Water Resources Management (2020) 34:3689–3708
https://doi.org/10.1007/s11269-020-02644-y

A Comparative Study of Linear Stochastic with Nonlinear Daily River Discharge Forecast Models

Hossein Bonakdari[1] · Andrew D. Binns[2] · Bahram Gharabaghi[2]

Received: 8 January 2020 / Accepted: 29 July 2020 /
Published online: 8 August 2020

Abstract

Accurate forecast of the magnitude and timing of the flood peak river discharge and the extent of inundated areas during major storm events are a vital component of early warning systems around the world that are responsible for saving countless lives every year. This study assesses the forecast accuracy of two different linear and non-linear approaches to predict the daily river discharge. A new linear stochastic method is produced by evaluating a detailed comparison between three pre-processing approaches, differencing, standardization, spectral analysis, and trend removal. Daily river discharge values of the Bow River with strong seasonal and non-seasonal correlations located in Alberta, Canada were utilized in this study. The stochastic term for this daily flow time series is calculated with an auto-regressive integrated moving average. We found that seasonal differencing is the best stationarization method for periodic effect elimination. Moreover, the proposed non-linear Group Method of Data Handling (GMDH) model could overcome the known accuracy limitations of the classical GMDH models that use only two inputs in each neuron from the adjacent layer. The proposed new non-linear GMDH-based method (named GS-GMDH) can improve the structure of the classical linear GMDH. The GS-GMDH model produced the most accurate forecasts in the Bow River case study with statistical indices such as the coefficient of determination and Nash-Sutcliffe for the daily discharge time series higher than 97% and relative error less than 6%. Finally, an explicit equation for estimation of the daily discharge of the Bow River is developed using the proposed GS-GMDH model to showcase the practical application of the new method in flood forecasting and management.

Keywords Discharge forecast · Water resources management · Pre-processing · Stochastic modelling · Time series

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11269-020-02644-y) contains supplementary material, which is available to authorized users.

✉ Hossein Bonakdari
hossein.bonakdari@fsaa.ulaval.ca

Extended author information available on the last page of the article

1 Introduction

One of the most important ways to reduce the damages of floods in large cities is to design a robust, reliable, and universal method for flood forecasting and early warning detections (Walton et al. 2019; Gholami et al. 2019). Due to climate change, water-related studies have attained increased importance by many researchers. Climate change forecasts have suggested that impacts on water resources will have devastating consequences on human and ecological health. In this manner, the planning and management of water resources will be the most critical issue faced by humankind (Gharabaghi and Sattar 2019). Climate change can increases the potential for extreme rainfall as well as the risk of flooding. Indeed, river flow forecasting is vital for the management and planning of river basins, including water allocation for agriculture, generation of hydroelectric energy, navigation planning, risk appraisal, droughts, and flood control (Khatibi et al. 2012). Floods are presently the dominant natural disaster and tend to have the most significant associated economic cost (Serinaldi et al. 2018). One-third of all natural disasters in Canada between 1950 and 2012 were the result of floods (Kelly and Stodolak 2013).

1.1 Data-Driven Models

Expanding upon data-driven models has an extended application in flood modelling, which has gained a heightened reputation in recent years. Amongst them, one of the most popular methods is the group method of data handling (GMDH) (Najafzadeh et al. 2015). The GMDH is a self-organized approach that is capable of introducing different explicit equations for practical applications. However, due to intricate non-linear patterns in time series, their use needs professional programming knowledge. Also, the critical question with these methods is the selection of the best input parameters for predicting the results with high accuracy (Ebtehaj et al. 2020).

1.2 Problem Statement

The use of stochastic-based models has generally been rejected due to their linear nature in the modelling of hydrological processes, and in most studies, inefficiencies have been reported (Mosavi et al. 2018). Recently, Bonakdari et al. (2019) indicated that considering an appropriate linear methodology could result in improved modelling compared to a non-linear approach (such as ANFIS and neural network) in terms of accuracy and simplicity. Therefore, the following fundamental questions arise: 1) Can the identification of deterministic and stochastic terms of the time series also provide a useful solution in modelling river discharge with strong seasonal and non-seasonal correlations with a stochastic model?; and 2) What is the performance of this linear methodology in comparison with the non-linear model? This study seeks to address these questions for the case of a highly complex daily time series data set with strong seasonal and non-seasonal correlations.

1.3 Scope of the Current Study

This study provides a comparison of a novel stochastic based linear methodology with a new encoding of the GMDH known as the generalized structure of the GMDH (GS-GMDH) for the daily discharge forecasting in the Bow River in Alberta, Canada. Both proposed linear and

non-linear approaches are encoded in MATLAB. The linear methodology assessed the existence of the deterministic and stochastic terms and suggested a method to remove the deterministic terms. In the non-linear GS-GMDH model, to overcome the limitation of the classical GMDH method (which uses only a second-order polynomial and lacks the use of nonadjacent inputs in each neuron), a second and third order polynomial and inputs from adjacent layers in each neuron. Finally, the performance of the best linear-based stochastic model is compared with the best GS-GMDH based model as a non-linear approach using multi-criteria statistical indices.

1.4 Hydrological Data Collection

The highest natural disaster in Canadian history in terms of economic losses was the June 2013 flood event that affected the city of Calgary, Alberta, where five lives were lost. As much as $6 billion CAD in economic losses were sustained (Pomeroy et al. 2016). This flood event resulted in one of the top causes of the domestic insurance misfortunes in Canada (Insurance Bureau of Canada 2017).

Moreover, the cost of infrastructure damages, recovery costs, and emergency response were $409 million, $323 million, and $55 million CAD, respectively. In addition to the June 2013 flood event, $186,831,824 was paid by insurance companies in response to 21,179 flood claims from the 2005 flood event that also affected Calgary (Dohy 2005).

The Bow River is located in Alberta, Canada, and flows through the city of Calgary. The headwaters are located in the Rocky Mountains at the Bow Glacier and merges with the Oldman River and eventually form the South Saskatchewan River (Fig. 1a). A hydrometric station had collected daily discharge data in the Bow River (05BH004, located at 51°03'00" N, 114°03'05" W) near Calgary from 2000 to 2018.

The Bow River drains a gross area of 7870 km^2. The average daily discharge of the Bow River is 90.7 m^3/s. The banks of this river overflow when the flow rate reaches 500 m^3/s and influences to structures, and overland flooding happens when the flow rate reaches 850 m^3/s (City of Calgary 2018). The daily discharge data related to the Bow River and statistical indices of these data for training and testing stages are presented in Fig. 1b and Table 1, respectively.

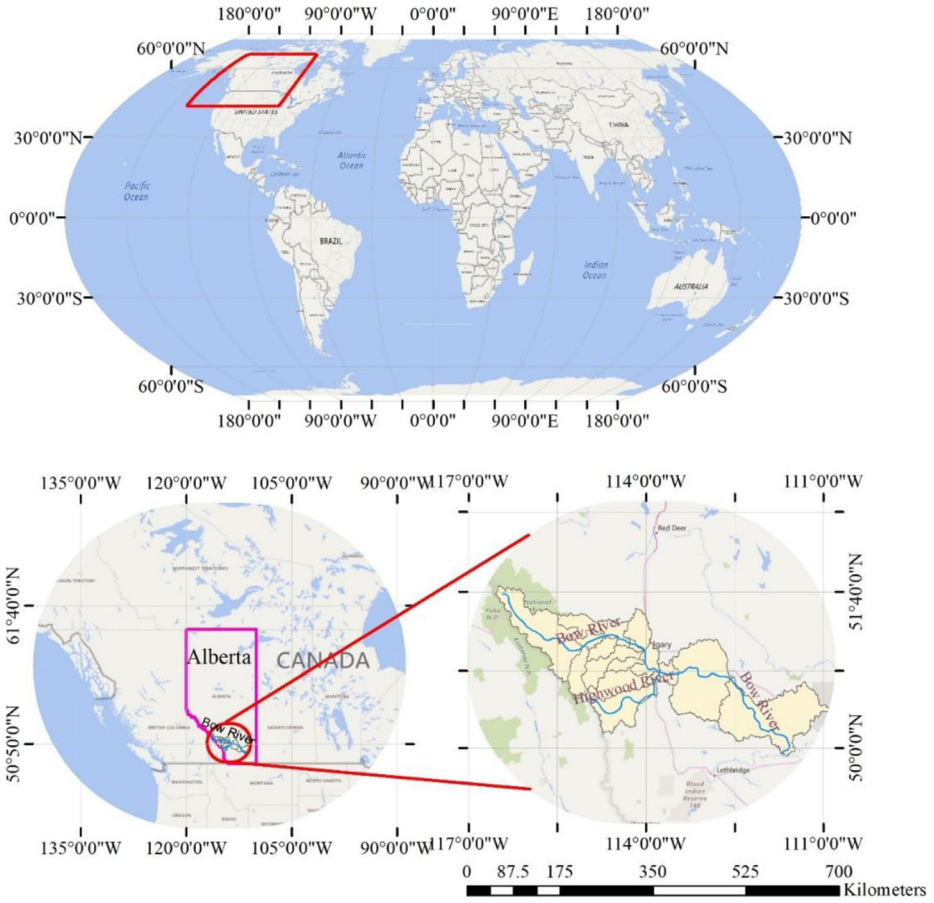
2 Theoretical Conceptions

2.1 Linear Modeling Conceptions

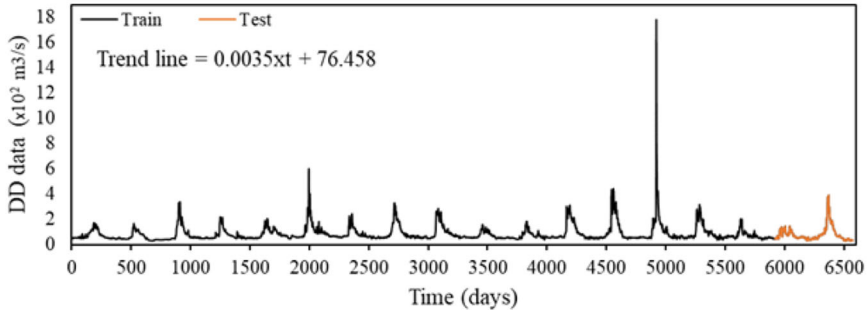
The autoregressive integrated moving average (ARIMA) is one of the most popular linear methods for predicting time series. This method may also be defined seasonally, which is then known as the Seasonal ARIMA (SARIMA). The ARIMA method is defined as:

$$\mathrm{ARIMA}(p,d,q) = \varphi(B)(1\ B)x(t) = \theta(B)\ \varepsilon(t) \tag{1}$$

where p and q are the order of autoregressive (AR) and moving average (MA), d is the differencing degree, φ and θ are the AR and MA parameters (respectively), $(1\text{-}B)^d$ is the dth



(a) Location of studied site



(b) Daily time series from January 2000 to January 2018

Fig. 1 The location and daily discharge of the Bow River near Calgary, Alberta, Canada

Table 1 Statistical Indices of Bow River daily discharge, divided into Total, Train, and Test Periods

Statistic	Nbr.	Min.	Max.	1st Q.	Median	3rd Q.	Mean	σ(n)	γ_1	γ_2
Total	6574	27.90	1750.00	54.60	64.20	93.50	86.38	64.95	6.81	114.85
Train	5917	30.10	1750.00	54.90	64.30	93.70	86.95	65.89	7.07	119.85
Test	657	27.90	392.00	49.60	62.80	91.30	81.27	55.63	2.64	8.35

Nbr., Number of data, Min. and Max., Minimum and Maximum of data, 1st Q. and 3rd Q., first and third Quarters, σ(n), Standard Deviation, γ_1, Skewness, γ2, Kurtosis

non-seasonal differencing operator, $x(t)$ is the raw time series, and the ε is the residual. Considering k as the non-seasonal parameters (φ and θ), the non-seasonal differencing is calculated as follows:

$$k(B)1\ k\ B\ kB^2{-}K_3B^3{-}\ldots K_nB^n \tag{2}$$

where n is the order of non-seasonal parameters (p and q). In the modelling of the time series using stochastic processes, the series should be subject to certain conditions. The Jarque-Bera (JB) (Jarque and Bera 1980) test is used to verify the normality of the series, and is defined as follows:

$$JB = n\left(\frac{S_K^2}{6} + \frac{(K_u{-}3)^2}{24}\right) \tag{3}$$

where K_u is elongation, S_K is skewed.

If the time series is regular, in the next step, the static term is evaluated, but if the series is not normal, it will normalize the series using the expression of Box and Cox (1964):

$$X_n(\lambda) = \begin{cases} \dfrac{(\mathcal{X} + \dashv)^{\lambda}}{\lambda} & \lambda{\neq}0 \\ \log(x + \dashv) & \lambda = 0 \end{cases} \tag{4}$$

where $X_n(\lambda)$ is the normalized time series, λ is the transform data, and α is a constant that $x_t + \alpha > 0$. We also evaluated the stationary of the time series, to make necessary transformations on the time series, if necessary before the series can be modelled. One of these tests, which is applied before the series, is the KPSS static time series test (Kwiatkowski et al. 1992) as follows:

$$S^2(l) = \frac{1}{n}\sum_{t=1}^{n} e_t^2 + \frac{2}{n}\sum_{j=1}^{1} w(j,l)\frac{1}{n}\sum_{t=j+1}^{n} e_t e_{t-s} \tag{5}$$

$$w(s,l) = 1{-}j/(l+1) \tag{6}$$

$$KPSS = \frac{1}{n^2}\sum_{t=1}^{n}\frac{S_t^2}{S^2(l)} \tag{7}$$

where S_t is Σe_t, l is the truncation lag. KPSS is a series static-statistic at level or trend. Each time series is formed from the four terms of trend, jump, period, and the stochastic term. The

existence of any of the first three terms in the time series causes the time series to become non-stationary. One of the most commonly utilized methods for time series stationary is differencing. In this method, the differential series is created by subtraction of two consecutive data values (i.e., Diff (t) = X (t) -X (t-1)). The trend and seasonal changes in the series are eliminated, and ultimately stationary the series can be obtained. Alternatively, the methods of differentiation (diff.), standardization (Std.), and spectral analysis (Sf.) can be used as time-series methods (Bonakdari et al. 2019). The non-parametric Mann-Kendal test is applied to test the process, to identify the gradual changes that occur over time in the time series (Jain and Kumar 2012). The standard of Mann-Kendall statistic (STD_{MK}), can be obtained as follows:

$$STDMK = \begin{matrix} (MK-1)\mathrm{var}(MK)^{-0.5} & Mk > 0 \\ 0 & MK = 0 \\ (MK+1)\mathrm{var}(MK)^{-0.5} & MK < 0 \end{matrix} \quad (8)$$

where MK is the Man-Kendall statistic, and $var(MK)$ represents the variance of MK. MK and var (MK) are calculated as:

$$MK = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{sgn}\left(X_j - X_i\right)$$

and

$$\mathrm{var}(MK) = \left(\left(2N^3 - 7N^2 - 5N\right) - \sum_{j}^{g} Obs_j \left(Obs_j - 1\right)\left(2Obs_j + 5\right) \right) / 18 \quad (9)$$

where X is data values, Obs_j is the number of observations at *the* j^{th} group, g is the number of identical groups, N is the number of samples, and sgn is the sign function. Gradual changes in the time series may occur alternately and seasonally, leading to a seasonal process in the time series. In this case, using the seasonal Mann-Kendall test, the seasonal process is identified as follows:

$$S_k = \sum_{i=1}^{N_k 1} \sum_{j=i+1}^{N_k - 1} \mathrm{sgn}\left(X_{ki} - X_{kj}\right) \quad (10)$$

$$SMK = \sum_{k=1}^{\omega} \left(S_k - sgn(S_k)\right) \quad (11)$$

$$\mathrm{var}(SMK) = 2 \sum_{k=1}^{\omega-1} \sum_{j=i+1}^{\omega} \sigma_{ij} + \sum_{k}^{\omega} \left(2N^3 k - 5N_k\right) / 18 \quad (12)$$

$$STD_{SMK} = MK\ \mathrm{var}(MK)^{-0.5} \quad (13)$$

where σ_{ij} is the covariance of statistic test in season i and j, and ω represents the number of the seasons in a year. If the probability of the statistics of these tests is higher than the significant level of 0.05, the time series has lacked any process. The jump the series can be tested with the following equation (Mann and Whitney 1947):

$$MW_U = \sum_{t=1}^{N_1}\left(Dg(X_{ordered}) - \frac{N_{m1(N_{m1}+N_{m2}+1)}}{2}\right) / \left((N_{m1}N_{m2}(N_{m1}+N_{m2}+1))^{0.5}/12\right) \quad (14)$$

In the relationship, $X_{ordered}$ is the series arranged according to the original $X(t)$, Dg ($X_{ordered}$) degrees of the $X_{ordered}$ function, N_{m1}, and N_{m2} is the number of members of the original series, $N_{m1} + N_{m2} = N_{total}$. The frequency in time series can be verified using the autocorrelation function (ACF) and the partial autocorrelation (PACF) diagrams. Another test that numerically examines time series is the Fisher test (Kashyap and Rao 1976). The test statistic is calculated as:

$$F^* = \frac{N(N-2)\left(a_k^2 + \beta_k^2\right)}{4\left(\sum_{z=1}^{k}\left(x(t) - a_z cos(\Omega_{z^t}) - \beta_z sin(\Omega_{z^t})\right)\right)} \quad (15)$$

where N is the number of sample data, F^* is the Fisher test statistic, α_z and β_z are Fourier coefficients, and Ω_z is the angular frequency obtained as follows:

$$a_z = \frac{2}{N}\left(\sum_{t=1}^{N} x(t) cos(2\pi f z^t)\right) z = 1, 2, \ldots, k \quad (16)$$

$$\beta_z = \frac{2}{N}\left(\sum_{t=1}^{N} x(t)(2\pi f z^t)\right) z = 1, 2, \ldots, k \quad (17)$$

$$f_z = \frac{z}{N}\Omega_z = \frac{2\pi z}{N}\ z = 1, 2, \ldots, k \quad (18)$$

In the above relationships, f_z is equal to the z-th harmonic of the base frequency. The periodicity of Ω_z is significant when the critical value F at the confidence level F (2, N-2) is lower than the F^* value.

$$F^* \geq F(2, N-2) \quad (19)$$

For a significant level of 0.05, the level of freedom in the denominator is equal to 3. The Ljung-Box test is used to check the validity of the modeling to verify the autonomy of the residuals of the time series (Ljung and Box 1978). The test statistic is calculated as follows:

$$\mathcal{Q}_m = N(N+2)\sum_{h=1}^{m}\frac{r_h}{N-1} \quad (20)$$

In this relationship, N is the number of samples, r_h is the correlation coefficient of the residues (εt) in delay h, m is equal to ln (N). If the probability of the Ljung-Box test statistic in the $\chi 2$ distribution is higher than the confidence level α (in this case PQ > α = 0.05), the residue series is independent, and the model is appropriate.

2.2 Group Method of Data Handling (GMDH)

The GMDH neural network arises from the bonding of different pairs through a quadratic polynomial by a set of neurons. The system describes a quadratic polynomial obtained by an

approximate function $\widehat{f}$ with output y from all neurons, for inputs $X = f(x_1, x_2, \ldots x_n)$ with the lowest error compared to the actual output of y. Therefore, for the observed sample M, including n inputs and one output, the results are represented in the form of:

$$y_i = f(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in})(i = 1, 2, \ldots, M) \tag{21}$$

In the GMDH method, a network that can predict the output value y for any input vector x can be calculated according to:

$$\widehat{y}_1 = \widehat{f}(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}) \quad (i = 1, 2, \ldots, M) \tag{22}$$

So that the mean square error between the observed values and the estimated values is minimized, as:

$$MSE = \frac{\sum_{i=1}^{M} \left(\widehat{y}_i - y_i\right)^2}{M} \rightarrow Min \tag{23}$$

The general formula of structure between the input and output variables can be represented using the polynomial function, as:

$$y = a_{0=} + \sum_{i=1}^{n} a_1 x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \tag{24}$$

The following second order form and two-variable polynomials are expressed as:

$$\widehat{y} = G(x_i, x_j) + a_0 + a_1 x_i + a_2 x_j + a_3 x_j^2 + a_4 x_j^2 + a_5 x_i x_j \tag{25}$$

The unknown coefficients a_i in the above equation are estimated by regression methods in such a way that the difference between the true output y and the estimated y values for each pair of input variables x_i and x_j is minimized. A set of polynomials is constructed using Eq. (25), all unknown coefficients are calculated by the least squares (LS) method. The coefficients of each neuron equation (for each function G_i) are obtained by minimizing its total error to adapt the inputs to all pairs of input-output sets optimally.

$$E = \frac{\sum_{i=1}^{M} (y_i - G_i)^2}{M} \rightarrow Min \tag{26}$$

In the GMDH algorithm, all dual neurons are constructed of n input variables, and unknown coefficients of all neurons are calculated using the LS method. Therefore, the number of neurons to build the second layer are $(n \;\; 2) = \frac{n(n-1)}{2}$, which can be represented as the following set:

$$\left\{ (y_i, x_{ip}, x_{iq}) | (i = 1, 2, \ldots, M) \& p, q \in (1, 2, \ldots, M) \right\} \tag{27}$$

From the quadratic form of the function expressed in the relationship (5), each M triple row is used; these equations can be expressed in the form of the following matrix:

$$Aa = Y \tag{28}$$

where A is the vector of unknown coefficients of the second order equation shown in Eq. (25) and:

$$a = \{a_0, a_1, \ldots a_5\} \tag{29}$$

$$Y = \{y_1, y_2 \ldots y_M\}^T \tag{30}$$

$$A = \begin{bmatrix} 1 & x_{1p} & x_{1p} & x_{1p}^2 & x_{1q}^2 & x_{1p}x_{1p} \\ 1 & x_{2p} & x_{2p} & x_{2p}^2 & x_{2p}^2 & x_{2px_{2q}} \\ & & \cdot & \cdot & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & \cdot & \cdot & \\ 1 & x_{Mp} & x_{Mp} & x_{Mp}^2 & x_{Mq}^2 & x_{mp}x_{Mq} \end{bmatrix} \tag{31}$$

The least squared method of multi-regression analysis calculates the equations in the form of the following equation:

$$a = \left(A^T A\right)^{-1} A^T Y \tag{32}$$

This equation generates a vector of coefficients of Eq. (25) for all three triangular M sets.

2.3 Generalized Structure of GMDH

Although GMDH has a great ability to model non-linear problems, this method is subject to some limitations, including 1) the use of a second-order polynomial; 2) inputs of each neuron are provided only from adjacent neurons, and 3) each neuron only has two inputs. Therefore, this method may not be accountable for issues of considerable complexity. Thus, in this study to address the problems presented, GMDH generalized structure (GS-GMDH) is introduced. In this method, neuron inputs can be two to three. In addition to the second-order polynomials, the third-order polynomial can also be used. The inputs of each neuron can be from adjacent layer neurons, and can also use the neurons of nonadjacent layers.

2.4 The Structure of the Proposed Models

In this study, two linear and non-linear methods for modelling the daily discharge of the Bow River are presented. In the non-linear process, the GMDH algorithm is used where, as explained in the previous section, the structure of this method has been modified so that it has advantages over the classical GMDH method. The linear method used is the ARIMA method, which has been used by several previous researchers. First we evaluated the normalization of the time series training data by the Jarque-Bera test. In the case of non-normality, the Box-Cox Transform normalization is completed. The stationary time series is then assessed using the KPSS test. Jump and period are other definite terms. By using the Mann-Whitney and Fisher tests (respectively), the existence of jump and period are examined, and by

differencing, standardization and spectral analysis are removed. After eliminating definite terms, the time series modelling is performed using the ARIMA method. After modeling, the independence of the residuals is evaluated using the Ljung-Box test. Following the verification step, the accuracy of the linear modelling results and the GS-GMDH methods are evaluated using the test data (Fig. 2).

3 Modelling Evaluation Measures

Due to the stochastic nature of the hydrological variables, the use of single criteria to assist in the execution of a statistical model is not enough. In this study, the coefficient of determination (R^2), as well as two relative indices (mean absolute percentage error (*MAPE*) and root mean square relative error (*RMSRE*)), are used to establish the efficacy of a linear and non-linear model.

$$R^2(\%) = 100 \times \left(\frac{\left(\sum_{1=1}^{n} \left(x_{obsi} - \overline{X}_{obst} \right) \left(X_{pi} - \overline{X}_{Pt} \right) \right)}{\sqrt{\sum_{i=1}^{n} \left(X_{obsi} - \overline{X}_{obst_t} \right)^2 \sum_{i=1}^{n} \left(X_{i=1} - \overline{X}_{Pt} \right)^2}} \right) \tag{33}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{X_{obs,i} - X_{p,1}}{X_{obs,i}} \tag{34}$$

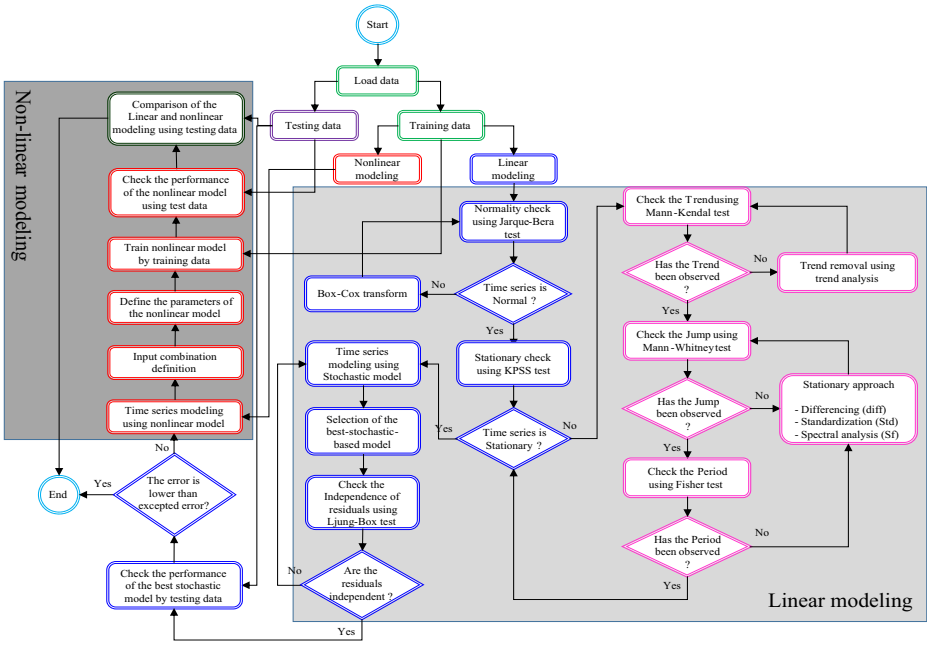


Fig. 2 The structure of the proposed model

$$RMRSE = (100) \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_{obs,i}-X_{P,i}}{X_{obs,i}}\right)^2} \tag{35}$$

The value of R^2 bounded by [0, 1] explains the covariance in the actual daily discharge data that can be described by the predicting model, but it originates from the linear assumptions (Krause et al. 2005). The R^2, *RMSRE,* and *MAPE* are insensitive to outliers (Legates and Mccabe 1999;). The Nash-Sutcliffe coefficient ($E_{\text{N-S}}$) is employed to overcome the limitation of the previously mentioned indices. Since neither the R^2, *RMSRE*, *MAPE,* nor $E_{\text{N-S}}$ consider the complexity of the model, the Akaike information criterion (*AIC*) is used to compare the performance of linear and non-linear models regarding accuracy and complexity simultaneously.

$$E_{N-S}(\%) = \left[\frac{\sum_{i=1}^{N}\left(X_{obs,i}-X_{P,i}\right)^2}{\sum_{i=1}^{N}\left(X_{obs,i}-\overline{X}_{obs}\right)^2}\right] \times 100 \tag{36}$$

$$AIC = N1n\left(\sum_{i=1}^{N}\left(X_{obs,i}-X_{P,i}\right)^2\right) + 2k \tag{37}$$

In the above equations k is the number of parameters, N number of samples, $X_{obs,i}$ and $X_{P,i}$ are respectively the i^{th} value of observed and predicted value.

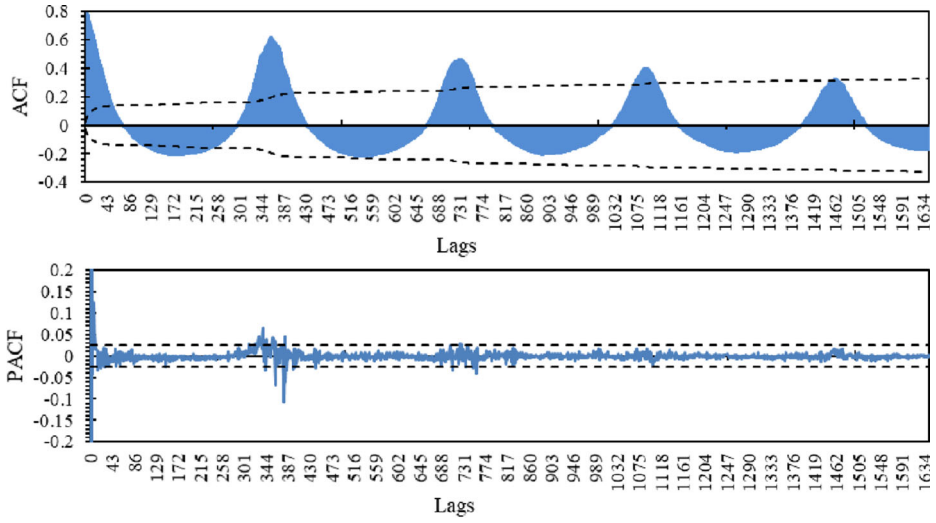
4 Development of Linear and Non-linear Modeling

4.1 Linear Modelling

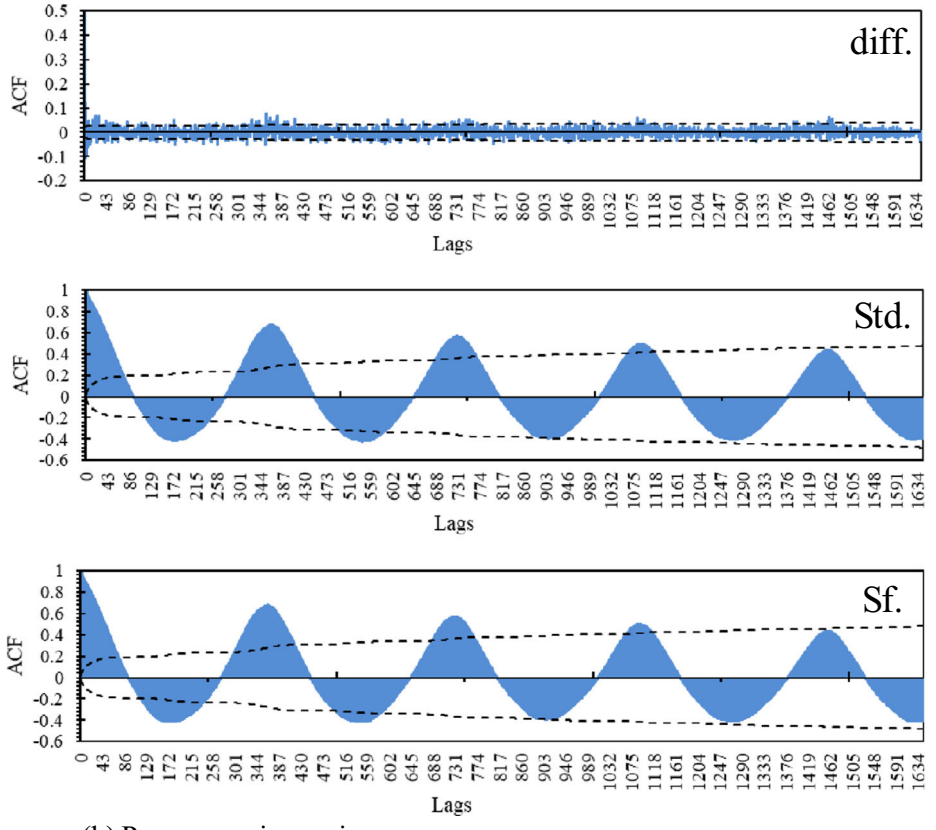
For linear modelling using the ARIMA model, the time series features need to be well identified and, if necessary, be static and normal using appropriate pre-processes. Therefore, at first, the correlations of the daily discharge (DD) series are plotted (Fig. 3a). It can be seen that the DD time series is volatile and has strong seasonal and non-seasonal correlations. Non-seasonal correlations of up to 52 primary lags and seasonal correlations of up to four lags with 365-day steps exist. The period should be eliminated by the appropriate methods in the residual time series.

Table 2 presents different test results for the numerical verification of the DD time-series features. In this table, it can be seen that the DD time series has seasonal and non-seasonal trends. Also, based on the fisher and JB test statistic, the all-time series are periodic and have no normal distribution. As the figure shows, the DD time series has a jump in the validation period, which is confirmed by the Mann-Whitney test. Despite these features, the DD time series is non-stationary based on the KPSS numerical test.

The ACF graphs of the series were re-drawn (Fig. 3b) to investigate the changes in the pre-processed series. In this figure, diff, Std and Sf represent differencing, non-seasonal standardization and spectral analysis and the changes from pre-processing in the series are clearly seen. The degree of seasonal and non-seasonal correlations in the series of differential equations has been greatly reduced, and the stationary of the pre-processed time series is evident. Standardization and spectral analysis methods have reduced the amount of seasonal and non-seasonal relations, but



(a) Daily discharge



(b) Pre-processing series

Fig. 3 Autocorrelation function plot of pre-processed BRDD data with proposed methods for N/4 of data: a) daily data, b) Pre-processing data with three methods

they have not been able to make the series stationary, and it can be seen that these correlations are still high. Therefore, the ARMA model cannot be used for modelling. Differencing is done to examine the possibility of data modelling using the ARIMA model.

The results are presented in Table 2 which shows that both the seasonal and non-seasonal trends and the jumps in the series have been eliminated. Although the periodic term has been created in standardization and spectral analysis methods, it can be seen that the series are considered stationary. Changes in the correlation diagrams of these series are shown in Fig. 4. Seasonal correlations have been eliminated, and the graphs have been taken up to a maximum of two lags. Therefore, using the ARIMA linear model with a maximum of the two non-seasonal parameters p and q and one differencing is very suitable.

Table 2. Test results of applied tests on BRDD data and pre-processed outcomes.

4.2 Nonlinear Modeling

Using the graphs presented in Fig. 2 and considering that in the GS-GMDH method, at least two variables should be considered as inputs, several models were considered as follows:

$$\mathrm{M1}: Q(t) = Q(t-1), Q(t-2)$$

$$\mathrm{M2}: Q(t) = Q(t-1), Q(t-2), Q(t-3)$$

$$\mathrm{M3}: Q(t) = Q(t-1), Q(t-2), Q(t-3), Q(t-4)$$

$$\mathrm{M4}: Q(t) = Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5)$$

Using the GS-GMDH method and considering the four models above, various relationships are proposed to predict the Bow River discharge, as shown in Table 3.

5 Results and Discussion

Figure 5 indicates the scatter plot of the ARIMA-based linear method (diff, Std, Sf) and GS-GMDH (M1-M4) techniques in daily discharge prediction. Comparison of GS-GMDH models

Table 2 Test results of applied tests on and pre-processed outcomes

Data	Tests	Trend		Jump	Period	Stationary	Norm.
Original data		MK%	SMK%	MW%	(F*)*	KPSS%	JB*
	DD	0.01	0.01	0.03	39,130	4.00	3,590,440.16
	diff	55.77	85.22	99.02	0	99.51	9.45
	Std	0.01	0.01	0.03	−146,729	0.21	9.45
	Sf	26.05	0.01	27.73	0	0.18	23.94
Subtracted data	*diff*	25.08	31.09	28.72	0	100.00	49,435.19
	Std	55.77	85.22	99.02	67,042	99.51	21,824.23
	Sf	50.89	79.73	93.18	266	98.92	21,817.62

*. *TEST statistics; Fisher critical value:* ***3****; JB critical value:* ***5.99***

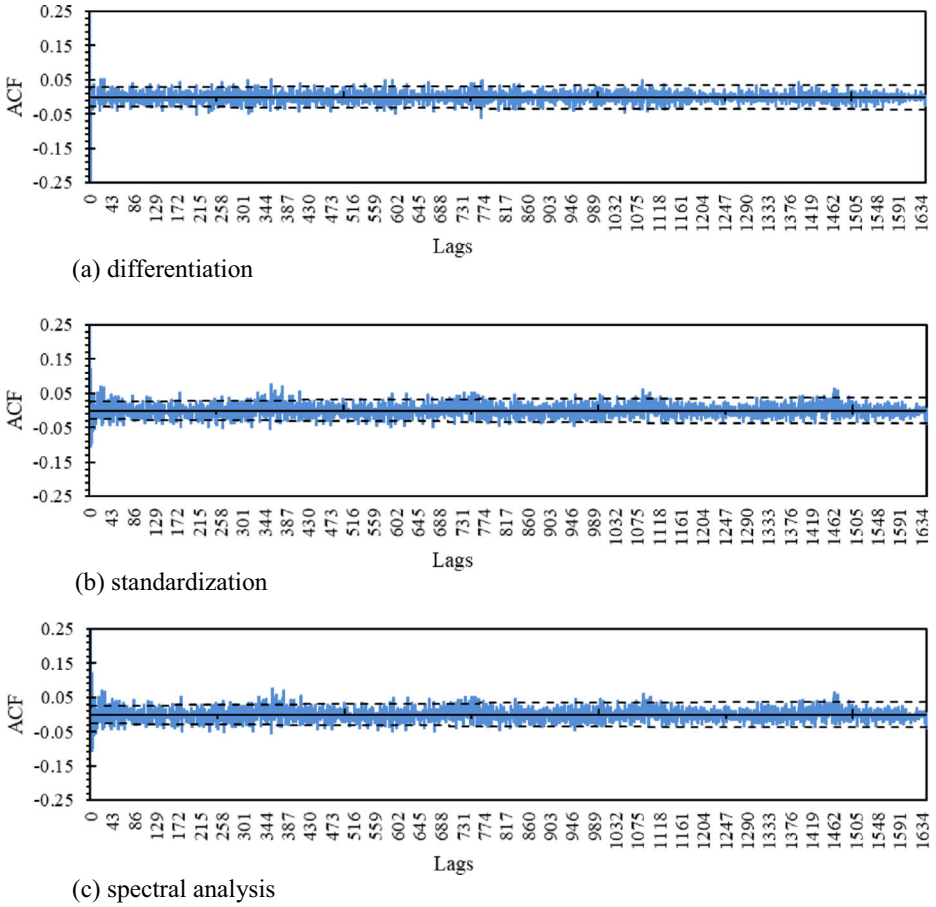


Fig. 4 Autocorrelation function plot of subtracted (ARIMA differencing operator) pre-processed BRDD data with proposed methods: a. diff., b. Std., c. Sf, for N/4 of data

with linear models shows a relatively similar function to the non-linear and linear method, with the difference that the maximum discharge for model testing in the GS-GMDH method has a better performance compared with linear methods (Std, Sf).

Figure 6a depicts the box plot of the observed and predicted daily discharge**.** It is observed that the performance of all methods (linear and non-linear) in different domains is approximately the same, so that the average values estimated with the average observed amounts are approximately equal. The scattering of these values (observed and estimated) in the first and third quantile is also similar. As observed in the scatter plot, the main difference between the performances of the models is in the peak discharges.

The qualitative comparison of the two sets of models presented in this study (Figs. 5 and 6a) depicts the good and similar performance of both models in estimating the Bow River daily discharge.

Figure 6b presents the box plot for a relative error of the ARIMA (diff, Std, Sf) and GS-GMDH (M1-M4) models for estimation of the Bow River daily discharge. The distribution of the relative error in non-linear and linear methods shows that the average value of the relative

Table 3 The proposed GS-GMDH equations for M1 to M4

Model	Equation	Eq. No.
M1	$Q(t) = 3.7962 + 1.1207 \times Q(t-1) - 0.222 \times Q(t-2) + 0.0063 \times Q(t-2) \times Q(t-1) - 0.0011 \times (Q(t-1))^2 - 0.0045 \times (Q(t-1))^2 - 5.7949E-05 \times Q(t-2) \times (Q(t-1))^2 + 6.023E-05 \times (Q(t-1))^2 \times Q(t-1) + 1.4284E-05 \times (Q(t-1))^3 - 1.8197E-05 \times (Q(t-2))^3$	38
M2	$Q(t) = 3.4087 + 1.0113 \times Q(t-1) - 0.1174 \times Q(t-2) + 0.0103 \times Q(t-3) + 0.0026 \times Q(t-1) \times Q(t-2) + 0.0033 \times Q(t-1) \times Q(t-3) - 0.0052 \times Q(t-2) \times Q(t-3) + 1.3633e-04 \times (Q(t-1))^2 - 0.0017 \times (Q(t-2))^2 + 0.0016 \times (Q(t-3))^2 - 2.4461e-05 \times Q(t-1) \times Q(t-2) \times Q(t-3) - 5.965E-05 \times Q(t-2) \times (Q(t-1))^2 + 7.6583E-05 \times Q(t-1) \times (Q(t-2))^2 - 3.2308E-06 \times Q(t-3) \times (Q(t-1))^2 + 4.2008E-05 \times Q(t-3) \times (Q(t-2))^2 + 1.4494E-05 \times Q(t-1\times)(Q(t-3))^3 - 3.2514E-05 \times Q(t-2) \times (Q(t-3))^2 + 1.397E-05 \times (Q(t-1))^3 - 3.5313E-05 \times (Q(t-2))^3 + 6.3676E-06 \times (Q(t-3))^3$	39
M3	$Q(t) = 0.0331 + 0.1525 \times Y_3 + 0.3472 \times Y_2 + 0.5012 \times Y_1 - 0.5904 \times Y_{2\times} Y_3 0.6234 \times Y_3 + 0.3568 \times Y_1 Y_2 - 0.0168 \times Y_3^2 + 0.1156 \times Y_2^2 - 0.4887 \times Y_1^2 - 0.0183 \times Y_1 \times Y_2 \times Y_3 + 0.0073 \times Y_2 \times Y_3^2 + 0.0026 \times Y_2^2 \times Y_3 + 0.0039 \times Y_1 \times Y_3^2 - 0.0168 \times Y_1 \times Y_2^2 + 0.0041 \times Y_1^2 \times Y_3 + 0.01129 \times Y_1^2 \times Y_2 - 0.0036 \times Y_3^3 - 2.1795E-05 \times Y_2^3 - 0.0045 \times Y_1^3$	40
	$Y_1 = 3.409 + 1.0113 \times Q(t-1) - 0.1174 \times Q(t-2) + 0.0103 \times Q(t-3) + 0.0026 \times Q(t-2) \times Q(t-1) + 0.0033 \times Q(t-3) \times Q(t-1) - 0.0052 \times Q(t-3) \times Q(t-2) + 0.00014 \times (Q(t-1))^2 - 0.0017 \times (Q(t-2))^2 + 0.0016 \times (Q(t-3))^2 - 2.446E-05 \times Q(t-3) \times Q(t-2) \times Q(t-1) - 5.965E-05 \times Q(t-2) \times (Q(t-1))^2 + 7.6582E-05 \times Q(t-1) \times (Q(t-2))^2 - 3.2306E-06 \times Q(t-3) \times (Q(t-1))^2 + 4.2E-05 \times Q(t-3) \times (Q(t-2))^2 + 1.4494E-05 \times Q(t-1) \times (Q(t-3))^2 - 3.2513E-05 \times Q(t-2) \times (Q(t-3))^2 + 1.397E-05 \times (Q(t-1))^3 - 3.5312E-05 \times (Q(t-2))^3 + 6.3672E-06 \times (Q(t-3))^3$	40–1
	$Y_2 = 3.7944 + 1.1208 \times Q(t-1) - 0.222 \times Q(t-2) + 0.0063 \times Q(t-2) \times Q(t-1) - 0.0011 \times (Q(t-1))^2 - 0.0045 \times (Q(t-1))^2 - 5.7948E-05 \times Q(t-2) \times (Q(t-1))^2 + 6.0229E-05 \times Q(t-1) \times (Q(t-1))^2 + 1.4283E-05 \times (Q(t-1))^3 - 1.8197E-05 \times (Q(t-2))^3$	40–2
	$Y_3 = 3.8837 + 0.9884 \times Q(t-1) - 0.269 \times Q(t-2) + 0.0285 \times Q(t-4) + 0.0022 \times Q(t-2) \times Q(t-1) + 0.0037 \times Q(t-4) \times Q(t-1) - 0.0047 \times Q(t-4) \times Q(t-2) + 0.0002 \times (Q(t-1))^2 - 0.0012 \times (Q(t-2))^2 + 0.008 \times (Q(t-4))^2 - 9.5629E-07 \times Q(t-4) \times Q(t-2) \times Q(t-1) - 5.8066E-05 \times Q(t-2) \times (Q(t-1))^2 + 6.6029E-05 \times Q(t-1) \times (Q(t-2))^2 - 5.955E-06 \times Q(t-4) \times (Q(t-1))^2 + 6.8678E-06 \times Q(t-4) \times (Q(t-2))^2 + 2.8258E-06 \times Q(t-1) \times (Q(t-4))^2 - 3.5653E-06 \times Q(t-2) \times (Q(t-4))^2 + 1.4104E-05 \times (Q(t-1))^3 - 2.3377E-05 \times (Q(t-2))^3 + 93633E-08 \times (Q(t-4))^3$	40–3
M4	$Q(t) = 0.5274 + 0.1539 \times Y_3 - 0.0.9668 \times Y_2 + 2.1095 \times Y_1 - 0.2516 \times Y_{2\times} Y_3 + 0.0049 \times Y_1 \times Y_3 - 0.4121 \times Y_1 \times Y_2 - 0.1242 \times Y_3^2 + 0.0827 \times Y_2^2 + 0.1971 \times Y_1^2 - 0.0123 \times Y_1 \times Y_2 \times Y_3 + 1.96E-05 \times Y_2 \times Y_3 + 0.0055 \times Y_2^2 \times Y_3 + 0.0076 \times Y_1 \times Y_3^2 + 0.0069 \times Y_1 \times Y_2^2 - 0.013 \times Y_1^2 \times Y_3 - 0.0003 \times Y_1^2 \times Y_2 - 0.0024 \times Y_3^3 - 0.004 \times Y_2^2 + 0.0003 \times Y_1^3$	41
	$Y_1 = 3.021 + 1.8997 \times Q(t-1) - 0.0268 \times Q(t-2) + 0.0407 \times Q(t-5) + 0.0004 \times Q(t-2) \times Q(t-1) + 0.0006 \times Q(t-5) \times Q(t-1) - 0.0056 \times Q(t-5) \times Q(t-2) + 0.0009 \times (Q(t-1))^2 - 0.0002 \times (Q(t-2))^2 + 0.0006 \times (Q(t-5))^2 + 2.0506E-05 \times Q(t-5) \times Q(t-5) \times Q(t-1) - 5.7532E-05 \times Q(t-2) \times (Q(t-1))^2 + 6.1046E-05 \times Q(t-1) \times (Q(t-2))^2 + 1.1519E-05 \times Q(t-5) \times (Q(t-1))^2 - 8.2515E-06 \times Q(t-5) \times (Q(t-2))^2 - 5.7532E-05 \times Q(t-2) \times (Q(t-1))^2 + 6.1046E-05 \times Q(t-1) \times (Q(t-2))^2 - 1.1519E-05 \times Q(t-5) \times (Q(t-1))^2 + 8.2515E-06 \times Q(t-5) \times (Q(t-2))^2 - 4.1869E-06 \times Q(t-1) \times (Q(t-5))^2 + 3.7425E-06 \times Q(t-2) \times (Q(t-5))^2 + 1.4335E-05 \times (Q(t-1))^3 - 1.9616E-05 \times (Q(t-1))^3 - 1.7397E-07 \times (Q(t-5))^3$	41–1
	$Y_2 = 3.8828 + 0.9884 \times Q(t-1) - 0.1269 \times Q(t-2) + 0.0285 \times Q(t-4) + 0.0022 \times Q(t-1) \times Q(t-2) + 0.0037 \times Q(t-1) \times Q(t-4) - 0.0047 \times Q(t-4) \times Q(t-2) + 0.0002 \times (Q(t-1))^2 0.0012 \times (Q(t-2))^2 + 0.008 \times (Q(t-4))^2 - 9.5623E-07 \times Q(t-1) \times Q(t-2) \times Q(t-4) - 5.8066E-07 \times (Q(t-1))^2 \times Q(t-2) + 6.6029$	41–2

Table 3 (continued)

Model	Equation	Eq. No.
	$E-05 \times (Q(t-2))^{2} \times Q(t-1) -5.9551E-06 \times (Q(t-1))^{2} \times Q(t-4) +6.8679E-06 \times (Q(t-2))^{2} \times Q(t-4)\ 2.8259E-06 \times (Q(t-4))^{2} \times Q(t-1) +3.5653E-06 \times (Q(t-4))^{2} \times Q(t-2) +1.4104E-05 \times (Q(t-1))^{3} -2.3377E-05 \times (Q(t-2))^{3} -9.363E-08 \times (Q(t-4))^{3}$	
	$Y_3 = 3.408 + 1.0113 \times Q(t-1)\ -0.1175 \times Q(t-2) +0.0103 \times Q(t-3) +0.0026 \times Q(t-1) \times Q(t-2) +0.0033 \times Q(t-3) \times Q(t-1) -0.0052 \times Q(t-3) \times Q(t-2) +0.0001 \times (Q(t-1))^{2} 0.0017 \times (Q(t-2))^{2} + 0.016 \times (Q(t-3))^{2} -2.4458E-05 \times Q(t-3) \times Q(t-2) \times Q(t-1) -5.9649E-05 \times Q(t-2) \times (Q(t-1))^{2} + 7.658E-05 \times Q(t-1) \times (Q(t-2))^{2} -3.2308E-06 \times Q(t-3) \times (Q(t-1))^{2} +4.2005E-05 \times Q(t-3) \times (Q(t-2))^{2} + 1.4493E-05 \times Q(t-1) \times (Q(t-3))^{2} + 3.2512E-05 \times Q(t-2) \times (Q(t-3))^{2} +1.3975E-05 \times (Q(t-1))^{2} -3.5311E-05 \times (Q(t-2))^{3} + 6.3672E-06 \times (Q(t-3))^{2}$	41–3

error for all purposes is less than 10%. Regardless of the outlier errors, the maximum relative error of the methods used is less than 20%. The performance of the models with respect to outlier relative errors shows that the maximum error is due to linear methods, and especially due to the diff method. The minimum value associated with the maximum error measured for outlier relative errors is related to the GS-GMDH (M1) method.

The performance evaluation of the ARIMA (diff, Std, Sf) and GS-GMDH (M1-M4) methods qualitatively confirmed the ability of these two methods for the prediction of Bow River daily discharge. For a closer comparison of these two models and determination of the superior model, several quantitative studies are required. The indices presented in Table 4

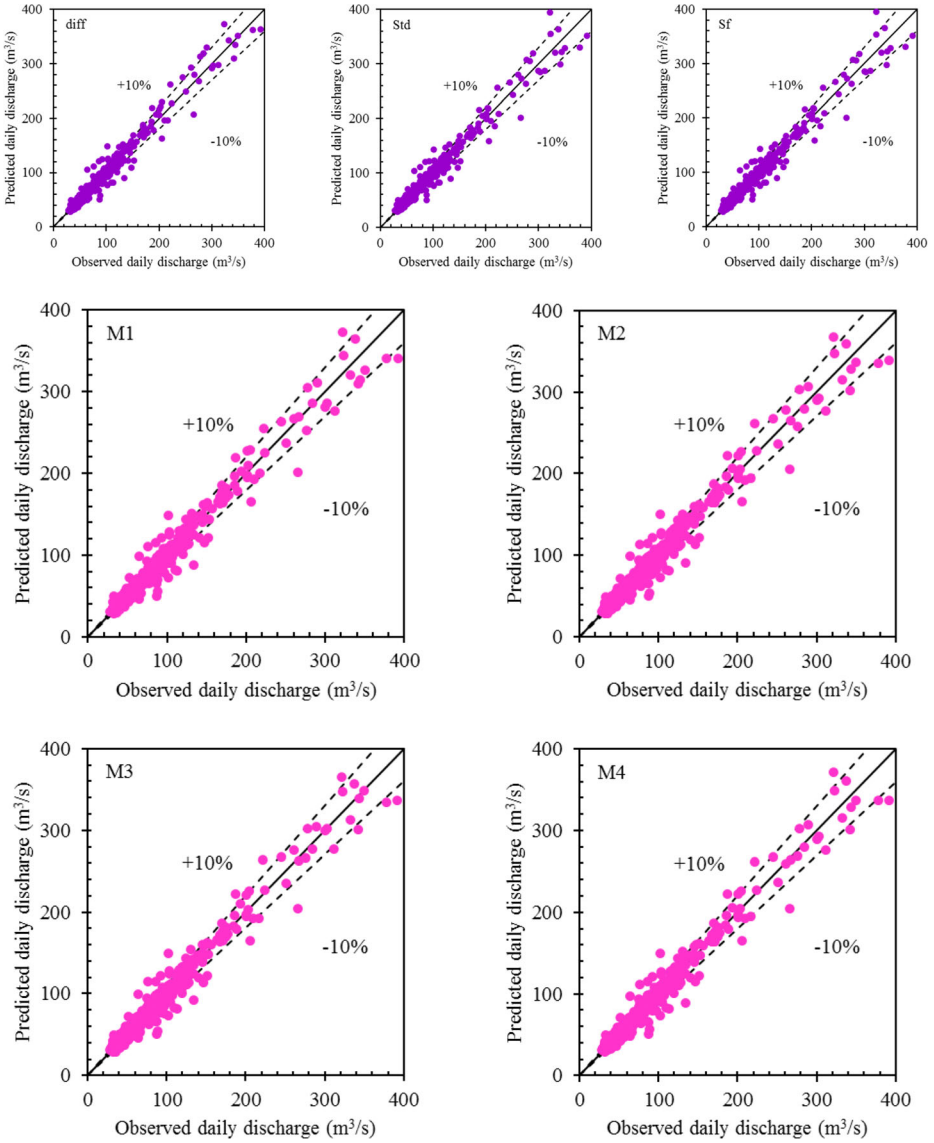
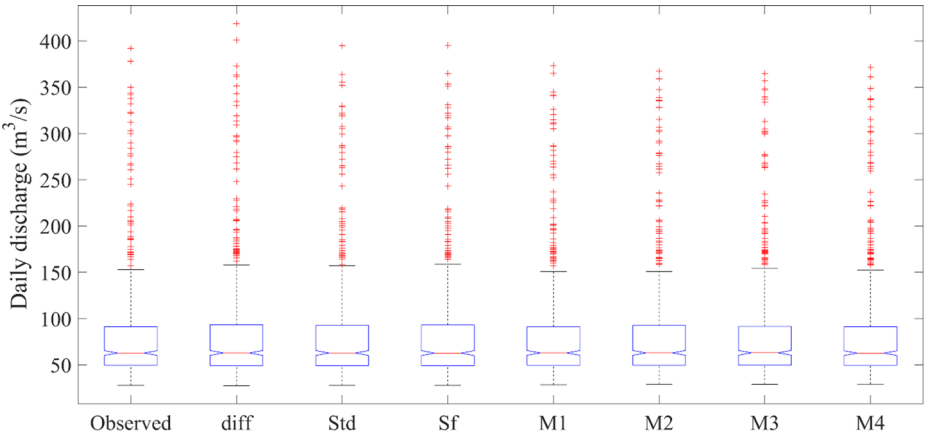
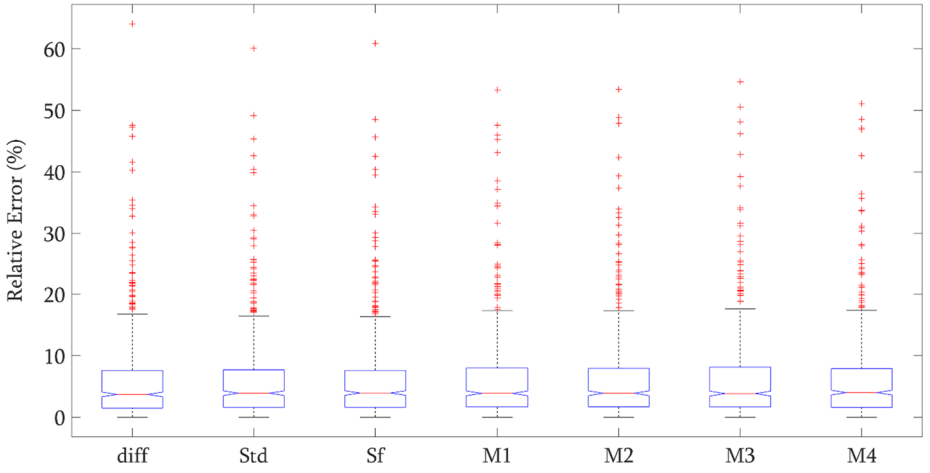


Fig. 5 Scatter plot of the linear (diff, Std, Sf) and non-linear (M1-M4) models in daily discharge prediction



(a) daily discharge



(b) pre-processed data

Fig. 6 The box plot for a relative error of the ARIMA: a) observed and predicted daily discharge; and b) pre-processed data (diff, Std, Sf) and GS-GMDH (M1-M4) models

confirm the significant performance of the proposed models in this study, which were qualitatively examined. The average relative error of these models is about 6%, and all models have a very high correlation coefficient. An obvious point in choosing the superior model is the use of an index that has a great deal of accuracy and simplicity.

The complexity of the model is evaluated using the AIC index. The more superior model will have the smaller lower and upper limits of this index. For linear models, the values of p and q used in the ARIMA model are considered as k in the definition of the AIC relationship, while for the GS-GMDH model, the coefficients are used to estimate the GS-GMDH model. In linear methods, the lowest AIC is the Std method. The AIC value in this method is slightly better than Sf, but its difference is significant compared to diff.

In non-linear methods, the values of all indices, except for AIC, are constant in all models, so that with increasing inputs, not only is the accuracy of the model not significantly changed,

but this also leads to the increased complexity of the model relative to the model with lower input parameters. Therefore, considering that the GS-GMDH (M1) method has the lowest AIC among both linear and non-linear methods, this model is selected as the superior model for predicting the Bow River daily discharge.

6 Conclusions

In this study, the accuracy of a linear stochastic model and non-linear GMDH daily discharge forecast models were compared. The linear stochastic method incorporates three input data pre-processing methods of differencing (diff), standardization (Std), and spectral analysis (Sf). In addition to the linear methodology, a non-linear method based on the GMDH was developed. A summary of the most notable results are listed as follows:

- The proposed GS-GMDH improved the results of classical GMDH by considering more than two input parameters in each neuron, admissibility of the input of each neuron from nonadjacent layers and employing second- and third-order polynomials to build the structure between the input and output variables.
- Comparison of the linear stochastic and the non-linear GMDH methods showed that all linear methods (diff, Std and Sf) and non-linear methods (M1-M4) have high accuracy in forecasting the Bow River daily discharge with an average relative error below 6%.
- This study showed that an appropriate pre-processing process can improve the results of a stochastic model and it can provide a similar forecast accuracy of the daily discharge compared to the more complex non-linear GMDH model.
- Comparison of all methods using an index that considers simultaneously the accuracy and simplicity of the model (AIC) showed that the GS-GMDH (M1) method has the best performance among all considered methods and can be used in practical applications.

Compliance with Ethical Standards

Conflict of Interest None

References

Bonakdari H, Moeeni H, Ebtehaj I, Zeynoddin M, Mahoammadian A, Gharabaghi B (2019) New insights into soil temperature time series modeling: linear or non-linear? Theor Appl Climatol 135:1155–1177

Table 4 Statistical indices for linear and non-linear methods

Model		R^2 (%)	*MAPE*	*RMSRE* (%)	*AIC*	$E_{N\text{-}S}$ (%)
ARIMA	diff	97.00	5.96	9.27	3050.26	96.93
	Std	97.02	5.94	9.09	2980.95	96.98
	Sf	97.02	5.92	9.08	2981.17	96.97
GMDH	M1	97.17	6.07	9.25	2957.99	97.09
	M2	97.15	6.15	9.34	2981.06	97.08
	M3	97.17	6.14	9.37	3057.89	97.11
	M4	97.21	6.04	9.14	3087.58	97.14

Box GE, Cox DR (1964) An analysis of transformations. J R stat Soc series B:211-252
City of Calgary (2018) Understanding river flow rates. Retrieved from Calgary: http://www.calgary.ca/UEP/Water/Pages/Flood-Info/Types-of-flooding-in-Calgary/Understanding-river-flow-rates.aspx
Dohy L (2005) Flood costs soaring in Alberta. Infomart, Postmedia Network Inc., Don Mills
Ebtehaj I, Zeynoddin M, Bonakdari H (2020) Discussion of "comparative assessment of time series and artificial intelligence models to estimate monthly streamflow: a local and external data analysis approach" by Saeid Mehdizadeh, Farshad Fathian, Mir Jafar Sadegh safari and Jan F. Adamowski J Hydrol 583:124614
Gharabaghi B, Sattar A (2019) Empirical models for longitudinal dispersion coefficient in natural streams. J Hydrol 575:1359–1361
Gholami A, Bonakdari H, Mohammadian M, Zaji AH, Gharabaghi B (2019) Assessment of geomorphological bank evolution of the alluvial threshold rivers based on entropy concept parameters. Hydrolog Sci J 64(7):856–872
Insurance Bureau of Canada (2017) Facts of the property and casualty insurance industry in Canada 2017 . Insurance Bureau of Canada
Jarque CM, Bera AK (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ Lett 6(3):255–259
Kashyap RL, Rao AR (1976) Dynamic stochastic models from empirical data. Academic Press, New York, USA
Kelly G, Stodolak P (2013) Why insurers fail, natural disasters and catastrophes. Casualty Insurance Compensation Corperation, Toronto
Khatibi R, Sivakumar B, Ghorbani MA, Kisi O, Kocak K, FarsadiZadeh D (2012) Investigating chaos in river stage and discharge time series. J Hydrol 414–415:108–117
Krause P, Boyle DP, Base F (2005) Comparison of different efficiency criteria for hydrological model assessment. Adv Geosci 5:89–97
Kwiatkowski D, Phillips PC, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J Econ 54(1–3):159–178
Jain SK, Kumar V (2012) Trend analysis of rainfall and temperature data for India. Curr Sci 102(1):37–49
Legates DR, Mccabe GJ (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 35:233–241
Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. Biometrika 65(2):297–303
Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18:50–60
Mosavi A, Ozturk P, Chau KW (2018) Flood prediction using machine learning models: literature review. Water 10(11):1536
Najafzadeh M, Barani GA, Hessami-Kermani MR (2015) Evaluation of GMDH networks for prediction of local scour depth at bridge abutments in coarse sediments with thinly armored beds. Ocean Eng 104:387–396
Pomeroy J, Stewart RE, Whitfield PH (2016) The 2013 flood event in the South Saskatchewan and Elk River basins: causes, assessment and damages. Can Water Resour J 41(1–2):105–117
Serinaldi F, Loecker F, Kilsby CG, Bast H (2018) Flood propagation and duration in large river basins: a data-driven analysis for reinsurance purposes. Nat Hazards 94:71–92
Walton R, Binns A, Bonakdari H, Ebtehaj I, Gharabaghi B (2019) Estimating 2-year flood flows using the generalized structure of the group method of data handling. J Hydrol 575:671–689

Affiliations

Hossein Bonakdari[1] · **Andrew D. Binns**[2] · **Bahram Gharabaghi**[2]

[1] Department of Soils and Agri-Food Engineering, Laval University, Québec G1V0A6, Canada

[2] School of Engineering, University of Guelph, Guelph, Ontario N1G 2W1, Canada