Available online at www.sciencedirect.com
ScienceDirect
Physics and Chemistry of the Earth 33 (2008) 141–150
PHYSICS and CHEMISTRY of the EARTH
www.elsevier.com/locate/pce

Monthly streamflow prediction in the Volta Basin of West Africa: A SISO NARMAX polynomial modelling

B.A. Amisigo [a,b,*], N. van de Giesen [c], C. Rogers [a], W.E.I. Andah [b], J. Friesen [c]

[a] *Centre for Development Research (ZEF), Bonn University, Germany*
[b] *CSIR-Water Research Institute (CSIR-WRI), Ghana*
[c] *Delft University of Technology, The Netherlands*

Received 14 April 2006
Available online 13 May 2007

Abstract

Single-input-single-output (SISO) non-linear system identification techniques were employed to model monthly catchment runoff at selected gauging sites in the Volta Basin of West Africa. NARMAX (Non-linear Autoregressive Moving Average with eXogenous Input) polynomial models were fitted to basin monthly rainfall and gauging station runoff data for each of the selected sites and used to predict monthly runoff at the sites. An error reduction ratio (ERR) algorithm was used to order regressors for various combinations of input, output and noise lags (various model structures) and the significant regressors for each model selected by applying an Akaike Information Criterion (AIC) to independent rainfall-runoff validation series.

Model parameters were estimated from the Matlab REGRESS function (an orthogonal least squares method). In each case, the sub-model without noise terms was fitted first followed by a fitting of the noise model. The coefficient of determination (*R*-squared), the Nash-Sutcliffe Efficiency criterion (NSE) and the *F* statistic for the estimation (training) series were used to evaluate the significance of fit of each model to this series while model selection from the range of models fitted for each gauging site was done by examining the NSEs and the AICs of the validation series.

Monthly runoff predictions from the selected models were very good, and the polynomial models appeared to have captured a good part of the rainfall-runoff non-linearity. The results indicate that the NARMAX modelling framework is suitable for monthly river runoff prediction in the Volta Basin. The several good models made available by the NARMAX modelling framework could be useful in the selection of model structures that also provide insights into the physical behaviour of the catchment rainfall-runoff system.

Keywords: NARMAX; NARX; Rainfall-runoff modelling; Non-linear models; Polynomial models; Dynamic models; Structure selection; Systems identification

1. Introduction

A proper basin scale water resources development and management scheme requires both existing and a good prediction of future stream flow series. Flood management and sizing of on-stream reservoirs require knowledge of both the magnitude and frequency of high flows while drought management, water withdrawals from streams and waste load carrying capacity of streams are determined from the magnitude and frequency of low flows (Tabrizi et al., 1998). An accurate simulation of river runoff series is, therefore, an important input to a comprehensive water resources management at the basin scale.

In the semi-arid Volta Basin of West Africa, existing monthly stream flow series for almost all the gauging stations are short and incomplete. In general these flow series are 20% incomplete with several stations having as high as 80% gaps (Taylor, 2004). Thus, stream flow simulation in this basin is essential not only as a means of filling in some of these gaps but also for the extension of these series in

* Corresponding author. Address: Centre for Development Research (ZEF), Bonn University, Germany.
E-mail address: barnyy2002@yahoo.co.uk (B.A. Amisigo).

1474-7065/$ - see front matter
doi:10.1016/j.pce.2007.04.019

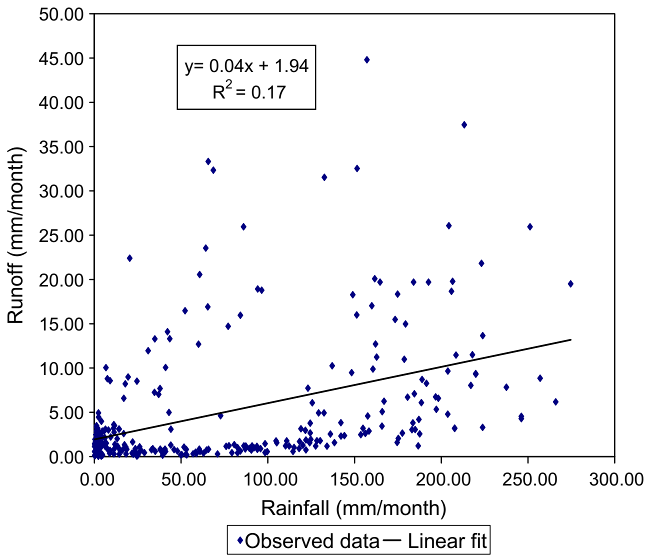

Fig. 1. Catchment runoff vs. rainfall for Bamboi on the Black Volta River.

order to provide adequate information for the water resources management of the basin.

The rainfall-runoff relationship in the Volta Basin has been recognised as highly non-linear (Andreini et al., 2000). This high non-linearity is clearly illustrated in Fig. 1 for one of the gauging stations used in this study. The figure is a plot of monthly catchment runoff vs rainfall and shows the high non-linearity in the relationship, at least at the monthly scale, with a very poor linear fit R-squared value of 0.17. Therefore, modelling the monthly rainfall-runoff system in this basin is a non-linear estimation problem.

2. The NARMAX polynomial model and system identification

A non-linear input–output model widely used in system engineering and found suitable for systems identification of a wide variety of non-linear systems including environmental systems is the NARMAX (Non-linear Autoregressive Moving Average with eXogenous Input) polynomial model (Billings and Leontaritis, 1982; Leontaritis and Billings, 1985b; Chen and Billings, 1989). This model is usually expressed as a non-linear polynomial function expansion of lagged input, output and noise terms, and, for the single-input-single-output (SISO) model, is represented as:

$$y(t) = f^d\begin{pmatrix} y(t-1), y(t-2), \ldots, y(t-n_y), u(t-n_k), \ldots, u(t-n_k-n_u), \\ e(t-1), \ldots, e(t-n_e) \end{pmatrix} + e(t) \tag{1}$$

where f^d is the polynomial of degree $d(d > 1)$; $y(t), u(t), e(t)$ are the output, input and white noise signals respectively at time t; n_y, n_u, n_e are the maximum output, input and noise lags, respectively; $n_{k(>0)}$ is the input signal time delay (measured in sampling intervals)

In non-linear systems identification in general, n_k is usually taken as at least 1. However, since in this study $u(t)$ is monthly rainfall, $n_k = 0$, i.e. $y(t)$ also a function of current input, would be considered. Thus for $n_y = n_u = n_e = 1, d = 2$ and $n_k = 0$, for example, the polynomial expansion in (1) for $y(t)$ is:

$$y(t) = \begin{pmatrix} \theta_1 + \theta_2 y(t-1) + \theta_3 u(t) + \theta_4 u(t-1) + \\ \theta_5 y(t-1)^2 + \theta_6 u(t)^2 + \theta_7 u(t)u(t-1) + \theta_8 u(t-1)^2 + \\ \theta_9 y(t-1)u(t) + \theta_{10} y(t-1)u(t-1) + \\ \theta_{11} y(t-1)e(t-1) + \theta_{12} u(t)e(t-1) + \theta_{13} u(t-1)e(t-1) + \\ \theta_{14} e(t-1) + \theta_{15} e(t-1)^2 \end{pmatrix} + e(t) \tag{2}$$

where $\theta_1, \theta_2, \ldots, \theta_{15}$ are parameters of the model.

A non-linear polynomial model of the form of (1), but without the noise terms, i.e,

$$y(t) = f^d[y(t-1), y(t-2), \ldots, y(t-n_y), u(t-n_k), \ldots, u(t-n_k-n_u)] + e(t) \tag{3}$$

is the NARX (Non-linear Autoregressive with eXogenous Input) model. This model is also general and can describe any non-linear system well (Stenman, 2002). In addition, it is not recursive as the regressors are independent of previous model outputs whereas in the NARMAX representation the noise terms can only be derived from previous model outputs. It is therefore more convenient to work with. However, the absence of a noise model in the structure means that a large number of regressor terms have to be included in order for the model to adequately represent both the system and noise dynamics (Stenman, 2002). Due to this limitation and also to avoid bias in the estimated parameters (Chiras et al., 2000), the full NARMAX model is preferred. However, to simplify the model selection process, the NARX part of the NARMAX model is fitted first in this study and then the noise terms fitted afterwards to obtain the full model.

2.1. Formulation of the model

For a given input–output series and any set of n_y, n_u, n_e, n_k and d, the polynomial represented in (1) above can generally be expressed as:

$$y(t) = \sum_{m=1}^{np} P_m \theta_m + e(t) \tag{4a}$$

where np is the number of terms in the polynomial expansion, P_m is the mth regressor term with $P_1 = 1$, and θ_m is the regression parameter for term m. The regressor terms are formed, as in (2), by various combinations of lagged values of the output and noise terms and both lagged and current (when $u_k = 0$) values of the input term.

In matrix form (4a) becomes:

$$y = P\theta + \varepsilon \tag{4b}$$

Here P is n by np and y n by 1 regressor and output matrices, respectively, Θ is np by 1 and ε is n by 1 parameter and white noise vectors, respectively, with n being the number of samples in the input–output series.