Wavelet Based regression models for prediction of Reservoir Inflow

B Krishna and Y.R.Satyaji Rao

Scientists, Deltaic Regional Center, National Institute of Hydrology, Siddartha Nagar, Kakinada-3, Andhra Pradesh.
E-mail: krishna_nih@rediffmail.com

Abstract *The accuracy of models used for any flood forecasting and warning system is critical since an accurate flood forecast with sufficient lead time can provide advanced warning. Data based forecasting methods are becoming increasingly popular in flood forecasting applications due to their rapid development times, minimum information requirements, and ease of real-time implementation. In data based flood forecasting, statistical models have traditionally been used, such as multiple linear regressions. In this paper, an attempt has been made to find an alternative method for prediction of inflow by combining the wavelet technique with linear and non linear regression models. The observed time series is decomposed into sub-series using discrete wavelet transform and then appropriate sub-series is used as an independent variable for the regression model. Several regression models have been developed to forecast the inflow into reservoir in one day advance. The calibration and validation performance of the developed models is evaluated with appropriate global statistics. The results were compared with the regression models with undecomposed data. The application of wavelet based regression models were found to be more effective as its prediction efficiency is more and its peak value is closer to observed value.*

Key words: *Inflow, Linear, Nonlinear Regression, Wavelet Decomposition*

1. Introduction:

Inflow is an important data for an optimal reservoir operation. The accuracy of models used for any flood forecasting and warning system is critical since an accurate flood forecast with sufficient lead time can provide advanced warning of an impending flood at an early enough stage such that flood damage can be reduced significantly. The importance of an accurate flow forecast, especially in flood-prone areas, has increased significantly over the last few years as extreme events have become more frequent and more severe due to climate change and anthropogenic factors. There are three main approaches that have been used flow modeling which include white-box model (physically based distributed models), gray box model (conceptual models) and the black-box model. The white and gray-box models aim to simulate the physical mechanisms underlying each component in the transformation of rainfall into flow, such as surface, subsurface and groundwater flow. A basin can also be represented by black-box models which associate basin inputs and desired outputs without detailed considerations on the physical processes. In this context, conventional statistical models such as linear-nonlinear regression types, multiple linear regression, autoregressive moving average and stochastic models were commonly used. These are popular in flood forecasting applications due to their rapid development times, minimum information requirements, and ease of real-time implementation.

Integrating the literature on existing forecasting techniques, [1] formulated Auto Regressive Moving Average (ARMA) models otherwise known as Box and Jenkins models. In the field of hydrology, Box and Jenkins models have been used for time series modeling of varied research interests some of which are [2] and [3]. But ARMA technique models only linear and stationary processes which would be a limitation in case of non-stationary and non-linear time series analysis. Consider a time series which is decomposed by some means to component time series that act as building blocks for time series. When these components were modeled independently to forecast the future components and finally reconstruct the forecasted components to arrive at the required future time series, the quality of predictions was better, although the method of decomposition was empirical in nature. Wavelets particularly Discrete Wavelet Transform (DWT) is one method that can decompose non-stationary and non-linear data into a set of simpler components, which could be modeled easily.

The wavelet transform is a mathematical tool that provides a time-scale representation of a signal in the time domain [4]. It is a useful tool for non-stationary processes such as hydrological time series. Wavelet transform, which is a pre-processing decomposed technique, showed successful performance in hydrological applications. Wavelets, due to their attractive properties, have been explored for use in time series analysis. Wavelet transforms provide useful decompositions of original time series, so that wavelet-transformed data improves the ability of a forecasting model by capturing useful information on various resolution levels. In the field of hydrology, wavelet analysis has recently been applied to examine the rainfall–runoff relationship in a Karstic watershed [5] and monthly reservoir inflow [6]. Several studies have been published that developed hybrid wavelet–regression models. [7] developed a coupled wavelet–autoregressive model for annual rainfall prediction. The wavelet transform is also integrated with multiple linear regression [8]. Each of these studies showed that different black box models trained or calibrated with decomposed data resulted in higher accuracy than the single models that were calibrated with an undecomposed and noisy time series.

The main purpose of the study presented is to examine the applicability and generalization capability of the wavelet based regression models for the prediction of the inflow values of Malaprabha reservoir.

2. Methods of analysis:

2.1 Wavelet Analysis:

Wavelet analysis is the breaking up of a signal into shifted and scaled versions of the original (or mother) wavelet. In wavelet analysis, the use of a fully scalable modulated window solves the signal-cutting problem. The window is shifted along the signal and for every position the spectrum is calculated. Then this process is repeated many times with a slightly shorter (or longer) window for every new cycle. In the end, the result will be a collection of time-frequency representations of the signal, all with different resolutions. Because of this collection of representations, it can be called as multiresolution analysis. By decomposing a time series into time–frequency space, one is able to determine both the dominant modes of variability and how those modes vary in time.

2.1.1 Discrete Wavelet Transform (DWT)

The basic aim of wavelet analysis is to determine the frequency (or scale) content of a signal and to assess and determine the temporal variation of this frequency content [9]. This property is in complete contrast to the Fourier analysis, which allows for the determination of the frequency content of a signal but fails to determine frequency time-dependence. However the short time Fourier transform maps a signal into a two-dimensional function of time and frequency, which can provide some information about both when and at what frequencies a signal event occurs. However, this method is only applicable to situations where the coherent time is independent of the frequency and the limited precision is determined by the size of the window [10]. Therefore, the wavelet transform is the tool of choice when signals are characterized by localized high frequency events or when signals are characterized by a large numbers of scale-variable processes. Because of its localization properties in both time and scale, the wavelet transform allows for tracking the time evolution processes at different scales in the signal.

The continuous wavelet transform of a time series f (t) is defined as

$$f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-b}{a}\right)dt \qquad (1)$$

Where $\psi(t)$ is the basic wavelet with effective length (t) that is usually much shorter than the target time series f (t). The variables are *a* and *b*, where *a* is the scale or dilation factor that determines the characteristic frequency so that its variation gives rise to a `spectrum'; and *b* is the translation in time so that its variation represents the `sliding' of the wavelet over *f(t)*. The wavelet spectrum is thus customarily displayed in time-frequency domain. For low scales i.e. when |a| << 1, the wavelet function is highly concentrated (shrunken compressed) with frequency contents mostly in the higher frequency bands. Inversely, when |a| >> 1, the wavelet is stretched and contains mostly low frequencies. For small scales, thus a more detailed view of the signal (known also as a “higher resolution”) whereas for larger scales a more general view of the signal structure can be expected.

The Discrete Wavelet Transform (DWT) is to calculate the wavelet coefficients on discrete dyadic scales and positions in time. Discrete wavelet functions have the form by choosing $a = a_0^m$ and $b = nb_0a_0^m$ in equation (1). where *m* and *n* are integers that control the wavelet dilation and shift respectively, and $a_0 > 1$, $b_0 > 0$ are fixed. The appropriate choices for a_0 and b_0 depend on the wavelet function. A common choice for them is a_0=2, b_0=1. In this study, the wavelet function is derived from the family of Daubechies wavelets with the 5 order.

The original signal X(n) passes through two complementary filters (low pass and high pass filters) and emerges as two signals as Approximations (A) and Details (D). The approximations are part of low pass filter, high-scale and low frequency components of the signal. The details are part of high pass filter, low-scale, and high frequency components. Normally, the low frequency content of the signal (approximation, A) is the most important part. It demonstrates the signal identity. The high-frequency component (detail, D) is nuance. The decomposition process can be iterated, with successive approximations being decomposed in turn, so that one signal is broken down into many lower resolution components (Figure 1). The selection of the optimal decomposition level is one of the keys to determine the performance of model in wavelet domain. Decomposition level is generally based on signal characteristics and experiences to selection. In this study, three decomposition levels were considered. For the present study, the wavelet function is derived from the family of Daubechies wavelets with the 5 order. Thus, input data were decomposed using Daubechies wavelet functions (mother wavelets) and sub-time series components (time series of 2-months mode (D1), 4-months mode (D2), 8-months mode (D3) and approximation mode (A3)) were obtained for the calibration and validation period.

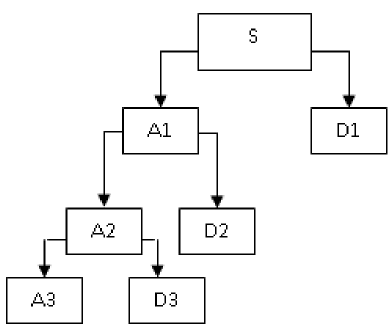

Figure 1. Wavelet Decomposition Tree

2.2 Regression Models:

ARMAX:

The Auto Regressive Moving Average with eXogenous inputs (ARMAX) model can be viewed as a simpler version of the ANN model with a linear threshold function as the transfer function and no hidden layer. In general the ARMAX model would be:

$$y(t) + a_1 y(t-1) + \ldots + a_{na} y(t-na) = b_1 u(t-nk) + \ldots + b_{nb} u(t-nk-nb+1) + e(t) + c_1 e(t-1) + \ldots + c_{nc} e(t-nc) \quad (2)$$

Where y is the output, u is the inputs considered in the study, and e is the error term. Where na, nb, nc and nk are the number of past outputs, inputs, error terms and time delays respectively contributing to the present output.

This model structure is represented by the notation ARMAX (na, nb, nc, nk). In this study, na and nb were each varies over the range from 0 to 4. For each of these combinations of na and nb, the optimal values for nc and nk and model parameters were estimated using MATLAB Identification Toolbox. The model which gives better performance statistical results was considered for the study.

Multiple Regression Models:

The decomposed details (D) and approximation (A) were taken as independent variables in regression models. The dependant variable will be the original value at one step ahead. The regression models can be described as LR1, LR2, NLR1 and NLR2. The LR1 model has constant, linear terms. The LR2 model has constant, linear and cross product terms. The NLR1 model has constant, linear, cross product and squared terms. The NLR2 model has constant, linear and squared terms. The regression coefficients were estimated by using the MATLAB statistical Toolbox. The performance statistics for calibration and validation period for these models are given in Table 1.

3. Study Area and Data Availability:

The river Malaprabha originated from Kanakumbi in the Western Ghat at an altitude of 792.48m and 16 km west of Jamboti in the Belgaum district of Karnataka State, India. The river flows in an easterly direction and joins the Krishna River. It is having a catchment area of 2,164 Sq. km and it lies between $15^0\ 30'$ to $15^0\ 50'$ N latitude and $74^0\ 15'$ to $75^0\ 0'$ E longitude along the border of Karnataka and Maharashtra states. The dam is constructed at a village Navilutheertha, Saundatti taluk, Belgaum district, Karnataka state, India having a latitude 15^0 49' North and longitude $75^0\ 6'$ East. The location of the reservoir is shown in Figure 2. The raingauges lies between the Khanapur stream-gauging station and location of dam (Navilutheertha) considered for calculating the average rainfall of the basin. The average annual rainfall for the study period was 687.18 mm. The average annual discharge at Khanapur gauge site is 8953 cumecs.

In the present study, the daily data of rainfall, stream flow at Khanapur gauging station and reservoir inflow for 11 years (from 1986 to 1996) were used. The model was calibrated using 7 years of data from

1986 to 1992 and validated by using the remaining 4 years of data from 1993 to 1996.

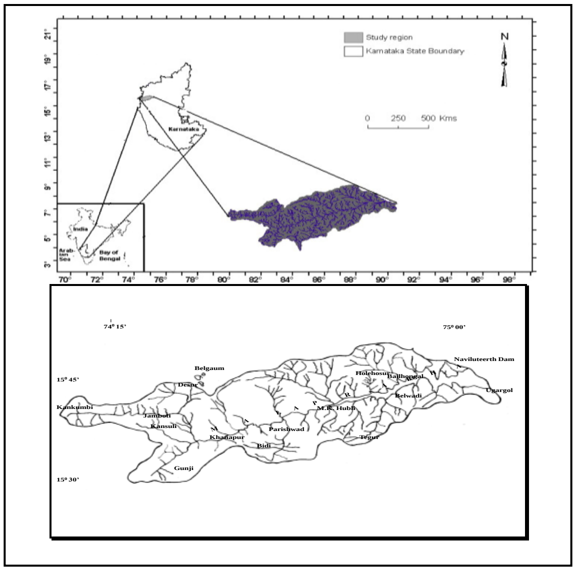

Figure 2. Location of study area

4. Input Selection:

The input vectors to models are selected based on the procedure described by [11]. The procedure uses cross-, auto-, and partial auto correlations between the variables in question along with their 95% confidence interval. By analyzing these correlogram plots, the significant lags of independent variables that are potentially influencing the output (dependant variable) can be identified. The following data sets identified as dependant variables to regression models were examined (i) daily inflow (at t_0 and t_0-1), daily rainfall (at t_0) and daily stream flow at Khanapur gauging station (at t_0) (ii) daily inflow (at t_0 and t_0-1), daily rainfall (at t_0) and daily stream flow at Khanapur gauging station (at t_0 and t_0-1) (iii) daily inflow (at t_0, t_0-1and t_0-2), daily rainfall (at t_0) and daily stream flow at Khanapur gauging station (at t_0 and t_0-1)

5. Model Evaluation:

To find out the optimal model developed in estimating reservoir inflow, different statistical indices are introduced. The indices employed are the root-mean-square error (RMSE) between the observed and forecasted values and the coefficient of efficiency (Nash-Sutcliffe) (CE).

$$\mathrm{RMSE} = \sqrt{\frac{\left[\sum_{t=1}^{N}(y_t^o - y_t^c)^2\right]}{N}} \qquad (3)$$

$$\mathrm{CE} = 1 - \frac{\left[\sum_{t=1}^{N}(y_t^o - y_t^c)^2\right]}{\left[\sum_{t=1}^{N}(y_t^o - \overline{y^o})^2\right]} \qquad (4)$$

where y_t^o and y_t^c are the observed and calculated values at time t respectively, and $\overline{y^o}$ and $\overline{y^c}$ are the mean of the observed and calculated values.

6. Results and discussion:

In this application, the first 7 year daily data (from 1986 to 1992) are used for training and the remaining 4 year (from 1993 to 1996) are used for testing. In this study, wavelet based regression models were employed to forecast inflow into Malaprabha reservoir with input combinations of antecedent inflow, rainfall and stream flow recorded at Khanapur gauging station. Further these were compared with standard regression models. For wavelet analysis, a decomposition level 3 and Daubechies wavelet of order 5 (db5) $\{D_1, D_2, D_3, A_3\}$. was considered for this study. Table 1 shows the performance of wavelet based regression models for datasets of dependant variables of (ii) i.e daily inflow (at t_0 and t_0-1), daily rainfall (at t_0) and daily stream flow at Khanapur gauging station (at t_0 and t_0-1) in calibration and validation periods. In addition, it also shows the performance of standard regression models.

Table 1. The root mean square errors (RMSE) and coefficient of efficiency (CE) statistics for the calibration and validation period

Method	**calibration**		**Validation**	
	RMSE	**CE**	**RMSE**	**CE**
ARMAX	25.26	87.93	41.10	76.70
LR1	22.19	90.69	31.81	86.08
LR2	19.84	92.56	**29.53**	**88.00**
NLR1	**19.59**	**92.74**	31.97	85.94
NLR2	21.36	91.38	32.42	85.54
WARMAX	18.78	93.33	25.11	91.33
WLR1	13.03	96.79	**16.73**	**96.15**
WLR2	7.24	99.01	27.69	89.45
WNLR1	**6.89**	**99.10**	37.04	81.13
WNLR2	11.72	97.40	18.84	95.12

From above table, it depicts that the performance of wavelet based regression models is much better than the performance of standard regression models in calibration and validation period. From the analysis it was found that efficiency index is more than 96% for Wavelet based LR1 model (WLR1) and for regression models (LR2) whereas it is 88 %. The wavelet based regression model (WLR1) with data set (ii) {i.e daily inflow (at t_0, and t_0-1), daily rainfall (at t_0) and daily stream flow at Khanapur gauging station (at t_0 and t_0-1)} have better performance (RMSE=16.73 and CE=96.15) than the regression model (LR2) (RMSE=29.53 and CE=88.00) during validation period. It may be due to the inputs used in wavelet based models are decomposed sub time series.

The error plot for models WLR1 and LR2 during validation period was shown in figure 3. It was observed that the WLR1 has low error than the LR2. The scatter plot for the models WLR1 and LR2 was shown in figure 4 and it was observed that the values predicted from the WLR1 model was more closer to 1:1 line than the LR2 model. From this analysis, it was worth to mention that the performance of wavelet based regression models was much better than the standard regression models in forecasting the reservoir inflow in one-day advance.

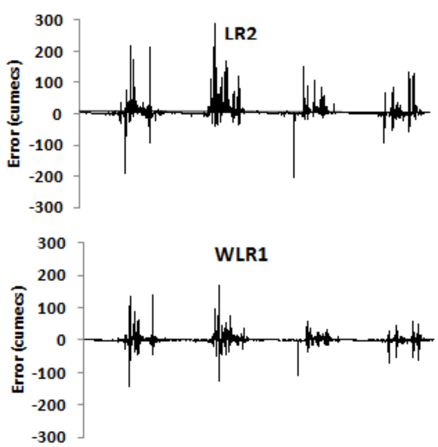

Figure 3. Distribution of error plots along the magnitude of flow during validation period

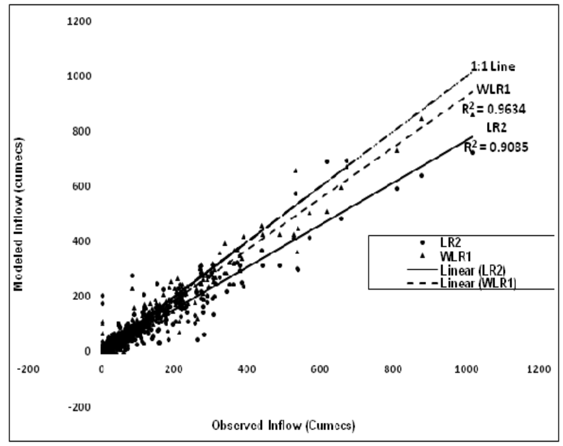

Figure 4. Scatter plots for reservoir inflow during validation period.

7. Conclusions:

The main purpose of the study presented is to examine the applicability and generalization capability of the wavelet based regression models for forecasting the inflow values of Malaprabha reservoir. Daily rainfall, antecedent inflow values and stream flow data at upstream gauging station used in this study. The observed time series are decomposed into sub-series using discrete wavelet transform and then appropriate sub-series is used as independent variables to the regression models in forecasting the reservoir inflow. Appropriate sub-series of the variable used as inputs to the ARMAX, linear and nonlinear regression models and original time series of the variable as output. The wavelet based regression models such as WARMAX, WLR1, WLR2, WNLR1 and WNLR2 was compared with the Standard regression models. Model parameters are calibrated using 7 years of data and rest of the data is used for model validation. From this analysis, it was found that efficiency index is more than 95% for Wavelet based regression models and 88 % for the standard regression models. This may be plausibly due to the variation in nonlinear dynamics of rainfall runoff process that mapped effectively by Wavelet based models. The study only used data from one area and further studies using more data from various areas may be required to strengthen these conclusions.

8. References:

[1] Box, G.E., Jenkins, G.M., Reinsel, G.C. "Time series analysis forecasting and control." San Francisco: Holden Day. 1970

[2] Han, P., Wang, P., Tian, M., Zhang, S., Liu, J., Zhu, D. "Application of the ARIMA Models in Drought Forecasting Using the Standardized

Precipitation Index" *Computer and Computing Technologies in Agriculture* VI. 352-358, 2013.

[3] Worrall, F., Swank, W.T., Burt, T.P. "Changes in stream nitrate concentrations due to land management practices, ecological succession, and climate: developing a systems approach to integrated catchment response" *Water Resources Research*. 39(7), 1177, 2003

[4] Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theory, 36(5): 961-1005, (1990).

[5] Labat, D., Ababou, R., Mangin, A. Wavelet analysis in Karstic hydrology 2nd Part: rainfall–runoff cross-wavelet analysis. Comptes Rendus de l'Academie des Sciences Series IIA Earth and Planetary Science 329, 881–887, 1999.

[6] Coulibaly, P., Anctil, F., Rasmussen, P. & Bobee, B. A recurrent neural networks approach using indices of low-frequency climatic variability to forecast regional annual runoff. Hydrological Processes 14, 2755-2777, (2000).

[7] Tantanee, S., Patamatammakul, S., Oki, T., Sriboonlue, V. & Prempree, T. Coupled wavelet-autoregressive model for annual rainfall prediction. Journal of Environmental hydrology 13(18), 1-8, (2005).

[8] Kisi, Ö. Neural networks and wavelet conjunction model for intermittent stream flow forecasting. Journal of Hydrologic Engineering, 14(8): 773-782, (2009).

[9] Heil, C.E. & Walnut, D.F. Continuous and discrete wavelet transforms. SIAM Review 31 (4), 628–666, (1989).

[10] Misiti, M., Misiti, Y., Oppenheim, G. & Poggi, J. Wavelet toolbox: For use with MATLAB, The MathWorks, Natick, Mass., 941, (2000).

[11] Sudheer KP, Gosain AK, Rangan DM, Saheb SM. Modeling evaporation using an artificial neural network algorithm. *Hydrological Processes* 16: 3189–3202, 2002