Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management (Proceedings of Symposium HS2004 at IUGG2007, Perugia, July 2007). IAHS Publ. 313, 2007.

Predictive models of reservoir storage-yield-reliability functions: inter-comparison of regression and multi-layer perceptron artificial neural network paradigms

ADEBAYO ADELOYE
School of The Built Environment, Heriot-Watt University, Edinburgh EH14 4AS, UK
a.j.adeloye@hw.ac.uk

Abstract In this study, functions for predicting the total (within-year plus over-year) reservoir capacity have been developed using: first classical multiple regression, and secondly artificial neural networks (ANNs). The basis of the models is the storage-yield-reliability (S-Y-R) analysis of 18 international rivers using the sequent-peak algorithm (SPA). The results showed that the regression model performed better than the ANN model. The relative superiority of the regression model was attributed to its use of the over-year capacity as an independent variable. In contrast, the ANNs use basic variables as inputs and thus offer more flexibility than the regression model, particularly at ungauged sites.

Key words artificial neural networks; storage-yield-reliability; sequent peak algorithm; over-year capacity; within-year capacity; multiple regression

INTRODUCTION

The determination of storage-yield-reliability (S-Y-R) functions for reservoir planning purposes is better carried out using observed runoff data at the project site with a sequential method of analysis such as the sequent peak algorithm, SPA (McMahon & Adeloye, 2005). However, there are many situations in the world where the required data are either unavailable or insufficient for the purpose. This is particularly true of developing countries where investment in hydro-meteorological data collection activities has been drastically cut or is non-existent and where the few existing networks are in a very bad state of disrepair. In such situations, the need for relatively accurate predictive equations for the S-Y-R during reservoir planning becomes paramount.

The functional relationship between reservoir capacity, yield and reliability is non-linear and complex, which complicates any attempt to develop it theoretically. Approximate expressions have been developed when the reservoir is dominated by over-year behaviour (Vogel & McMahon, 1996); however, because of the large number of variables involved, no such analysis has been attempted when within-year dominates the behaviour of the reservoir. An over-year system is one which stores water for meeting demand over several years and thus takes a long time to empty or refill. Conversely, within-year systems are smaller reservoirs which are required to meet seasonal discrepancies between the inflow and demand; they therefore empty or refill at least once in any year. In practice, many reservoir systems and their associated rivers have within-year behaviour, which makes it important that an attempt is made to develop generalized S-Y-R functions for such systems.

In this study, generalized functions for predicting the total (within-year plus over-year) storage, given the yield and reliability, have been developed using first, classical multiple regression and secondly, artificial neural networks. The basis of the models is the S-Y-R analysis carried out using the sequent-peak algorithm (McMahon & Adeloye, 2005) and data for 18 international rivers. In the next section, further details about the methodology are given. This is followed by the presentation of the results. Finally, a summary of the study is given.

METHODOLOGY

Multiple linear regression analysis is simple and involves relating a single output variable to a number of input variables. The output variable selected for the regression was the total (within-year plus over-year) capacity. For a linear model to be plausible, the input variables must be those that have high linear correlation with the output variable. On the contrary, feed-forward, back-propagation artificial neural networks (ANNs) are able to predict multiple output variables simultaneously and in principle can map any function, no matter how complex, without the need to specify the functional form *a priori* (Parasuraman & Elshorbagy, 2007). ANNs are therefore unrestricted in the choice of input variables and were used to directly model the intrinsically nonlinear S-Y-R function. The ANN will be used to simultaneously predict the total and over-year capacities.

Data and S-Y-R analysis for deriving output variables

The study used monthly runoff data from 18 international catchments as summarized in Table 1. The first 15 rivers were used for model development and the remaining three for independently testing the models. The data records range from 20 years to 69 years with the catchments which vary in size from 101 km^2 to 19 654.4 km^2. The *Cv* of the annual flows in Table 1 varies from 0.180 to 1.47, which covers most of the regions of the world and its 15 monthly flow regimes as identified by Haines *et al*. (1988).

Output variables, i.e. the total and over-year capacities, were obtained through the S-Y-R analysis of the runoff data using the modified SPA (McMahon & Adeloye, 2005). The analysis used first the annual runoff data in order to determine the over-year capacity for six different failure situations. Monthly data were then used to calculate the total storage. For the analysis, ten annual yield ratios from 0.1 to 1 with a step of 0.1 were considered. For each river, the SPA was run for 0, 1, 2, 3, 4 and 5 failure year situations, implying an annual reliability range of 73% to 100%. This resulted in 60 values of over-year and total capacities (10 demand ratios multiplied by 6 failure scenarios) for each river, providing 900 values of each of the capacities for model development and 180 values for testing the models.

To identify the input variables, a correlation analysis was carried out. Possible input variables considered are those known to affect reservoir storage capacity, such as the coefficient of variation of annual flow *Cv*, the time-based reliability R_t, the yield ratio *D* and the length of record *L* (years) (Montaseri & Adeloye, 1999). Additionally,

Table 1 Details of catchments used in the study.

River	Gauging station	Country	*Cv*	*MAR* (10^6 m^3)	Record length (years)	Area (km^2)	max($Cv_{monthly}$)*	range(Mmr)*	min(Mmr)*
Sites for Model Development									
Oglio	Iseo	Italy	0.180	1758.665	20	1784.8	0.789	0.078	0.048
Earn	Kikell bridge	UK	0.189	648.285	34	590.5	0.648	0.103	0.034
Dee	Erbistok rectory	UK	0.201	1000.258	32	1040	0.782	0.108	0.035
Chiese	Idro	Italy	0.207	798.349	40	617.00	0.755	0.085	0.049
Sarca	Torbole	Italy	0.220	151.396	19	2200.00	0.901	0.055	0.061
Hatchie	Bolivar	USA	0.363	2178.401	55	3833.2	1.256	0.140	0.020
Homochitto	Eddiceton	USA	0.395	238.167	46	466.2	1.395	0.155	0.023
Paria	Lees Ferry	USA	0.404	26.760	61	3651.9	1.492	0.150	0.020
Richmond	Casino	Australia	0.653	699.459	61	1790	1.912	0.181	0.016
Ongaparinga	Clarendon weir	Australia	0.683	81.471	69	445	2.433	0.251	0.003
Werribee	Ballan	Australia	0.707	21.458	30	101	3.676	0.214	0.005
Renoster	Koppies dam	South Africa	0.991	112.362	40	2196	4.338	0.196	0.004
Vis	Harderug	South Africa	1.004	18.517	33	1463	3.646	0.237	0.008
Mareetsame	Neverest	South Africa	1.012	3.377	37	566	6.083	0.222	0
Brak	Bellair dam	South Africa	1.072	2.277	40	546	0.193	0.159	0.029
Sites for Independent Testing									
Prins	Prins River Dams	South Africa	1.47	3.85	42	761	4.437	0.172	0.018
Mulgrave	Gordonvale	Australia	0.51	886.34	35	554	1.997	0.246	0.015
Coruh	Karsikoy	Turkey	0.26	5922.41	38	19654.4	0.409	0.207	0.347

*See definitions in Table 2.

Table 2 Summary of the variables and their meanings (MAR is mean annual runoff)

Input variables	
Cv	Coefficient of variation of annual runoff
Cv^2	Square of the *Cv*
R_t	Time-based reliability (%)
D	Yield as ratio of MAR
max($Cv_{monthly}$)	Maximum Cv of monthly flow
range(Mmr)	Range of the mean monthly runoff (as a ratio of MAR)
min(Mmr)	Minimum mean monthly runoff (as a ratio of MAR)
L	Length of data record (years)
Output variables	
K_a	Over-year capacity as a ratio of MAR
K_t	Total capacity as a ratio of MAR

consideration was given to Cv^2, because of its strong influence on the over-year capacity (McMahon & Adeloye, 2005), and statistics of monthly flows which are thought to affect the within-year capacity, e.g. the minimum mean monthly runoff as a ratio of the mean annual runoff (min(Mmr)), the maximum monthly *Cv* (max($Cv_{monthly}$)) and the range of mean monthly runoff (range(Mmr)). The range is

given by max(Mmr) – min(Mmr), where max(Mmr) is the maximum mean monthly runoff expressed as a ratio of the mean annual runoff. A summary of the variables is shown in Table 2.

As expected, both the demand ratio and *Cv* of the annual runoff have significant correlation with the output variables. The very high correlation coefficient (r = 0.995) between the total storage capacity and the over-year capacity makes the latter a good predictor of the former for the regression model. Strangely, both the reliability and length of data record have low correlations with the output variables. Consequently, these two variables were not included as input variables for the regression models although they were included in the ANNs.

MODELS

Based on the above consideration, a regression model for the total storage capacity ratio was formulated as:

$$K_t = a + bCv + cD + dK_a + \xi \quad (1)$$

where a, b, c and d are model parameters, ξ is the error term and all the other variables are as defined in Table 2. The parameters were calibrated using all 900 output cases.

For the ANN modelling, the over-year and total storage capacities were considered as output variables to be predicted simultaneously, using five input variables, i.e.:

$$[K_a, K_t] = f\left(C_v^2, R_t, D, \max(Cv_{monthly}), \min(Mmr)\right) \quad (2)$$

where $f(.)$ denotes a functional form. The ANN also comprised one hidden layer. The determination of the number of neurons in the hidden layer used the early stopping rule (Sarle, 1995). To achieve this, the 900 exemplars were split into three sets for training (50%), validation (25%) and testing (25%) respectively. The number of neurons was initially set to a low value and was then progressively increased until the validation error reversed following a sustained sequence of decline. The validation error was measured in terms of the root mean square error, RMSE. On this basis, the optimal number of neurons in the hidden layer was found to be 21. The back-propagation training was achieved using the Levenberg-Marquardt algorithm, principally because of its fast convergence rate for moderately-sized networks (Reddy & Williamowski, 2000). All the ANNs were implemented using the ANN Toolbox of Windows Matlab (version 5, The Mathworks Inc., Natick, MA, USA) software.

RESULTS AND DISCUSSION

The calibrated coefficients of the regression model are shown in Table 3. The associated *t*-statistics (= parameter estimate divided by the standard error of estimate) are generally high, implying that the parameters are well estimated and are statistically different from zero. The model residuals were checked for normality and constant variance and they were found to be adequate. The particular form of the final model becomes:

$$K_t = -0.222 + 0.322Cv + 0.6D + 1.025K_a \quad (r^2 = 0.988) \quad (3)$$

Table 3 Summary statistics for the regression model.

Parameter	Estimate	Standard error	*t*-Statistic	Signif. at 5% (Y/N)*
a	–0.222	0.030	–7.370	Y
b	0.322	0.031	10.217	Y
c	0.600	0.043	13.791	Y
d	1.025	0.011	93.238	Y

*Null hypothesis that parameter is zero can be rejected at the 5% significance level (Y or No?).

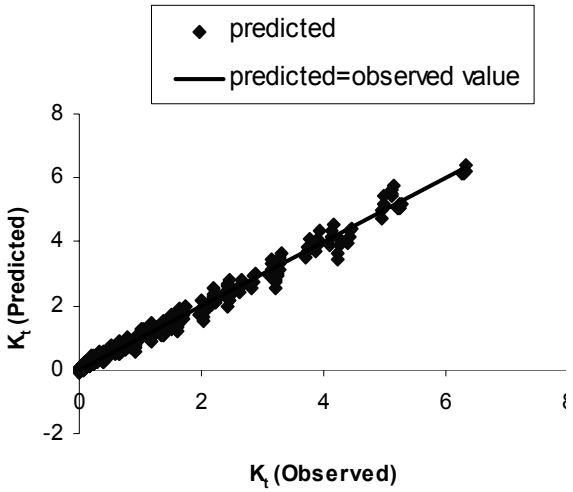

Fig. 1 Performance of regression model for predicting total (within-year plus over-year) capacity.

Figure 1 shows the observed and predicted total storage capacity during calibration and further confirms the excellent performance of the regression model.

The summary of the ANN model is shown in Table 4. In general, the model was able to predict both output variables well and the performance of the model was equally good during training and validation. As was the requirement of the early stopping approach, a quarter of the exemplars were reserved for testing the generalisation capability of the networks and the results of the testing are also shown in Table 4. The performance during testing is also good. Figure 2(a) and (b) shows the comparison between the observed and predicted output variables during testing and further confirm the excellent performance of the ANN model.

INDEPENDENT TESTING: COMPARING THE REGRESSION AND ANN MODELS

The testing of the ANN model carried out previously was part of its training, in that the S-Y-R exemplars used were drawn from the same pool as the exemplars used for model training and validation. As a consequence, that testing cannot be regarded as truly independent. Furthermore, although the regression model performed well during calibration, it has hitherto not been independently tested. It is therefore necessary to

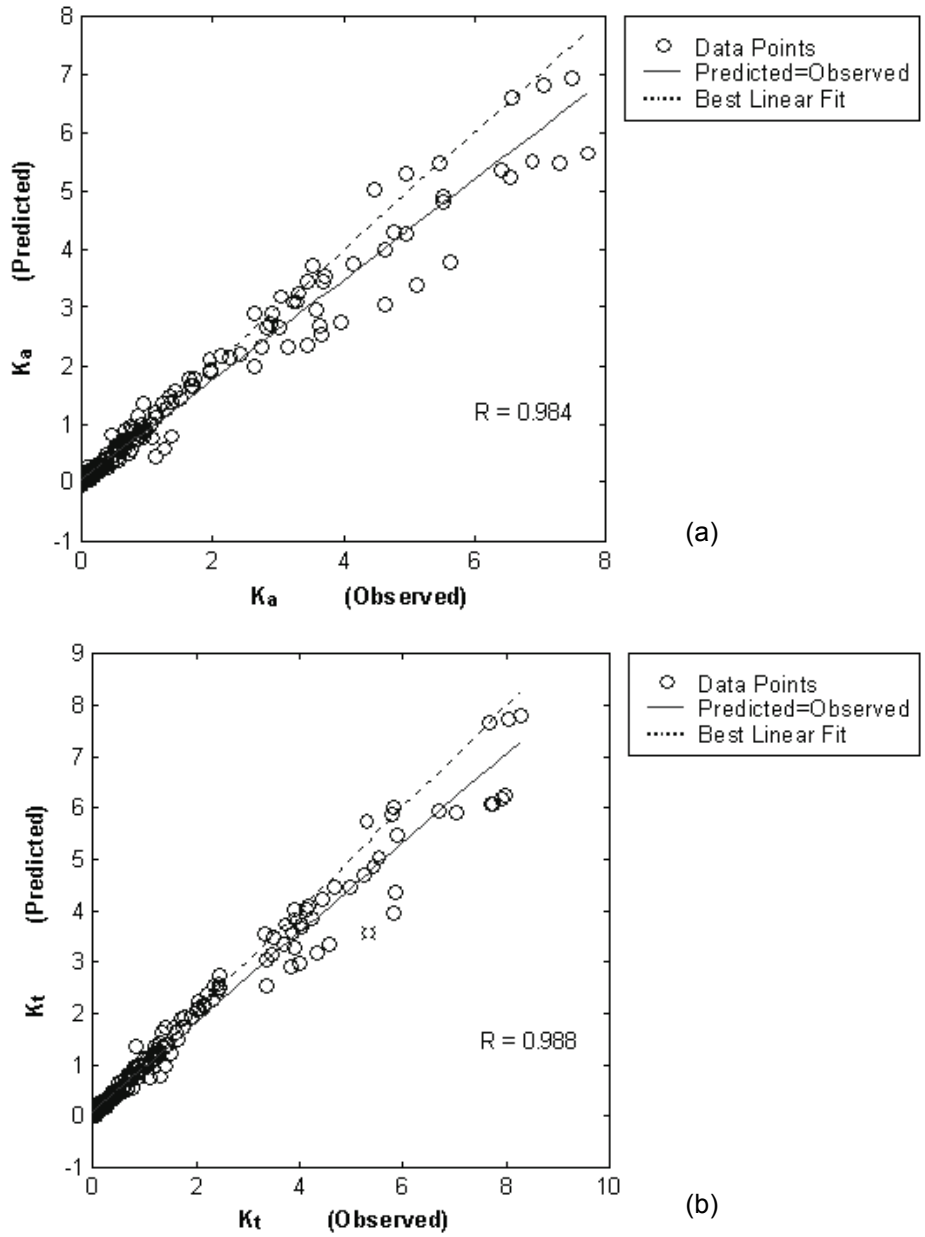

Fig. 2 Performance of the ANN model during testing for: (a) over-year capacity, K_a and (b) total capacity, K_t.

investigate how the two models will perform when applied to streamflow data not used at all as part of the training and calibration. Thus, the 180 exemplars derived by the S-Y-R analysis of the last three rivers in Table 1 were used to independently test both the regression and the ANN models.

For the total capacity K_t, a common output variable for both models, the performance of the regression model (r = 0.998) was far superior to that of the ANN (r = 0.8125) during independent testing. In comparison with the performance of the ANN presented in Table 4, the ANN is actually performing worse during independent testing than during the testing phase of the training. This is a problem with ANNs as

Table 4 Performances indexes for ANN model. Architecture: input-hidden-output: 5-21-2 nodes.

	MSE	*r*	Average MSE	Average *r*	RMSE
Training					
K_a/MAR	0.0039	0.9986	0.0044	0.9987	0.0663
K_t/MAR	0.0048	0.9987			
Validation					
K_a/MAR	0.1241	0.9843	0.1315	0.9858	0.3626
K_t/MAR	0.1389	0.9873			
Testing					
K_a/MAR	0.1514	0.9841	0.1503	0.9861	0.3877
K_t/MAR	0.1492	0.988			

observed by Flood & Kartam (1994), who noted that ANNs are typically very bad extrapolators. ANNs model surfaces using a set of data points by minimizing an error function; so when they are presented with data outside the training input data space, they are unable to relate the new data with the modelled error surface and may, therefore, not perform satisfactorily. It should be noted that the *Cv* of River Prins and min(Mmr) of River Coruh are outside the range used for the calibration.

Apart from the above, a further factor that might have caused the superior performance of the regression model during independent testing is that it used the over-year capacity as an input variable. The correlation between the total capacity and over-year capacity is much higher than with any other variable; thus using the K_a as an input variable is bound to result in a much better prediction of the total capacity than the use of any of the other input variables in Table 2.

The implication of the above is that the regression model should be used if the over-year capacity is known. This will certainly be the case for gauged catchments where the available annual runoff data can be analysed using the SPA to determine the over-year capacity. Although the total capacity can also be obtained by applying the sequent peak algorithm to the monthly or shorter-term runoff data at the site if available, the use of the regression model will be much quicker, thus representing a saving in analysis time. Further, given the very good performance of the regression model as demonstrated in this work, it is unlikely that much will be lost in terms of accuracy of the estimated total capacity.

However, for ungauged sites, the use of the regression model is foreclosed because the over-year capacity will be unknown. In such situations, the ANN model would be used. The inputs to the ANN model are basic streamflow parameters that can be readily estimated from catchment characteristics as described by Adeloye *et al*. (2003). In addition, the ANN model estimates simultaneously the total and over-year capacity with reasonable accuracy.

CONCLUSION

Generalised S-Y-R functions have been developed using multiple regression analysis and ANNs. The regression model was used to predict the total capacity, using the over-year capacity as an independent variable. A feed-forward back propagation ANN was trained to simultaneously predict the total and over-year capacity. The final network

has one input layer (with five neurons) and one hidden layer with 21 neurons. Results showed that both models performed well during development; however, the regression model was better when used on independent data sets. The reason for this was attributed to its use of the over-year capacity as an input variable. While this is an advantage, it does limit the application of the regression model to gauged sites. For ungauged sites, the ANN model is better suited because it uses inputs that can be readily obtained at ungauged sites using catchment characteristics or by pooling estimates from gauged sites within the region containing the ungauged site.

REFERENCES

Adeloye, A. J., Lallemand, F. & McMahon, T. A. (2003) Regression models for within-year capacity adjustment in reservoir planning. *Hydrol. Sci. J.* **48**(4), 539–552.

Flood, I. & Kartam, N. (1994) Neural networks in civil engineering. I: Principles and understanding. *J. Comp. Civ. Engng ASCE* **8**(2) 131–148.

Haines, A. T., Finlayson, B. L. & McMahon, T. A. (1988) A global classification of river regimes. *Appl. Geog.* **8**, 255–272.

McMahon, T. A. & Adeloye, A. J. (2005) *Water Resources Yield*. Water Resources Publications, Littleton, Colorado, USA.

Montaseri, M. & Adeloye, A. J. (1999) Critical period of reservoir systems for planning purposes. *J. Hydrol*. **224**, 115–136.

Parasuraman, K & Elshorbagy, A. (2007) Cluster-based hydrologic prediction using genetic algorithm-trained neural networks. *J. Hydrol. Engng ASCE* **12**(1), 52–62.

Reddy, J. M. & Williamowski, B. (2000) Adaptive neural networks in regulation of river flows. In: *Artificial Neural Networks in Hydrology* (ed by R. Govindaraju, & A. R. Rao), Chapter 8. Kluwer Academic, Dordrecht, The Netherlands.

Sarle, W. S. (1995) Stopped training and other remedies for overfitting. In: *Proc. 27th Symp. Interface Computing Science and Statistics*, 352–360.

Vogel, R. M. & McMahon, T. A. (1996) Approximate reliability and resilience indices for over-year reservoirs fed by AR(1) Gamma and normal flows. *Hydrol. Sci. J.* **41**(1), 75–96.