Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 8 (9): 1744-1749, 2008
ISSN 1812-5654

Flood Estimation at Ungauged Sites Using a New Hybrid Model

[1]Mahsa Hassanpour Kashani, [1]Majid Montaseri and [2]Mohammad Ali Lotfollahi Yaghin
[1]Department of Water Engineering, Faculty of Agriculture, University of Urmia, Urmia, Iran
[2]Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran

Abstract: As flood forecasting in ungauged basins has been an area of extensive research, new techniques have been introduced to minimize the forecast errors and to issue more accurate forecasts. The use of Artificial Neural Networks (ANNs) in flood forecasting is new and still in the evolution stage. In this study, MLP and Elman networks and also a new nonlinear regression model are applied and combined with each other for T-year flood estimation in western basins of Urmia Lake. At first, these networks used physiographic and climatic data selected from the regression model, to train. Finally, the best structure of the networks is chosen based on correlation coefficient between observed and estimated discharges. In order to train the models well, the return period variable is considered as one of the input variables of them. The obtained results have proved the ability of the hybrid model to predict T-year flood events and the effect of networks types on prediction precision.

Key words: MLP network, Elman network, new regression model, hybrid model, flood estimation, ungauged basins

INTRODUCTION

Floods are natural disasters that bring devastation and wide spread miseries to life. Though complete immunity from floods is not possible, yet the damage potential due to floods can be reduced through various structural and non-structural measures. The structural measures are capital intensive compared to non-structural measures, which are less expensive. Non-structural measures are basically used to provide flood forecasts and warning to people who reside in different flood zones (Kumar *et al*., 2001).

Floods are natural phenomena and inherently complex to model. The UK Flood Estimation Handbook (FEH) notes that many flood estimation problems arise at ungauged sites for which there are no flood peak data. In such cases, the hydrologist is faced with the difficult task of estimating flood event magnitudes from basin properties and/or regional climatology. The FEH recommends that, wherever possible, such estimates should be based on the transfer or analogous data from sites that are hydrologically similar in terms of basin area, rainfall and soil type (Dawson *et al*., 2006).

The quantity of runoff resulting from a given rainfall event depends on a number of factors. Different factors have been used to generate the stream flows. It is possible with standard statistical regression techniques to forecast flood based on basin descriptors-for example, derived from basin area, wetness and base flow index (Dawson *et al*., 2006). However, Robson and Reed (1999) state that flood estimates made from basin descriptors are, in general, grossly inferior, to those made from flood peak data (Robson and Reed, 1999).

The aims of the present investigation are thus twofold: (1) to explore the potential application of a new regression model, Artificial Neural Networks (ANNs) and a hybrid model solutions to the problem of flood estimation in ungauged basins; (2) to compare ANNs model prediction skill with that of the regression and the hybrid model.

ANNs have been used to perform hydrological modelling operations for over a decade. Since the advent of effective training algorithms for neural networks in the mid 1980s (Rumelhart and McClelland, 1986), neural solutions have been applied to a wide range of hydrological problems, such as rainfall-runoff modelling and river discharge (or stage) forecasting (for a review of forecasting applications (see Abrahart *et al*., 2004; Govindaraju, 2000). There have, however, been relatively few studies involving the application of ANNs to flood estimation at ungauged sites. For example, at the regional scale, Liong *et al*. (1994) investigated flood quantile prediction for ungauged basins in Quebec and Ontario (Liong *et al*., 1994). Muttiah *et al*. (1997), investigated 2-year peak storm discharge predictions for river basins in the United States (Muttiah *et al*., 1997). Hall and Minns (1998) related the scale and location parameters of the Extreme Value Type 1 (EV1 or Gumbel) distribution for

Corresponding Author: Mahsa Hassanpour Kashani, Department of Water Engineering, Faculty of Agriculture, University of Urmia, Urmia, Iran Tel/Fax: +989144008020

annual floods to six basin characteristics in two flood regions of the UK in subsequent experiments (Hall and Minns, 1998). Hall *et al.* (2000) used between four and twelve input basin characteristics to predict the same two EV1 parameter outputs using data from sites in Sumatra and Java (Hall *et al.*, 2000) whereas Dastorani and Wright (2001) found that seven basin inputs were sufficient to predict the index flood for selected basins in the UK (Dastorani and Wright, 2001). This study discusses the application of a hybrid model consisted of ANNs and a new regression model to estimate 5, 10, 15, 20, 25, 50, 100, 200, 500 and 1000 year flood event magnitudes using eleven input variables at such sites.

MATERIALS AND METHODS

Study region and data: The Urmia Lake basin is located in the North-west of Iran. This basin lies between 35° 40' and 38° 29' N and between 44° 13' and 47° 53' E and has a drainage area of 52700 km^2. The climate of the basin is continental semi-arid. The average annual precipitation and the average annual stream flow are about 398 mm and 4.5×10^9 m^3, respectively (Hassanpour, 2007). In this research, we applied MLP network, Elman network and a new nonlinear regression model to forecast T-year flood events in western region of Urmia Lake in 2006. The study region is consisted of six sub-basins. Physiographical and climatic data and information were obtained from West Ajarbayjan Water Authority. These data included basin drainage area, basin perimeter, Longest drainage path, basin mean slope, river mean slope, gravelius factor, form factor, time of concentration, 2-year rainfall and mean altitude of basin above sea level.

Nonlinear Regression Model (NRM): In the present study, a new regression model which does not have been applied within the field of regional flood modeling, is defined as follows:

$$Q_{T_r} = \alpha T_r^{\beta} \quad (1)$$

where, Q_{T_r} is T-year flood event magnitude (m^3/s), T_r is the return period (year), α and β are the parameters which are defined based on the physiographical and climatic data as follows:

$$\alpha = a_0 + a_1A + a_2P + a_3L_r + a_4S_r + a_5S_b + a_6F_g + a_7F_b + a_8T_c + a_9H + a_{10}R_2 \quad (2)$$

$$\beta = b_0 + b_1A + b_2P + b_3L_r + b_4S_r + b_5S_b + b_6F_g + b_7F_b + b_8T_c + b_9H + b_{10}R_2 \quad (3)$$

where, A is the basin drainage area (km^2), P is the basin perimeter (km), L_r is the Longest drainage path (km), S_b is the basin mean slope (%), S_r is the river mean slope (%), F_g is the gravelius factor, F_b is the form factor, T_c is the time of concentration (h_r), R_2 is the 2-year rainfall (mm), H is the Mean altitude of basin above sea level (m) and a_i and b_i are the model constant coefficients.

Multi-Layer Perceptron Network (MLP): Although there are now a significant number of network types and training algorithms, this research will focus on the Multi-Layer Perceptron (MLP) and Elman networks. Figure 1 and 2 provide an overview of the structure of these networks, respectively.

In this case, the ANN has three layers of neurons (nodes) -an input layer, a hidden layer and an output layer. Each neuron has a number of inputs (from outside the network or the previous layer) and a number of

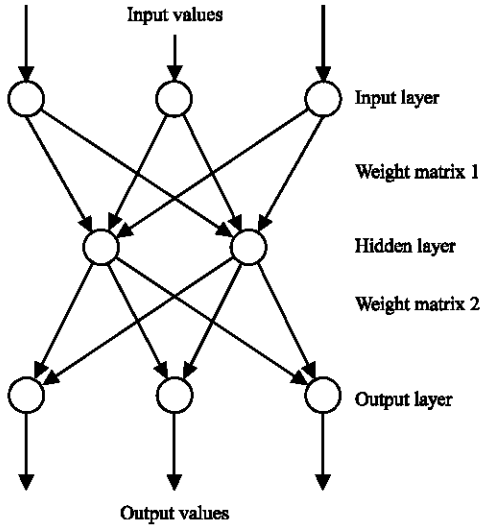

Fig. 1: Multi-Layer Perceptron (MLP) network

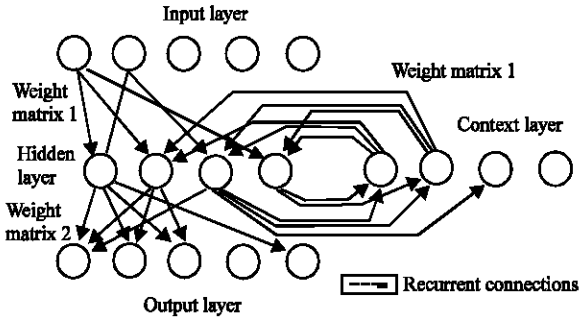

Fig. 2: Elman network

outputs (leading to the subsequent layer or out of the network). A neuron computes its output response based on the weighted sum of all its inputs according to an activation function (in this case the tangent sigmoid). Data flows in one direction through this kind of network-starting from external inputs into the first layer (the predictors), that are transmitted through the hidden layer and then passed to the output layer from which the external outputs (predictands) are obtained. The network is trained by adjusting the weights that connect the neurons using a procedure called error back propagation. In this procedure, the network is presented with a series of training examples (predictors and their associated predictands) and the internal weights are adjusted in an attempt to model the predictor/predictand relationship. This procedure must be repeated many times before the network begins to model the relationship (Dawson *et al*., 2006).

Elman network: Elman Networks are a form of recurrent Neural Networks which have connections from their hidden layer back to a special copy layer. This means that the function learnt by the network can be based on the current inputs plus a record of the previous state(s) and outputs of the network. In other words, the Elman net is a finite state machine that learns what state to remember (i.e., what is relevant). The special copy layer is treated as just another set of inputs and so standard back-propagation learning techniques can be used (something which isn't generally possible with recurrent networks). At each time step, a copy of the hidden layer units is made to a copy layer. Processing is done as follows:

- Copy inputs for time t to the input units
- Compute hidden unit activations using net input from input units and from copy layer
- Compute output unit activations as usual
- Copy new hidden unit activations to copy layer

RESULTS AND DISCUSSION

Nonlinear regression model: The parameters of the nonlinear model (α and β), are calculated using Excel software and the Linest function. In order to determine the best variables that have the most effect on the good performance of the multiple regression model, T-year flood magnitudes are estimated using each variable. The first variable is chosen based on the maximum correlation coefficient between observed and estimated discharges. The second effective variable is also chosen by testing the effect of both the rest variables and the selected variable on the model and etc. This will be continued until there is no difference between two successive maximum correlation coefficients or the last obtained correlation coefficient is enough high (Table 1). As can be seen from Table 1, the correlation coefficients related to No. 5 and 6, are almost the same and the one of No. 6 is excellent, so, there is no need to add the rest variables to the input variables and it is better to remove them. According to Table 1, the more the input variables, the better the obtained results. As can be seen from No. 4, the presence of S_b and S_r in the model is not reasoned because of their similar influences on T-year flood magnitudes. Also, this model shows good and almost similar results when it uses more than (or equal to) four input numbers. So, the data of F_b, A, S_r and H are recommended as the best variables for flood estimation in west basins. In the next two sections, we examine that MLP and Elman network whether confirm these 4 variables as the best ones or not.

Table 1: The best variables chosen by the nonlinear regression model

No.	Best variables	Correlation coefficient
1	F_b	0.712
2	F_b-A	0.838
3	F_b-A-S_r	0.886
4	F_b-A-S_r-H	0.936
5	F_b-A-S_r-H-S_b	0.949
6	F_b-A-S_r-H-S_b-L_r	0.953

MLP network: The selected data from the regression model are considered as the neural network's inputs. They were normalized in the range of [-1, 1] because of using the tangent sigmoid transfer function in the hidden layer, as follows:

$$P_n = \frac{2(P_i - P_{min})}{(P_{max} - P_{min})} - 1 \quad (4)$$

where, P_i is the inputs, P_{min} and P_{max} are minimum and maximum input respectively and P_n is the normalized inputs. In order to improve generalization, all data were divided into three sets: training set (75% of data), validation set (30% of training set) and test set (25% of data). Twelve training algorithms such as: Basic gradient descent (gd), Gradient descent with momentum (gdm), Adaptive learning rate (gda, gdx), Resilient back propagation (rp), Fletcher-Reeves conjugate gradient algorithm (cgf), Polak-Ribiére conjugate gradient algorithm (cgp), Powell-Beale conjugate gradient algorithm (cgb), Scaled conjugate gradient algorithm (scg), BFGS quasi-Newton method (bfg), One step secant method (oss) and Levenberg-Marquardt algorithm (lm) were applied to train the network. In this research, we left most of the training parameters of these algorithms at the default values, except the performance goal and maximum number of epochs to train that we modified their default values and

sat them equal to 0.008 and 120, respectively. After training, network was tested based on the test data and a linear regression was performed between the network outputs and the corresponding targets by putting the entire data set through the network (training, validation and test) to measure performance of the trained network. It returns to three parameters. The first two, m and b, correspond to the slope and the y-intercept of the best linear regression relating targets to network outputs. If we have a perfect fit (outputs exactly equal to targets), the slope would be 1 and the y-intercept would be 0. The third variable returned by regression analysis is the correlation coefficient (r-value) between the outputs and targets. It is a measure of how well the variation in the output is explained by the targets. If this number is equal to 1, then there is a perfect correlation between targets and outputs. Finally, the optimum number of the hidden neurons for each algorithm and input numbers was determined based on the maximum correlation coefficient value (Table 2). It can be seen that all algorithms don't show good results with one input variable (No. 1), in other words, the form factor (and return period) variable is not sufficient for precisely flood forecasting. Gd and gdm algorithms were recognized as weak algorithms, because of their bad results compared with other algorithms and the regression model. Rp and cgp have shown the best results compared with other algorithms of back propagation and conjugate gradient algorithms, respectively. bfg algorithm is the best algorithm with average correlation coefficient of 0.916. Maximum average correlation coefficient of all algorithms with 1 to 6 input numbers was equal to 0.933 related to the network with four input variables (No. 4). So, it can be concluded that MLP network is able to estimate flood using just four variables and there is no need for more inputs variables. In other words, MLP network is able to recognize the effective variables needed for flood forecasting. More input variables do not always cause to a good performance of MLP network.

Elman network: Elman network was trained as similar as MLP, but only gdx algorithm was applied to its training. Table 2, also shows the results of this network. It can be seen that, this network with six input variables (No.6) has the best performance because of the maximum correlation

Table 2: The results of MLP and Elman networks

		gd			gdm			gda			gdx			gdx[1]			rp		
Algorithms	No.	r	Neuron	Epochs	r	Neuron	Epochs	r	Neuron	Epochs	r	Neuron	Epochs	r	Neuron	Epochs	r	Neuron	Epochs
Back	1	0.691	13	120	0.606	5	48	0.711	13	101	0.725	17	120	0.712	12	118	0.698	16	69
propagation	2	0.664	11	120	0.687	9	18	0.841	8	110	0.856	17	120	0.891	9	85	0.788	11	80
	3	0.669	3	120	0.719	5	120	0.879	5	112	0.887	13	120	0.893	10	120	0.870	12	50
	4	0.554	1	120	0.847	9	48	0.888	5	120	0.940	9	120	0.930	10	120	0.888	8	79
	5	0.772	6	120	0.673	3	31	0.928	11	101	0.927	1	120	0.950	10	120	0.875	8	120
	6	0.657	9	120	0.658	6	120	0.910	4	105	0.787	3	120	0.903	7	120	0.906	3	20

		cgf			cgb			cgp			scg		
		r	Neuron	Epochs	r	Neuron	Epochs	r	Neuron	Epochs	r	Neuron	Epochs
Conjugate	1	0.716	13	17	0.717	5	37	0.714	17	21	0.713	17	28
gradient	2	0.935	9	63	0.913	4	43	0.937	14	40	0.907	10	48
	3	0.910	5	27	0.907	14	29	0.901	12	42	0.893	12	32
	4	0.901	9	41	0.942	9	52	0.922	9	48	0.939	9	49
	5	0.924	1	20	0.946	2	53	0.961	6	30	0.922	1	17
	6	0.929	4	46	0.939	4	43	0.957	6	69	0.924	1	32

		oss			bfg		
		r	Neuron	Epochs	r	Neuron	Epochs
Quasi-	1	0.716	16	32	0.731	5	68
Newton	2	0.924	13	41	0.931	13	52
	3	0.918	13	77	0.947	11	44
	4	0.936	10	58	0.945	12	31
	5	0.944	6	91	0.973	9	34
	6	0.942	8	50	0.972	7	40

		lm		
		r	Neuron	Epochs
Levenberg-	1	0.723	7	18
Marquardt	2	0.947	4	36
	3	0.937	8	12
	4	0.951	5	9
	5	0.937	4	40
	6	0.966	3	8

[1]gdx is relevant to Elman network

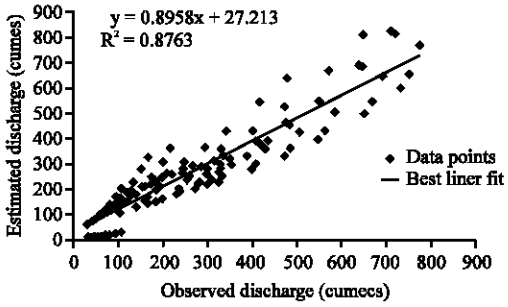

Fig. 3: Comparison of the observed and the estimated flow of the regression model (No. 4)

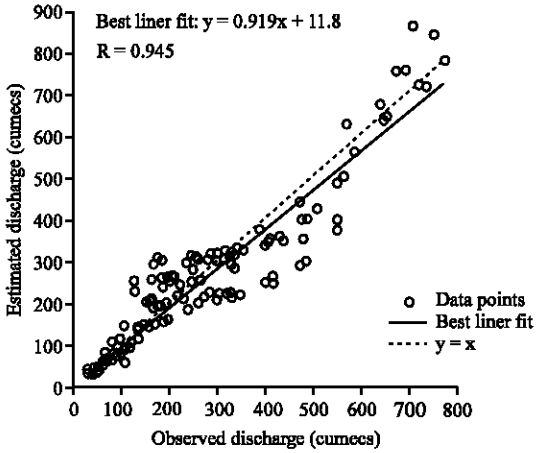

Fig. 4: Comparison of the observed and the estimated discharge of bfg algorithm (No. 4)

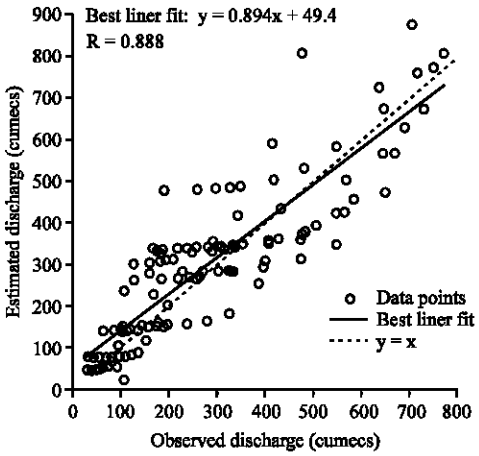

Fig. 5: Comparison of the observed and the estimated discharge of Elman network (No. 4)

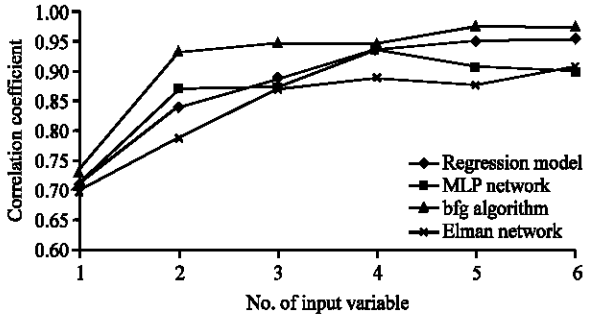

Fig. 6: Correlation coefficient of observed and estimated discharge with respect to the number of input variables for different models

coefficient value (0.906) and there is an almost remarkable difference between it and the one of No. 4 (0.888). But, its result in No. 5 that the S_b factor is added to the inputs, is worse than No. 4. So, Elman network could recognize the best variables for flood forecasting. According to the Table 2, it can be seen that Elman network has a low convergence speed. Also, this network needs more hidden neurons (10) than MLP (8). Also, gdx algorithm of MLP network shows better results than the one of Elman network.

Because of too many graphs obtained from these models, Fig. 3-5 compare the observed flow with the estimated flow of the nonlinear regression model, the best algorithm of MLP (bfg) and Elman network using only four variables (No. 4).

In order to determine the best variables for flood forecasting in west basin easily, a plot of the correlation coefficient of observed and estimated discharge with respect to the number of input variables for different models was provided in Fig. 6.

As it can be seen from this Fig. 6, it confirms the previous results of the models, but some new facts are as follows:

MLP network just using two input variables outperforms the regression model. In other words, this network just confirms the form factor and the drainage area variables as the input variables. Therefore, these variables are suggested as the at least data needed for flood forecasting. So, for flood forecasting in west basin at least two input variables (F_b-A) and at most four input variables (F_b-A-S_r-H) are suggested.

According to the plot of bfg algorithm, it can be seen that there is a remarkable improved performance with two inputs rather than one. Although, the best result is obtained using five inputs, but, it is not very different from the result of using four inputs. Also, this algorithm outperforms the nonlinear regression model.

Elman network shows a weak performance compared with MLP and the nonlinear regression model because: (a)

may be, it does not agree with the regression model about all of its selected variables (No. 1 to No. 6). In this case, it won't be a good network for solving hydrological problems or (b) may be it needs more hidden neurons than MLP, or (c) some of its training parameters (for example, learning rate, momentum constant) do not have the best values.

CONCLUSION

The performance of the new nonlinear regression model for flood forecasting was very satisfactory. As the number of input variables of the regression model increases, the values of correlation coefficients increase. In other words, the regression model needs more variables to estimate flood precisely and it is difficult to recognize the best variables using the regression model. According to the results of this model, F_b-A-S_r and H variables are suggested to estimate T-year flood magnitudes.

MLP network (except some weak algorithms) outperforms Elman and the regression model; also, bfg algorithm of this network is suggested to estimate flood in west basin. According to the results of this network, it can be seen that this network confirms the suggested variables of the regression model for flood forecasting and also all of its selected variables as the network input variables. In order to estimate flood precisely using MLP network and few data, it is better to apply all of the training algorithms. The regression model (in determining MLP's input variables) and MLP network (in determining the best variables for flood forecasting and estimating T-year floods) as a hybrid model showed satisfactory results. The return period variable as one of the input variables of all models was very effective on precisely flood estimation.

Elman network has a low convergence speed and needs more hidden neurons than MLP. Elman network did not outperform MLP and the regression model, so, it is not suggested to estimate T-year flood events in west basins of Urmia Lake.

REFERENCES

Abrahart, R.J., P.E. Kneale and L. See, 2004. Neural Networks for Hydrological Modelling. Taylor and Francis, London.

Dastorani, M.T. and N.G. Wright, 2001. Application of artificial neural networks for ungauged catchment flood prediction. Floodplain Management Association Conference, San Diego, CA, March.

Dawson, C.W., R.J. Abrahart, A.Y. Shamseldin and R.L. Wilby, 2006. Flood estimation at ungauged sites using artificial neural networks. J. Hydrol., 319 (1-4): 391-409.

Govindaraju, R.S., 2000. Artificial neural networks in hydrology II. Hydrological applications. J. Hydrol. Eng., 5 (2): 124-137.

Hall, M.J. and A.W. Minns, 1998. Regional flood frequency analysis using artificial neural networks. Proceedings 3rd International Conference on Hydroinformatics, Vol. 2. Balkema, Rotterdam, pp: 759-763.

Hall, M.J., A.W. Minns and A.K.M. Ashrafuzzaman, 2000. Regionalisation and data mining in a data-scarce environment. BHS 7th National Hydrology Symposium, Newcastle, UK., pp: 3.39-3.43.

Hassanpour, K.M., 2007. Simulation of rainfall-runoff process using artificial neural network model in Western and Southern basins of Urmia Lake. M.S. Thesis, Urmia University, Iran.

Kumar, S., R. Kumar, B. Chakravorty, C. Chatterjee and N.G. Pandey, 2001. An artificial neural network approach for flood forecasting. 16th National Convention of Computer Engineering, Patna.

Liong, S.Y., V.T.V. Nguyen, W.T. Chan and Y.S. Chia, 1994. Regional Estimation of Floods for Ungaged Catchments with Neural Networks. In: Developments in Hydraulic Engineering and their Impact on the Environment, Cheong, H.F., N.J. Shankar, E.S. Chan and W.J. Ng (Eds.). Proceedings 9th Congress of the Asian and Pacific Division of the International Association for Hydraulic Research, Singapore, pp: 372-378.

Muttiah, R.S., R. Srinivasan and P.M. Allen, 1997. Prediction of two-year peak stream discharges using neural networks. J. Am. Water Res. Assoc., 33 (3): 625-630.

Robson, A.J. and D.W. Reed, 1999. Flood Estimation Handbook Vol. 3: Statistical Procedures for Flood Frequency Estimation. Institute of Hydrology, Wallingford.

Rumelhart, D.E. and J.L. McClelland, 1986. Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 1, MIT Press, Cambridge.