Universal BHP distribution and nonlinear prediction in complex systems using the Ruelle-Takens embedding

R. Gonçalves[a] and A. A. Pinto [b]

[a]Faculdade de Engenharia da Universidade do Porto,
4200-465 Porto, Portugal

[b]Escola de Ciências, Universidade do Minho,
Campus de Gualtar, 4465- Braga, Portugal

We use nonlinear dynamics techniques in a complex non-deterministic dynamical setting. Our object of study is the observed riverflow time series of the Portuguese Paiva. The Ruelle-Takens delay embedding of the daily riverflow time series revealed an intermittent dynamical behavior due to precipitation occurrence. The laminar phase occurs in the absence of rainfall. The k-nearest neighbor (k-nn) method of prediction revealed good predictability in the laminar regime. In the presence of rainfall the k-nn method revealed to be misleading. We present some new insights between the quality of the prediction in the laminar regime, the embedding dimension, and the number of nearest neighbors considered. We find, unexpectedly, that the BHP distribution is an approximation of the empirical distribution of the relative first difference.

1. Introduction

A direct link between the real world and deterministic dynamical systems theory is the analysis of real systems time series in terms of nonlinear dynamics with noise. Great advances have been made to exploit ideas of dynamical systems theory in cases where the system is not necessarily deterministic but it displays a structure not captured by classical stochastic methods. The application of dynamical systems methods found a firm ground on the reconstruction theorem of Ruelle-Takens [20] and in the probabilistic justification due to Sauer, Yorke and Casdagli ([18]). The motivation for applying methods of deterministic dynamics in riverflow time series lies in the natural tendency of river systems to present recurrent behavior (see [14], [19], [2], [11], [5]). We start by doing a Takens delay coordinates reconstruction of the daily flow series indicating that Paiva river is an intermittent dynamical system. This intermittent dynamical behavior is characterized by a laminar and an irregular phase. The laminar phase occurs in the absence of rainfall and the irregular phase occurs under the action of rain. Hence, the forcing of the dynamical system is not of a deterministic type because rainfall is stochastic.
We present some new insights between the quality of the prediction, the embedding dimension, and the number of nearest neighbors considered (see [12]). The nearest neighbor method of prediction reveals good predictability in the laminar regime. We compute the

mean of the relative predicted first difference, for different regimes and embedding dimensions. The nearest neighbor method of prediction does not exhibit good predictability in the irregular phase indicating the stochastic predominance over the deterministic in the irregular phase. Since 75% of data is laminar, we warn that the use of nonlinear deterministic prediction methods can be just misleading when both dynamical regimes are considered. In the laminar regime, the nearest neighbor method of prediction indicates that the dynamics can be approximated by a one to three dimensional dynamical system. The prediction results revealed that it is essential to know the current runoff to predict future values. In [3], we use these results to reconstruct an approximation of the one-dimensional dynamics of the runoff using different prediction estimators. In [3], we discovered that the empirical distributions of the relative first difference, for some runoff regimes, exhibit a good fit to the distribution BHP [1]).

2. Data analysis

The relevant data for this work consist of the time series of mean daily runoff of the Paiva river, measured at Fragas da Torre section, district of Beira-Alta. The sample period runs from October, 1st 1946 to September, 30th of 1999 for a total of 19358 observations (see chronogram of figure 1). The Paiva river is not an intermittent river in the sense that at the referred location and in the 53 years of observation the surface stream never disappeared.

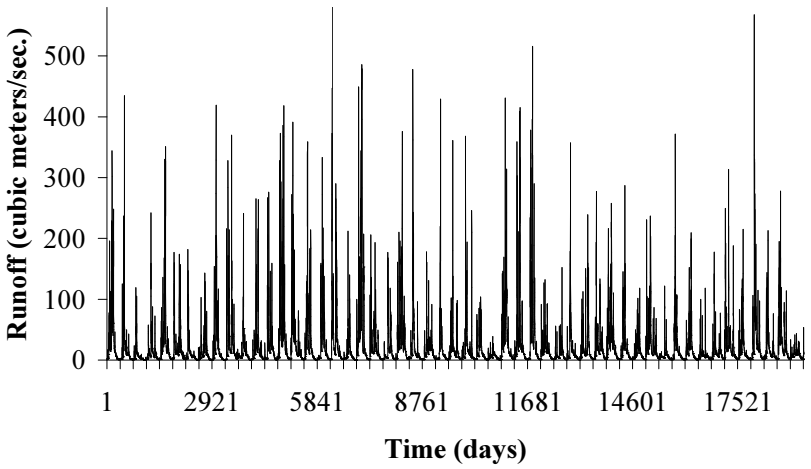

Figure 1. Chronogram of the daily mean riverflow of Paiva measured at Fragas da Torre 1946-99.

The daily river flow descriptive statistics, (see Tab. 1 and figure 1) shows the strong asymmetry of the data. We can see that there is a quite a difference in the flow between the summer and winter meaning that there are no glaciers or dams that may give water to

Table 1
Descriptive statistics for the daily mean riverflow series of Paiva (1946-99) measured at Fragas da Torre.

Statistic	Value
Mean	20.73 m^3/s
Median	$5.66 m^3/s$
Skewness	5.3
Kurtosis	45.98
Maximum	920.0 m^3/s
Minimum	$0.06 m^3/s$

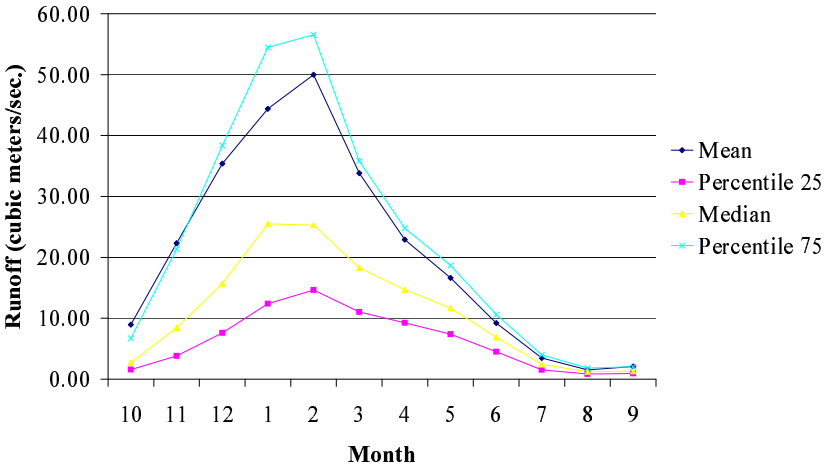

Figure 2. Evolution of Monthly Paiva daily mean runoff time series statistics.

the river in the summer (table 1, figure 1). The monthly river flow descriptive statistics, (figure 2) shows structural differences specially between the summer and winter revealing the dominant regimes of each season. The spring and autumn months reveal a kind of mixed behavior in terms of runoff statistics revealing a coexistence of the two regimes.

The sample autocorrelation function (figure 3) is characterized by the usual seasonality of this kind of data but also by an irregular behavior specially for the winter where the correlation is positive.

The average and the standard deviation for each day of the year (see figure 5) [1] show a strong statistical irregularity for each day of the year that increases with the runoff value.

[1]The 29th of February of each year were deleted.

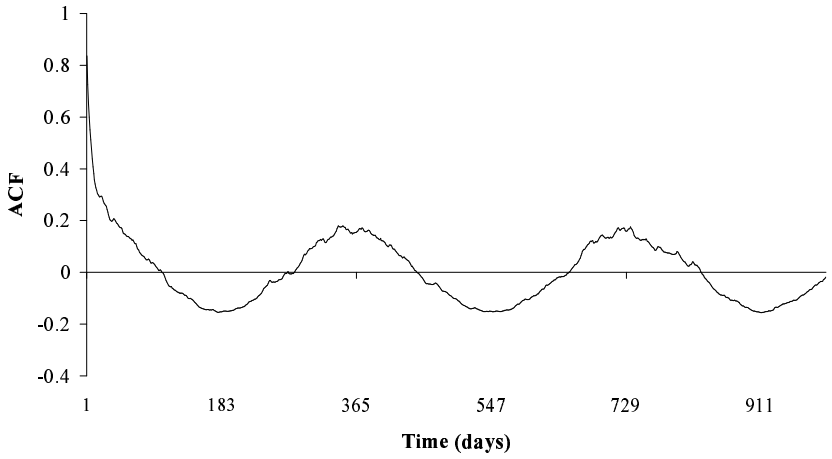

Figure 3. Sample Autocorrelation Function of the daily mean runoff series (1946-99) of the Paiva river at Fragas da Torre.

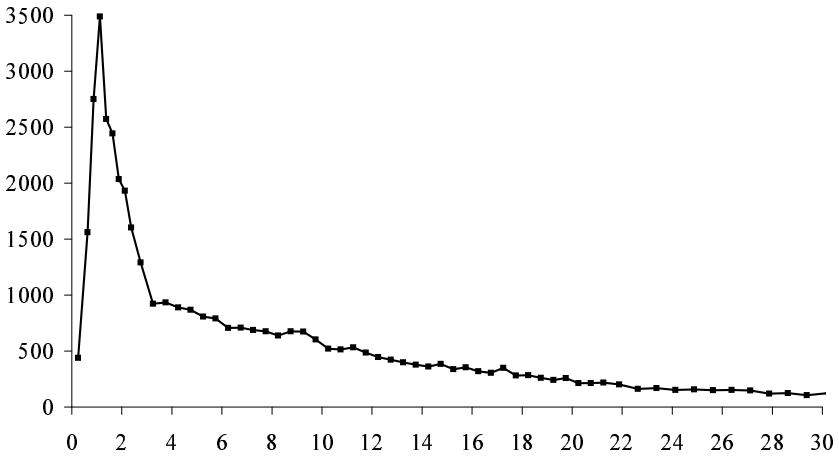

Figure 4. Histogram of the mean daily runoff series of Paiva river.

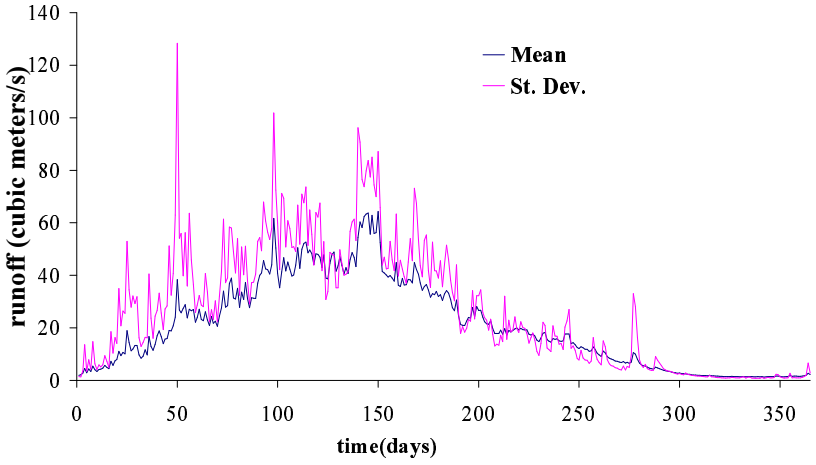

Figure 5. Mean and standard deviation for each day of the year.

3. Intermittent dynamics of Paiva

The dynamic characterization includes invariant estimation and in this direction we do a correlation-integral analysis for all the data and then we consider only the runoffs less than $20m^3/s$ which represents about 75% of the data corresponding mainly to the laminar phase, i. e., periods without rain. The Correlation Integral of a system is by definition the probability of finding fraction of points observed of a set of data is given by Eq. 1.

$$C_N^{(m)}(\varepsilon) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \Theta\left(\varepsilon - \left\| X^i - X^j \right\|\right) \tag{1}$$

where $X^t = (X_t, X_{t+1}, \ldots, X_{t+m-1})$ is a reconstructed vector which elements are values of the time series, $\{X_t\}_{t=1}^N$, N is the number of data points of the series, Θ the Heaviside function, ε the neighborhood radius and m the embedding dimension of the reconstructed phase space. The sample CI is a statistic used in the correlation dimension estimation, it was proposed by [4]. The sample CI is a statistic used in the the observed fraction of reconstruction vectors (RV) at a distance smaller than ε. The sum, (1), is computed for a set of distances, $\varepsilon_1, \ldots, \varepsilon_n$ evenly spaced on a logarithmic scale. A *scaling range* is said to exists if for such a range of values the sample correlation integral behaves like a power law or the same to say like a line on a $\log - \log$ scale. In practice there is a cut-off on the radius size due to data size restrictions. Defining $d(N, \varepsilon) = \partial \ln C_N^{(m)}(\varepsilon) / \partial \ln \varepsilon$, we have that

$$D_C = \lim_{\varepsilon \to 0^+} \lim_{N \to \infty} d(N, \varepsilon) \tag{2}$$

Hence, $d(N, \varepsilon)$ is the slope of the CI curve for a certain range, and D_C is then the estimate of the correlation dimension.
In figure 6, we present the correlation integral slopes. We can distinguish three different behaviors in the correlation-integral curve for different ranges of the radius, ε. For the runoff values larger than $30m^3/s$ no scaling range exists. For the runoffs in the interval $[5-30m^3/s]$ there is a scaling range which point towards a one-dimensional attractor. This dimension is not fractal and indicates that the behavior of riverflow for that range is close to that of a curve. This is a sign of the existence in the reconstructed phase-space of a one-dimensional manifold to which all the laminar phase orbits are close, i. e. the orbits are mainly contained in a small neighborhood of a one-dimensional curve. The singular value decomposition (SVD) analysis may be used to compare the two phases. The information given by this analysis is highly relevant for the understanding of the correlation integral curve. We take as vector variable the reconstruction vectors, $(X_t, X_{t+1}, \ldots, X_{t+m-1})$, where X_t is the daily mean runoff at day t. Using the *SPAD* statistical package to perform the SVD one computes the principal directions of the data set and their corresponding weights in the process the the principal factors are also calculated for the covariance matrix of the laminar phase for different embedding dimensions. The percentage(%) of the total variance explained by the three largest eigenvalues of the Covariance Matrix for the laminar phase are presented in Table 2.

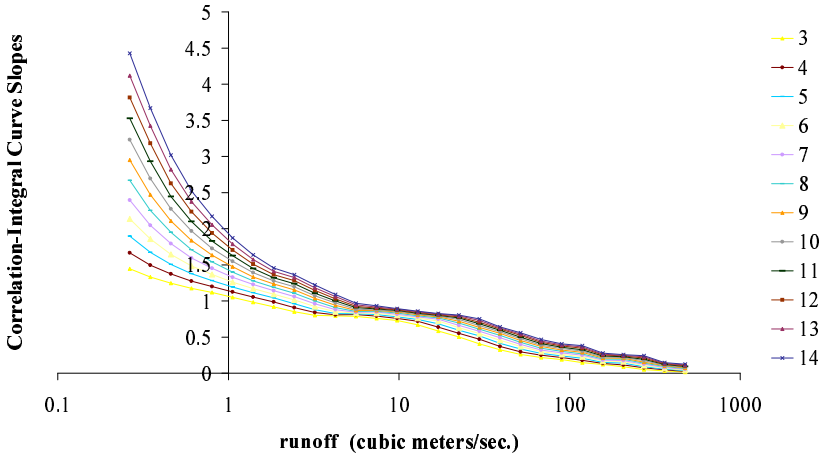

Figure 6. Slopes of the sample correlation integral curve of the Paiva river data and for several embedding dimensions.

Table 2
Percentage(%) of the total variance explained by the three largest eigenvalues of the Covariance Matrix for the laminar phase.

Dimension	%	%	%
3	96.34	2.69	0.96
4	95.01	3.23	1.13
5	93.82	3.70	1.33
6	92.75	4.13	1.50
7	91.75	4.52	1.66

The correlations between the principal components and the original variables are presented on the table 3. In the laminar phase there exists a principal component explaining more 90% of the variance. This is explained by the laminar dynamics being close to a segment line, figure 7. According to the usual criteria to quantify the number of significant eigenvalues, (see [17]) the reconstruction vectors (*individuals*) of the last two data sets are close to the one-dimensional case.

Table 3
Correlation between the 1st Principal Component and the original variables at the laminar phase and for several embedding dimensions.

Variable	3	4	5	6	7
X_t	0.98	0.97	0.96	0.95	0.94
X_{t+1}	0.99	0.98	0.98	0.97	0.96
X_{t+2}	0.98	0.98	0.98	0.98	0.97
X_{t+3}	-	0.97	0.98	0.98	0.97
X_{t+4}	-	-	0.96	0.97	0.97
X_{t+5}	-	-	-	0.95	0.96
X_{t+6}	-	-	-	-	0.94

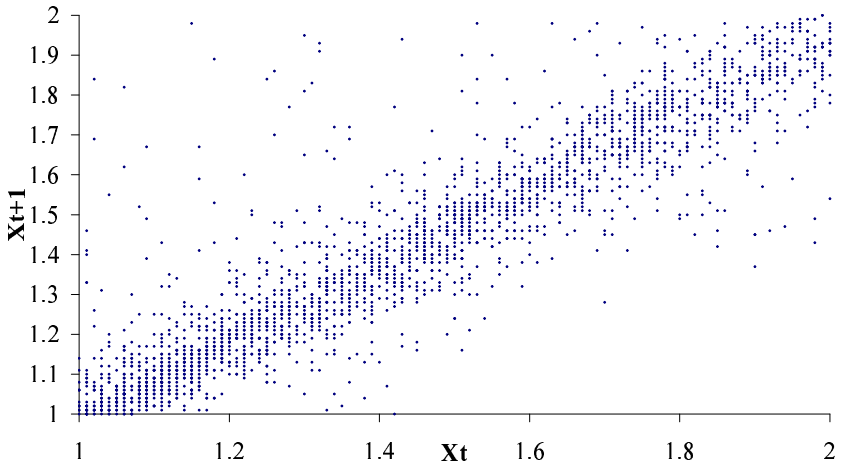

Figure 7. Phase space portrait for runoff values between 1 and 2 cubic meters/s.

4. Nonlinear prediction

4.1. Prediction results

Several authors used nonlinear prediction methods for river flow data locally in the phase space, [7], [8], [15], [11] and [5] among others. In this work, we use a different version of the k-nn neighbors method proposed by [9] to predict the next day runoff. Since our goal is to predict the runoff value during the laminar regime (absence of rain) we will filter appropriately the reconstruction vectors. Hence given the dimension, m, of the Ruelle-Takens embedding we consider only the reconstruction vectors $(X_t, X_{t+1}, \cdots, X_{t+m-1})$ satisfying the following δ-relative non-increasing rule, see Eq. 3.

$$X_{t+i-1} \leq X_{t+i}(1+\delta),\ 1 \leq i \leq m \tag{3}$$

where δ is a fixed positive value. In our approach the predicted value is the average of the phase-space images of the neighbors defined as below. Other authors, [11] reported that predictors based on the phase space average give better results than other local linear functions. Instead of using all the neighbors within a fixed radius we have used the the k-nn nearest neighbors with $k = 10$.

Table 4
Mean Square Error for the hydrological year 1997/98 one step ahead prediction and for different embedding dimensions (m) and time delays, (τ).

τ/m	3	4	5	6	7	8	9	10
1	8.98	7.36	7.08	7.39	8.38	8.35	8.63	8.70
2	10.87	12.33	11.72	12.84	12.55	14.01	14.30	14.53
3	12.05	11.43	13.04	13.60	13.28	13.49	13.82	15.50
4	8.93	11.01	13.35	13.37	11.99	12.36	12.42	13.10
5	18.63	17.08	20.32	19.11	17.76	21.08	20.67	20.53
6	16.57	16.74	16.92	18.15	14.64	15.09	14.58	14.84
7	20.95	22.27	26.13	26.85	26.64	25.25	25.66	26.20
8	17.88	17.57	17.96	16.72	17.68	20.37	20.54	18.38

The best Mean Square Error[2] (MSE) result was found for an embedding dimension 5 and for time delay of one day, (see table 4).

Here we should mention the paper on river flow prediction, [5] where the authors also obtained results of the same magnitude for the *MSE* for different embedding dimensions. As the results show, predicting locally in the phase space can be misleading if the system has an intermittent behavior.

[2]The Mean Square Error of Prediction (MSE) is defined by MSE $= \frac{1}{n}\sum_{i=1}^{n}\left(X_t - \hat{X}_t\right)$. When several models are proposed for the same data the ultimate choice of one may depend on goodness of fit such as the MSE.

Table 5
Mean Relative Error (MRE) and MSE for the one-step ahead prediction for the years 1995-99 and also for the laminar regime. Several embedding dimensions were considered.

m	MRE	MSE
1	0.0486	2.915
2	0.0328	1.004
3	0.0301	0.888
4	0.0310	1.034
5	0.0341	1.155
6	0.0378	1.193
7	0.0427	1.375

Table 6
Percentage of T_m for each neighbors set considered RV_1 - RV_8.

RV_i	1	2	3	4	5	6	7	8
Percentage (%)	0.5	1	5	10	12.5	15	20	25

4.2. Empirical distribution of the relative first difference

Here we will describe the empirical distribution of the relative first difference for the laminar phase. In this analysis, instead of using all the neighbors within a fixed radius, we use a fixed number of close neighbors, $RV_1, \ldots, RV_8$, see Tab. 6. We denote the total numbers, T_m, of filtered reconstruction vectors depending on the embedded dimension m considered. We will only consider reconstruction vectors whose future points to a non-increasing runoff. As a rule of thumb, we choose a 15% of increasing tolerance for runoffs lower than $4m^3/s$ and 5% for runoffs above that threshold. This distinction was based on the observed effects of measurement error.
In Figure 8, we present the total number of neighbors versus the Logarithm of the last coordinate of the reconstruction vector for embedding dimensions 1. The resulting curve show that there is a plateau for values in the range of $3.2m^3/s$ to $3.6m^3/s$.
The figure 8 gives us the variation of the relative distance of the predicted value, using the close neighbors method with respect to the RV's, runoff regimes. The relative distance of neighbors in the phase space increases with the runoff regime and also with the embedding dimension. Moreover, the mean for several regimes and number of neighbors (RV_i) considered increases approximately linearly with m.

In Figure 10, we show the histogram of the one step-ahead prediction for the neighbor sets RV_1 and for the runoff regimes considered, on top of the BHP pdf. The fit is good for the predicted case and for the regime 3 to 9 m^3/s.

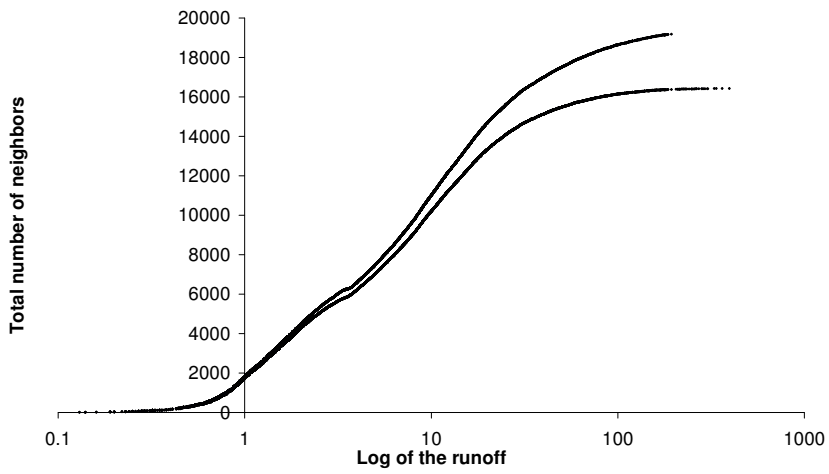

Figure 8. Total number of neighbors in the phase space vs the Logarithm of the last value of the reconstruction vector for embedding dimension 1, upper curve, and embedding dimension 3, lower curve.

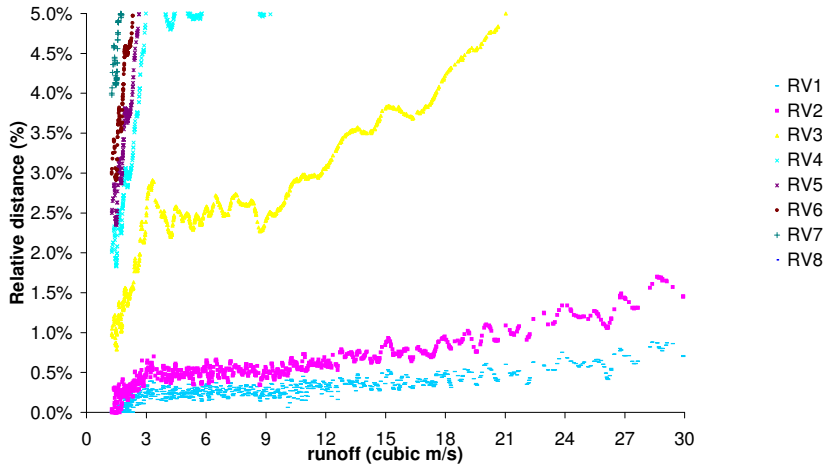

Figure 9. Relative distance (%) of RV's as function of the last coordinate for different neighborhoods and for dimension 1.

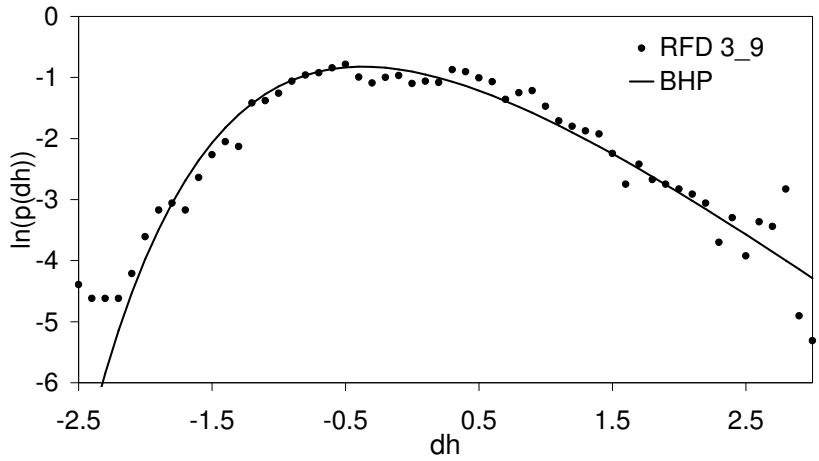

Figure 10. The universal BHP distribution on top of the Log of the predicted relative first difference histogram curve for the regime 3 to $9m^3/s$. The embedding dimension is 1.

5. Conclusions

A dynamical analysis of the Paiva river data was performed using the Ruelle-Takens method of dynamical reconstruction. We concluded that the Paiva river is an intermittent system. The laminar phase takes place in the absence of rainfall and the irregular phase occurred under the action of rain. The nearest neighbor method of prediction revealed good predictability in the laminar regime. However, since 75% of data is laminar, the use of nonlinear deterministic prediction methods can be just misleading when both dynamical regimes are considered. We studied the dependence of the nearest neighbors runoff predictor on the embedding dimension and on the relative average distance of the nearest neighbors with respect to the runoff value. The prediction results revealed that it is essential to know the current runoff to predict future values. We noticed small improvements in prediction when the former two runoffs are used. For a special regime, the distribution of the relative first difference predictor exhibited a good fit to the universal distribution BHP.

REFERENCES

1. Bramwell, S.T., Holdsworth, P.C.W., & Pinton, J.F. (1998) Universality of rare fluctuations in turbulence and critical phenomena *Nature* **396**, 552–554.
2. R. Gonçalves, A. A. Pinto & F. Calheiros in *Differential equations, Chaos and Variational problems*, (Birkhäuser, Staicu, Vasile (Ed.), 231, 2007).
3. Gonçalves, R., Pinto, A. and Stollenwerk, N. The BHP universal pdf as a measure of uncertainty of the rivers Paiva and Douro. In preparation.
4. Grassberger, P. and Procaccia, I., *Measuring the strangeness of strange attractors*, Physica 9D, **9**, (1983), 189–208.
5. Islam, S. and Sivakumar, B. Characterization and prediction of runoff dynamics: a nonlinear dynamical view. *Advances in Water Resources*, **25**, 179–190, (2002)
6. Jayawardena, A., Gurung, A., Noise reduction and prediction of hydrometereological time series dynamical systems approach vs stochastic approach. *Journal of Hydrology*, **228**, 242–64, (2000)
7. Jayawardena, A. W. and Lai, F., Analysis and prediction of chaos in rainfall and stream flow time series, *Journal of Hydrology*, **153**, 23–52, (1994)
8. Jayawardena, A., Li, W. and Xu, P., Neighbourhood selection for local modelling and prediction of hydrological time series. *Journal of Hydrology*, **258**, 40–57, (2002)
9. Kantz, H. and Schreiber, T., *Nonlinear Time Series Analysis*. Cambridge Univ. Press, (1997)
10. Kennel, M., Brown, R. and Abarbanel, H., *Determining minimum embedding dimension for phase space reconstruction using a geometrical construction.*, Phys. Rev. E **48(3)**, (1992), 1752–1763.
11. Liu, Q., Islam, S., Rodriguez-Iturbe, I. and Le, Y. Phase-space analysis of daily streamflow: characterization and prediction. *Advances in Water Resources*, **210**, 463–475, (1998)
12. Lorenz, E., Atmospheric predictability as revealed by naturally ocurring analogues. *Journal of Atmospheric Sciences*, **20**, 636-646, (1969)
13. Porporato, L. and Ridolfi, L., Clues to the existence of deterministic chaos in river flow. *Int. J. of Mod. Phys. B.*, **10**, 1821–1862, (1996)
14. Porporato, L. and Ridolfi, L. Nonlinear analysis of a river flow time sequences. *Water Resources Research*, **33**,1353–1367 (1997)
15. Porporato, A. and Ridolfi, L., Multivariate nonlinear prediction of river flows, *Journal of Hydrology*, **248**, 109–122, (2001)
16. Rodriguez-Iturbe, I., De Power, F., Sharifi, M. and Georgakakos, K., Chaos in rainfall, *Water Resouces Association* **25(7)**, 1667–75, (1989)
17. Saporta, G.*Probabilités, Analyse des Donnes Et Statistique.* Editions Technip, Paris, (1990)
18. Sauer, T., Yorke, J. and Casdagli, M., Embedology. *Journal of Statistical Physics*, 65, 579-616, (1991)
19. Sivakumar, B., Chaos in hydrology: important issues and interpretations. *J. of Hydrology*, vol 227, 1-20, (2000)
20. Takens, F., Detecting strange attractors in Turbulence. *In Lecture Notes in Mathematics.*, **898**,366–81, Springer, Rand D. A., Young L. Editors (1980)