approximation coefficients come from a fractional ARIMA process.

In this paper, simulation is performed which is based on Abry and Sellan algorithm [17] that is used to generate samples of wfBm for the Hurst exponent (H=0.06) and 2440 data points of the noise reduced rainfall time series and a fractal dimension estimated by the R/S analysis to be 1.94 is exhibited in the following figures (fig. 11 and 12) show the same fractal dimension estimate. The similarity is obvious.

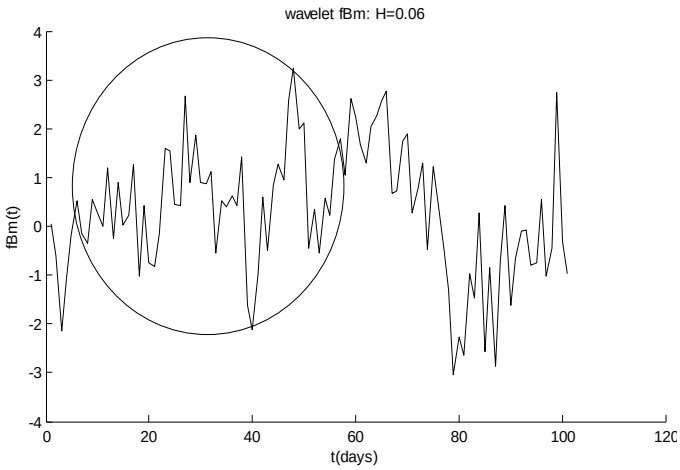

Fig. 11: A sample of wfBm with parameter H=0.06 and D=1.94

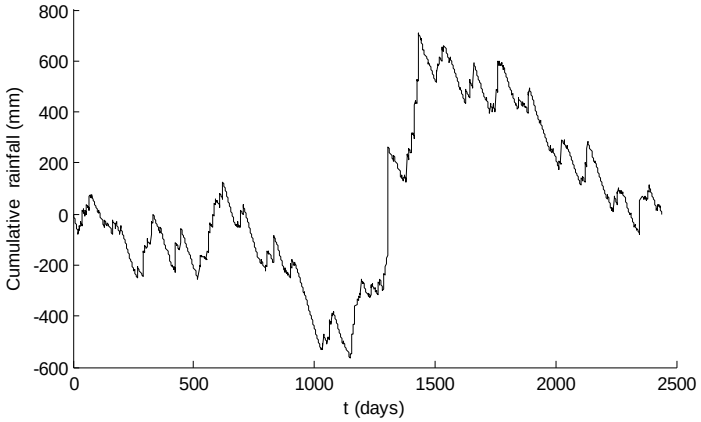

Fig. 12: A plot of rainfall time series from Lahore, Pakistan

3. CONCLUSION

Since the value of the fractal dimension of phase space depends on noise level. So noise must be reduced because it increases the value of fractal dimension. In our case, by comparing the algorithms of fractal dimension, Takens algorithm and R/S analysis are approximately equal to 1.95 of our noise reduced monsoon daily rainfall time series of twenty years which is unexpectedly low. Because the correlation dimensional estimate provides the lower bound for the number of independent variables for necessary to model the system and assists one to determining the appropriateness of the model, in this case the minimum number of independent variables and the number of equations required to model the rainfall dynamical system are 2 which is an important estimation for modeling the system.

Our rainfall time series shows the self-similarity i.e. the daily fluctuations exhibit the same statistical behavior over different scales of time. Analysis needs to be made for different limit of scales to be self-similar. For considering the simulation for rainfall time series, wfBm is possible as a model.

4. REFERENCES

[1] Takens F., Detecting strange attractors in turbulence, Dynamical Systems and Turbulence, Warwick 1980 lecture Notes in Math. (1981) 366-381.

[2] Packard N. H., Crutchfield J. P., Farmer J.D., Shaw R.S., Geometry from a time series, Phys. Rev. Lett. 45 (1980) 712–716.

[3] Gibson J.F., Farmer J.D., Casdagli M., Eubank S., An analytic approach to practical state space reconstruction, Physica D. 57 (1992) 1–30.

[4] Grassberger P., Procaccia I., Measuring the strangeness of strange attractors, Physica D. 9 (1983a) 189–208.

[5] Grassberger P., Procaccia I., Characterization of strange attractors, Phys. Rev. Lett. 50 (1983b) 346–349.

[6] Kantz, H., Schreiber, T., Nonlinear Time Series Analysis, Cambridge University Press Cambridge (1997).

[7] Sauer T., Yorke J., Casdagli M., Embedology, J, Suitable delay times for continuous systems, Stat. Phys. 65 (1991) 579.

[8] Fraser A. M., Swinney H. L., Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33, 1134 (1986).

[9] Hegger R., Kantz H. H., Schreiber T., 'Practical implementation of nonlinear time series methods, The TISEAN package Chaos 9 (1999).

[10] Kennel M.B., Brown R., Abarbanel H.D.I., Determining embedding dimension for phase-space reconstruction using a geometrical construction, Phys. Rev. A. 45 (1992) 3403.

[11] Grassberger P., Hegger R., Kantz H., Schaffrath C., Schreiber T., On noise reduction methods for chaotic data, Chaos 3 (1993) 127.

[12] Provenzale L.A., Smith R. and Murante G., Distinguishing between low-dimensional dynamics and randomness in measured time series, access this article Physica D. 58 (1992) 31-49.

[13] Theiler J., Lacunarity in a best estimator of fractal dimension, Phys. Lett. A. 135 (1988) 195.

[14] Takens F., Braaksma B. L. J., Broer H. W., Dynamical Systems and Bifurcations, Lecture Notes in Math. Spring. (1985) 1125.

[15] Hurst, H., 1951, "Long Term Storage Capacity of Reservoirs," Transactions of the American Society of Civil Engineers, 116, 770-799.

[16] Mandelbrot B.,Van Ness J. W., Fractional Brownian motions: Fractional noises and applications, SIAM Rev. 10, No. 4 (1968), 422-437.

[17] Abry, P.; F. Sellan (1996), "The wavelet-based synthesis for the fractional Brownian motion proposed by F. Sellan and Y. Meyer: Remarks and fast implementation," *Appl. and Comp. Harmonic Anal.*, 3(4), pp. 377-383.