A joint stochastic-deterministic model (cross-correlation transfer functions, artificial neural networks, polynomial regression) for monthly flow predictions

Milan Stojković*, Srdan Kostić*, Jasna Plavšić**, Stevan Prohaska*

* Institute for the development of water resources “Jaroslav Černi”, Beograd, Serbia
** Faculty of Civil Engineering, University of Belgrade, Belgrade, Serbia
milan.stojkovic@jcerni.co.rs, srdan.kostic@jcerni.co.rs, jplavsic@grf.bg.ac.rs
stevan.prohaska@jcerni.co.rs

Abstract—A procedure for modelling of mean monthly flow time series using records of the Great Morava River (Serbia) is presented in this paper. The main assumption of the conducted research is that a time series of monthly flow rates represents a stochastic process comprised of deterministic, stochastic and random components. The former component can be further decomposed into a composite trend and two periodic components. The deterministic component of a monthly flow time-series is assessed by spectral analysis, whereas its stochastic component is modelled using cross-correlation transfer functions, artificial neural networks and polynomial regression. The results suggest that the deterministic component can be expressed solely as a function of time, whereas the stochastic component changes as a nonlinear function of climatic factors (rainfall and temperature). The results infer a lower value of Kling-Gupta Efficiency in the case of transfer functions, whereas artificial neural networks and polynomial regression suggest a significantly better match between the observed and simulated values. It seems that transfer functions fail to capture high monthly flow rates, whereas the model based on polynomial regression reproduces high monthly flows much better because it is able to successfully capture a highly nonlinear relationship between the inputs and the output.

I. INTRODUCTION

There is a number hydrological models for flow estimation based on long-term flow statistics [6] or which use the conventional mathematical approach based on dependence among inputs and the output at a few successive lags [1]. The former approach preserves the characteristics of the hydrologic process in the long-run, while the latter models can reproduce the short-term hydrologic pattern. The motivation of thus research is to introduce an approach that preserves both long-term and short-term characteristics of hydrologic time-series on two time scales (annual and monthly), using a disaggregated time series.

The underlying assumption is that time series can be decomposed into deterministic, stochastic and random parts. The deterministic part consists of the composite trend, macro-periodic component and seasonal component. These components describe low-frequency and seasonal variability, while high-frequency variability is represented by the stochastic component. In the present research, the stochastic component is modelled by cross-correlation transfer function (TFs) that describe linear relations among the flows and climatic data. For the same purpose, artificial neural network (ANN) and polynomial regression (PLR) are used as techniques capable of modelling the complex relations between monthly flows and climatic inputs. In such cases, ANN and PLR serve as tools for examining the accuracy of the approach based on transfer functions, with the ultimate goal of deriving the most reliable model for the stochastic component of monthly flow time-series.

II. METHODOLOGY

The present research relies on the basic assumption that monthly flow time series \( Q_t \) represent the sum of three components [7]:

\[
Q_t = Det_t + Stoch_t + error_t \rightarrow \\
Q_t = [Q_{tr} + Q_{ts} + Q_{et}] + [Y_t] + [e_t], \quad t = 1, 2, ..., N (1)
\]

where \( Q_{tr} \) is the composite trend, \( Q_{ts} \) is the long-term periodic component, \( Q_{et} \) is the seasonal component, \( Y_t \) is the stochastic component and \( N \) denotes the sample size. Addend \( e_t \) is the error term (random time series).

We model the deterministic component separately from the stochastic component. The composite trend and macro-periodic component constitute the annual deterministic component, which is modelled on an annual time-scale, whereas the seasonal component is modelled on a monthly time scale. After the annual deterministic component is modelled, it is downscaled to monthly intervals, so as to subtract it from the monthly flow time-series. The first and second order residuals are derived in this way. The latter contains a significant seasonal component, which is modelled further using the monthly flow time-series.

The composite trend is modelled using the Linear Moving Window (LMW) method [4]. Once the composite linear trend \( Q_{tr} \) has been determined, it is downscaled from the annual to a monthly time scale. The first-order
residuals that represent the long-term periodicity are identified by assessing the time series in the frequency domain using spectral analysis (i.e. Fourier transformation). Consequently, the time series is smoothed by the locally-weighted scatterplot smoothing method (LOESS), prior to identifying the long-term periodic component. Once the long-term periodic component is obtained, second-order residuals Q*, which contains seasonal periodicity, are computed. This component is also modelled in the frequency domain assuming that the annual cycle does not vary over time and that the seasonal periodic component is repeated each year.

Furthermore, the stochastic component was modelled by TFS, which are generally used for modelling of dynamic systems to transform a given input into the needed output. If it is assumed that a pair of observed input time series Xt and output time series Yt is available for the same time interval, the output Yt of a dynamic system can be expressed using a linear filter [2]. Once the necessary parameters are assessed, the output series yt, driven by input time series xt, (rainfall) and zt (temperatures), can be obtained as follows [7]:

\[ y_t = \frac{\alpha_0(B)}{\delta_1(B)} x_t + \frac{\omega_0(B)}{\delta_2(B)} x_{2t} + \frac{\theta(B)}{\phi(B)} a_t. \]  

The transfer function parameters are given by \( \alpha(B) = \alpha_0 B^r \cdots \alpha_B \), and \( \omega(B) = \omega_0 B^s \cdots \omega_B \), where \( r \) denotes the number of significant terms in the TF between the input and output variables, \( s \) is analogous to the order \( p \) of the AR(p) model and \( t \) is the random time series.

In order to develop a reliable model of the stochastic component, the authors use a three-layer feed-forward neural network with the Levenberg-Marquardt learning algorithm, which is commonly considered to be the fastest method for training moderate-sized feed-forward neural networks [3], and it is the first choice for solving problems of supervised learning, which is the case in the present analysis.

In the last step, the authors apply the polynomial regression approach and the least squares method of approximation [5], to formulate a prediction model for the stochastic component, as a function of the same input parameters as the case of TF and ANN model. The model has the following general form:

\[ Y = f + \sum_{i=1}^{s} \beta_i Z_i + \sum_{i=1}^{t} \sum_{j=1}^{r} \beta_{ij} Z_i Z_j + \sum_{i=1}^{s} \sum_{j=1}^{r} \sum_{k=1}^{u} \beta_{ijk} Z_i Z_j Z_k \]  

where \( Y \) is the response variable (i.e. stochastic component of monthly flow rates), which can be defined as a linear or nonlinear function of the examined input factors \( X_t, X_2 \) and \( Y \) (i.e. air temperature, rainfall and previous monthly flow rates). \( Z_t \) is any of the examined input factors. Coefficients \( \beta_0, \beta_1, \beta_0 \) and \( \beta_{ij} \) correspond to the coefficients of the intercept, linear, quadratic and cubic terms (or interactions), respectively.

III. RESULTS

Time series of the annual flow rates are used to model the trend and the long-term periodical component at h.s. Ljubčevski Most. The analysed time series cover the period from 1931 to 2012. The LMW method is employed for determining the linear trend of annual flows. The obtained results imply that annual trend consists of a periodic component with low amplitude and large oscillation period (Figure 1a). In the next step, the determined composite trend QT at annual time scale is downscaled from the annual to monthly time scale, in order to define first order residuals, which is needed for modelling of monthly seasonal component. The first-order residuals are then smoothed by the local regression LOESS method.

Having assessed the parameters of smoothed residuals, the macropериодical component is modelled by the means of the Fourier transform (Figure 1a). The obtained results indicate that the statistically most significant period is 20.5 years, followed by the periods of 16.4, 10.3, 27.3 and 13.7 years. As it could be seen in Fig. 1a, fluctuation of macropериодical component is observed for the whole time series. The largest value of annual flows is observed in period 1960-1980, while the lowest value is recorded during the last decade of the 20th century.

The second-order residuals Q* are determined using the downscaled composite trend and the downscaled long-term periodic component at monthly time scale. These residuals demonstrate seasonal cycles with significant periods of 12 months. One should note that the seasonal period of 6 months is also significant regardless of its lower share in the explained variance when compared to the annual cycle. As in the case of the long-term periodic component, the seasonal periodic component is modelled by applying the Fourier transform. The seasonal component is shown in Figure 1(b) along with the observed monthly flows for 1931-2012.

Once the trend component, the long-term periodic component and the seasonal periodic component are removed from the monthly time series, the resulting third-order residuals are used to model the stochastic component Yi. The functional dependence between the input time series (including monthly time series of rainfall X1t and air temperature X2t over the river basin) and the output time series (the stochastic component Yi) is further assessed using three different models (TF, ANN and PLR). One should note the authors assume that monthly temperatures and rainfalls predetermine the stochastic component, since they were the only measured and reliable numeric information within the investigated river basin.

By summing all the modelled components (deterministic and stochastic) for the calibration period (1950-1990), the authors obtain the resulting monthly flow rates for h.s. Ljubčevski Most, which are then compared to the observed ones, as it is shown in Figures 2.

As it could be seen in Figures 2 there is a good match between the recorded and the modelled monthly flow rates. Values of the Kling-Gupta Efficiency (KGE) of the modelled monthly flow rates for TF, ANN and PLR are 0.736, 0.841 and 0.891, respectively. Hence, best matching is achieved for the case of PLR, which also
shows better agreement in domain of higher monthly flow rates, whereas TF fails to explain high flows.

Predictive power of derived models is tested by using the records in validation phase which have not been utilised during the calibration, for the period 1991-2012. The modelled flow rates compared with observations are given in Figure 3.

IV. CONCLUSION

The results indicate that the models’ efficiency during the calibration phase exhibited satisfactory matching with observations. It was shown that the best agreement with observed monthly flows was achieved with PLR and that the least reliable model was TF. It can be suggested that the main advantage of the proposed approach, when compared to other available methods, is relatively quick

Figure 1. Observed monthly flow rates together with composite trend, macroperiodical component (a) and seasonal component (c) of the Great Morava River at h.s. Ljubičevski Most, 1931-2012.

Figure 2. Observed and modelled monthly flow rates of the Great Morava River at h.s. Ljubičevski Most for the calibration period (1950-1990): (a) TF, (b) ANN, (c) PLR.

Figure 3. Predictive power of proposed models using (a) TF, (b) ANN and (c) PLR for the validation phase (1991-2012).
assessment of model parameters based on hydrologic and meteorological records in the river basin. Therefore, the proposed method does not require more sophisticated inputs, which are needed for physically-based hydrologic models. Using the proposed approach, it is possible to model low-frequency and seasonal frequency variability from hydrologic records for different climates, owing to the fact that the suggested method is able to deal with a broad range of frequencies, amplitudes and phase shifts. In the case of arid regions with frequent discontinuities in flow time-series, intra-annual changes should be modelled by incorporating the intermittent part into the seasonal component.

The developed model of monthly flow rates from Eq. (1) can be used for seasonal forecasting. The model of the stochastic component with the best performance (PLR) can be applied for monthly prediction of this component based on seasonal rainfall and air temperature forecasts. The deterministic component of monthly flow rates, which characterises multi-annual and annual changes, can be extrapolated for several time steps ahead, since it can be expressed solely as a function of time. Also, the uncertainties of flow prediction can be reduced by analysing modelling errors, especially in the domain of high flows. That way, the distribution of error terms can be assessed by providing a confidence interval of the seasonal forecast.

Acknowledgment

This study was supported by the Ministry of Education, Science and Technological development of the Republic of Serbia as a part of the project TR 37013 "System development to support optimal sustainability of the high dams in Serbia, 2010-2017".

References