



A new evolutionary time series model for streamflow forecasting in boreal lake-river systems

Ali Danandeh Mehr^{1,2} · Sahand Ghadimi¹ · Hannu Marttila¹ · Ali Torabi Haghighi¹

Received: 28 June 2021 / Accepted: 12 January 2022 / Published online: 24 January 2022
© The Author(s) 2022

Abstract

Genetic programming (GP) is an evolutionary regression method that has received considerable interest to model hydro-environmental phenomena recently. Considering the sparseness of hydro-meteorological stations on northern areas, this study investigates the benefits and downfalls of univariate streamflow modeling at high latitudes using GP and seasonal autoregressive integrated moving average (SARIMA). Furthermore, a new evolutionary time series model, called GP-SARIMA, is introduced to enhance streamflow forecasting accuracy at long-term horizons in a lake-river system. The paper includes testing the new model for one-step-ahead forecasts of daily mean, weekly mean, and monthly mean streamflow in the headwaters of the Oulujoki River, Finland. The results showed that a combination of correlogram and average mutual information (AMI) analysis might yield in the selection of the optimum lags that are needed to be used as the predictors of streamflow models. With Nash-Sutcliffe efficiency values of more than 99%, both GP and SARIMA models exhibited good performance for daily streamflow prediction. However, they were not able to precisely model the intramonthly snow water equivalent in the long-term forecast. The proposed ensemble model, which integrates the best GP and SARIMA models with the most efficient predictor, may eliminate one-fourth of root mean squared errors of standalone models. The GP-SARIMA also showed up to three times improvement in the accuracy of the standalone models based on the Nash-Sutcliffe efficiency measure.

Keywords Discharge · Genetic programming · SARIMA · Time series modeling · Forecasting

1 Introduction

Predicting floods and streamflow, in general, is one of the most critical tasks of hydrological modeling. This is quite a difficult modeling task, owing to the highly nonlinear, time- and spatially varying nature of the underlying process (Cheng et al. 2020). In addition, it is time-consuming and costly to measure the processes that affect streamflow, particularly in tributaries and snow-fed rivers, which means that the use of remotely sensed data is inevitable for accurate forecasts (Yang et al. 2007). The available data is often noisy, incomplete, or entirely missing. In addition, there is often an urgent need for high-quality modeling results

(Havlíček et al. 2013). Hence, it is more satisfactory to use univariate artificial intelligence (AI) techniques in which the preceding streamflow records are merely used to construct a predictive model (Zhang et al. 2018).

In recent decades, there has been considerable research on the use of AI techniques such as artificial neural networks (ANNs), extreme learning machine, and support vector machines to develop predictive models and identify the underlying hydrological pattern amongst a set of empirically observed variables (Govindaraju 2000; Raghavendra and Deka 2014; Yaseen et al. 2019; Boucher et al. 2020). Although the task is known as *system identification*, modelers have failed to discover a physically interpretable relationship for the desired phenomenon in many cases. This is mainly due to the black-box characteristics of most of the AI techniques which may model the process through implicit networks of data and parameters. To tackle the problem, recent studies have recommended gray-box techniques such as genetic programming (GP) (Giustolisi 2004; Nourani et al. 2014; Bozorg-Haddad et al. 2017; Herath et al. 2021).

✉ Ali Danandeh Mehr
ali.danandehmehr@oulu.fi

¹ Water, Energy and Environmental Engineering Research Unit, University of Oulu, FI90014 Oulu, Finland

² Department of Civil Engineering, Antalya Bilim University, Antalya, Turkey

GP is an emerging AI technique that applies evolutionary algorithms to identify explicit relationships for a given process (Koza 1992). It has different variants including (but not limited to) monolithic GP, gene expression programming, linear GP, multistage, and multigene GP. In all types, a population of random solutions (programs) is formed at the outset and then, the genetic items of each program are progressively changed to achieve the desired solution. Hydrologists have frequently used GP as a symbolic regression tool (Danandeh Mehr et al. 2018; Mohammad-Azari et al. 2020). Examples include the use of GP for rainfall-runoff modeling (Babovic and Keijzer 2002; Havlíček et al. 2013), groundwater simulation (Fallah-Mehdipour et al. 2014), forecasting meteorological variables (Kisi and Shiri 2011; Citakoglu et al. 2020), water quality prediction (Bozorg-Haddad et al. 2017), soil temperature modeling (Kisi et al. 2017), and spatial distribution of flow depth in fluvial rivers (Yan et al. 2021).

The current GP literature shows that several studies have also attempted to apply GP for univariate streamflow forecasting (e.g., Sivapragasam et al. 2008; Guven 2009; Wang et al. 2009; Al-Juboori and Guven 2016; Danandeh Mehr and Demirel 2016). Overall, its ability to extract explicit formulas has been reported as its main advantage over other AI techniques (Karimi et al. 2016, 2019; Herath et al. 2021). However, it may fail to model the streamflow process, particularly in long-term forecasts. To increase the predictive accuracy of GP, the most recent studies suggest hybrid GP models that can better tackle nonstationary features of streamflow time series (Danandeh Mehr et al. 2018). The key objective of the present study is therefore to improve the efficiency of univariate streamflow forecasting models through introducing a new hybrid GP model. In this study, we, first, developed a set of GP models for one-step-ahead streamflow forecasting in a lake-river system in cold climate conditions for a catchment in North-Eastern Finland. The models cover both short- (daily) and long-term (weekly and monthly) forecasting horizons and were compared with seasonal autoregressive integrated moving average (SARIMA) models developed as the benchmark. A new ensemble model, called GP-SARIMA, is additionally introduced to enhance the predictive accuracy of the standalone models for monthly streamflow forecasting. To select effective predictors, the study benefits from both autocorrelation and average mutual information (AMI) techniques.

1.1 Main contributions

The primary contributions are twofold. First, this study, for the first time, investigates the predictive capabilities of GP and SARIMA models for streamflow forecasting in a boreal lake-river system. Next, a new ensemble evolutionary model, called GP-SARIMA, is proposed for monthly streamflow

forecasting that is superior to standalone GP and SARIMA models and meets both accuracy and simplicity conditions. Compared to metaheuristic optimized AI models existing in the literature, the proposed model is explicitly having a less computational burden that makes it more appropriate to be implemented in practice.

2 Methodology

2.1 Overview of GP

GP is an evolutionary modeling approach in which random computer programs are created and improved to solve a given problem (Koza 1992). The computer programs have a tree structure comprising a root/function node, inner nodes, branches, and terminal nodes (leaves). Figure 1 demonstrates a GP tree and the associated mathematical expression. The main steps required to develop a GP-based forecasting model include (i) selection of input variables, (ii) educated guess about modeling functions (mathematical or Boolean), and (iii) appropriate tuning of evolutionary operators (Hrnjica and Danandeh Mehr 2019). Skilled decision-making during these steps helps the GP algorithm to evolve precise models and decrease the time of computations.

Regardless of the kind of problem, the GP algorithm starts with the random establishment of the initial programs known as potential solutions. At that point, the programs are sorted based on their goodness of fitness, and the ones demonstrating higher suitability are chosen as parents subjected to the evolutionary operations of crossover and mutation (Koza 1992). During crossover, two top parents combine their branches and create two offspring that may show higher fitness than their parents. In mutation operation, only a single parent is chosen, and an offspring is created by substituting some of its genetic materials with the new materials. Among the initial parents, the individual(s) showing the highest fitness is transferred directly to the new set of programs. The modeler defines the probability of GP

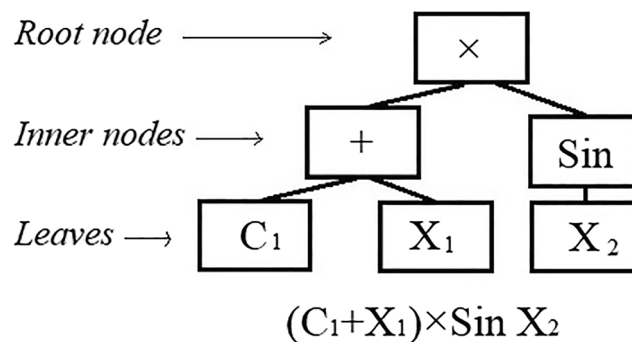


Fig. 1 An exemplary genome and its mathematical expression

operations. For a symbolic regression task, as is the case in this study, a high crossover rate is generally selected so that it is substantially greater than the mutation and reproduction rates. Since there is no universal way to determine these rates, one may use a trial-and-error procedure to optimize their values. For details about the GP algorithm and its applications in hydrology, the reader is referred to Danandeh Mehr et al. (2018).

2.2 Overview of SARIMA

Classic autoregressive time series modeling techniques such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and seasonal autoregressive integrated moving average (SARIMA) could be used as alternatives for univariate streamflow modeling (Terzi and Ergin 2014; Valipour 2015; Mehdizadeh and Sales 2018). The pertinent literature shows SARIMA outweighs its counterparts as it can handle both potential trend and periodicity features in streamflow series (Moeni et al. 2017). However, its performance is highly sensitive to selecting a correct periodic term in the model calibration stage.

A SARIMA model is structured by combining additional seasonal terms into ARIMA structure (Box et al. 2015). The model is commonly expressed as SARIMA(p, d, q)(P, D, Q) $_m$ in which $p, d,$ and q are non-seasonal components; $P, D,$ and Q are the seasonal backshifts and the letter m denotes the number of samples in a year (e.g., $m = 12$ for monthly data). Equation (1) expresses an example of a first-order SARIMA model without a constant for a set of quarterly data (i.e., $m = 4$).

$$SARIMA(1,1,1)(1,1,1)_4 : (1 - \phi_1 B)(1 - \Phi_1 B^4)(1 - B)(1 - B^4) \tag{1}$$

$$y_t = (1 + \theta_1 B)(1 + \Theta_1 B^4)\epsilon_t$$

where ϕ_1 and θ_1 are the parameters of non-seasonal and Φ_1 and Θ_1 are the parameters of seasonal components of the model. The term ϵ_t is white noise (Bender and Simonovic 1994).

It is seen that the additional seasonal terms are simply multiplied by the non-seasonal terms. Like ARIMA modeling, the seasonal backshift parameters (Φ and Θ) can be determined through either a correlogram analysis or from an analytical stationarity test such as augmented Dickey-Fuller (ADF). For more details about parameter tuning in SARIMA, the interested reader is referred to Bender and Simonovic (1994).

2.3 The proposed evolutionary GP-SARIMA model

Predictive performance is of the utmost importance to a hydrological model. Ensemble learning algorithms

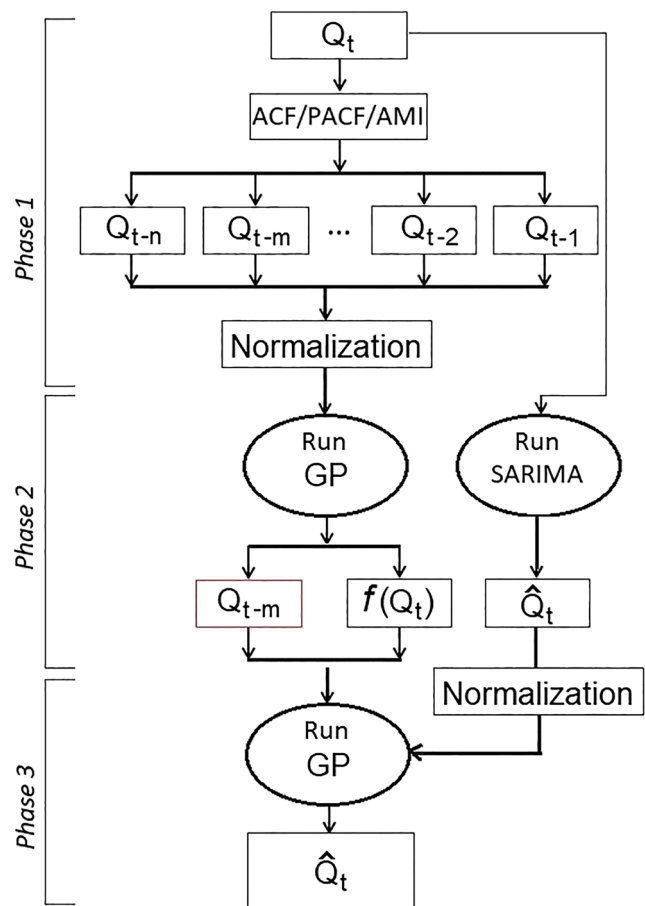


Fig. 2 Flowchart of the proposed GP-SARIMA streamflow forecasting model

typically combine the forecasts from multiple models and are designed to outweigh any contributing ensemble member. Applications of different types of ensemble AI techniques in hydrology have been recently reviewed by Zounemat-Kermani et al. (2021). This study introduces the ensemble GP-SARIMA model in which the prediction process is composed of three main stages (Fig. 2). The correlogram and mutual information analysis are implemented in the first stage to determine the potential lags (predictors). Then, both target and input vectors are normalized to secure the development of dimensionally accurate solutions. In the second stage, the ad hoc modeling phase, the SARIMA, and GP techniques are run to evolve initial solutions. To this end, Gretl and GPdotNET v5.0 (Hrnjica and Danandeh Mehr 2019) tools can be used, respectively. In this phase, the most effective input (Q_{t-m}) is determined concerning the average impact of each input in the best solutions as suggested by Uyumaz et al. (2014). The GP result in this phase is a dimensionless vector of streamflow, but the SARIMA forecasts streamflow with the same dimension of the input series (here m^3/s). In the last phase, the GP engine is rerun so that the

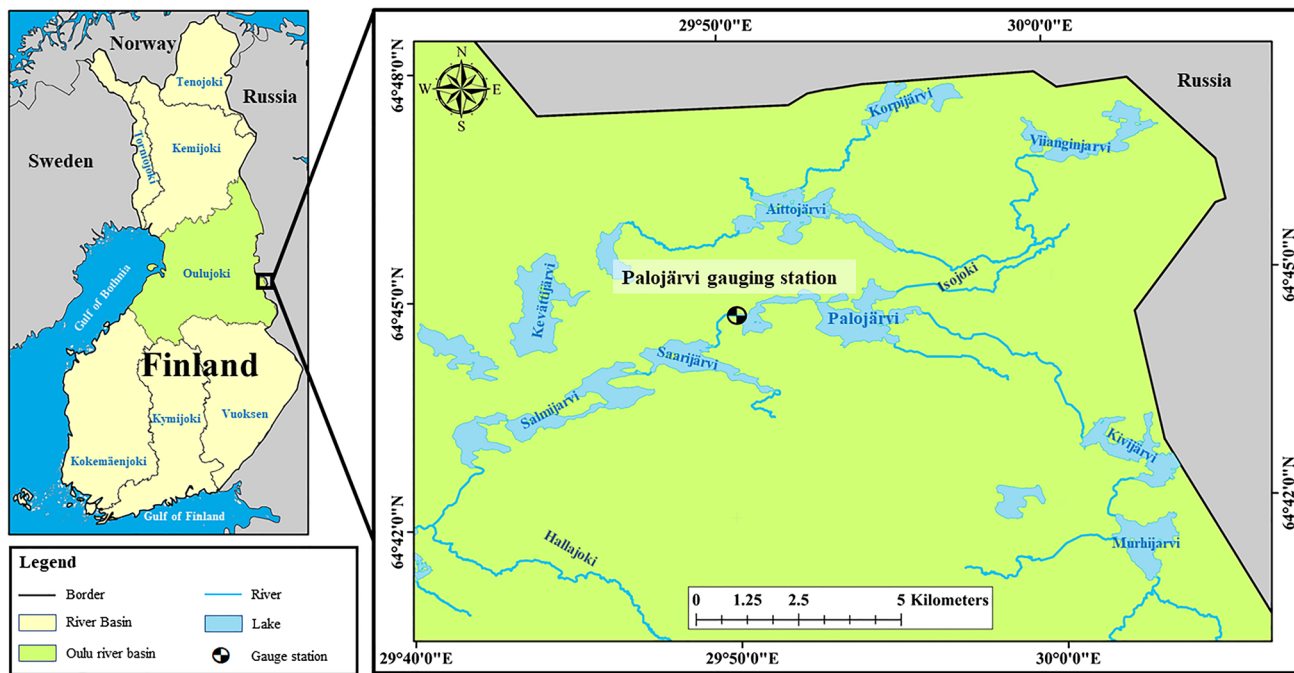


Fig. 3 Study area and location of Palojärvi gauging station

most influential input, the initial GP forecast, and the normalized SARIMA forecast are used as the new predictors. This acts as post-processing to reduce the initial models’ errors. As both GP and SARIMA are explicit models, the ensemble model remains explicit; however, the results are dimensionless and need to be denormalized. Compared to hybrid simulation-metaheuristic optimization models (e.g., Yaseen et al. 2017), the new model has a less computational burden. This makes it faster than metaheuristic optimization models. However, it increases the likelihood of trapping GP in local optima, and thus, the modeler needs to control the model against the overfitting problem.

2.4 Average mutual information (AMI)

A set of optimal time delays (lags) leads a predictive AI model to a robust solution. On the contrary, inefficient or redundant lags may result in poor or complex models. Many optimal lag selection methods fail to perform properly owing to the inherent hypothesis of linearity or intense redundancy between the lags (Darudi et al. 2013). In previous studies, autocorrelation analysis of the streamflow series has been commonly employed to identify the optimum lags (e.g., Rezaie-Balf et al. 2019). However, the information distilled from autocorrelation analysis merely represents collinearity among the current and preceding discharge amount. Thus, the method might fail to extract

efficient inputs in a nonlinear process (Danandeh Mehr and Gandomi 2021). To cope with this drawback, the average mutual information (AMI) that could be judged as a nonlinear generalization of the autocorrelation function was additionally considered in this study. This criterion (Eq. 2), aka auto mutual information, is generally used to find time delayed coordinates that are as independent of each other as possible (Fraser and Swinney 1986).

$$AMI(Q_t, Q_{t+\tau}) = \sum_{i,j} P_{i,j}(\tau) \log\left(\frac{P_{i,j}(\tau)}{P_i P_j}\right) \tag{2}$$

where P_i is the probability of Q_t in bin i of the histogram constructed from the data points and $P_{i,j}(\tau)$ is the probability that Q_t is in bin i and $Q_{t+\tau}$ is in bin j . As merely the joint probability $P_{i,j}(\tau)$ depends on τ , and thus, the AMI function also depends on how the histograms are constructed, i.e., the width and position of the bins.

3 Study area and data

Essential to the fulfillment of a hydrological model is its stochastic feature. Construction of regulators such as a dam is just one way to lose the stochasticity of streamflow. Therefore, the implementation of the new models should be assessed on a case-by-case basis. When deciding on a catchment, it should be taken into account that flow measurements

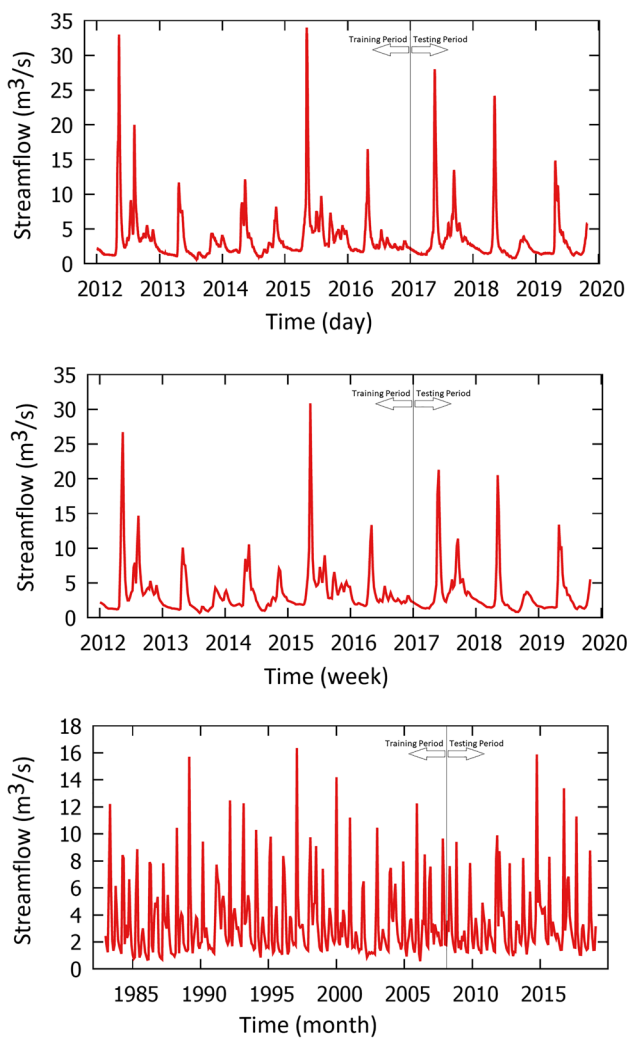


Fig. 4 Observed Streamflow at Palojärvi gauging station

are not regulated or adjusted before or at the location of a stream gauge. Accordingly, implementation of the GP and SARIMA for daily, weekly, and monthly univariate streamflow modeling was carried out using observations from the Palojärvi gauging station located on the unbuilt headwater of the Oulujoki River system in North-Eastern Finland (Fig. 3). The region is northern boreal with seasonal snow and soil frost and does not contain any glaciers or permafrost. Oulujoki river catchment is strongly seasonally affected. Limited baseflow during winter months and spring floods during the snow melting period at April-May, and summer baseflow during July-August. Having a length of about 107 km, the Oulujoki is one of the largest lake-river systems (catchment area = 22,500 km²) in Finland (Salojärvi et al. 1982). At the point of Palojärvi gauging station (7,186,061 N, 3,635,282 E), the catchment area is about 264 km². The upstream area from the gauging station has a high lake/pond percentage that significantly affects the runoff regime. Daily streamflow

data at the station is recorded since 1983 by Finnish Environment Institute (SYKE), and data is openly available at the national OIVA-database.

The observed daily, weekly, and monthly streamflow hydrographs used in this study are shown in Fig. 4. Of the total observations, the first 70% and last 30% were used to train and test the evolved models. Table 1 represents the associated statistical features. Prior to importing the datasets to the GP engine, we normalized the predictor/target vectors so that they are within the range of 0.0 to 1.0 (i.e., min-max normalization approach).

4 Criteria for performance appraisal

Numerical metrics utilized to evaluate models' performance have been reviewed by Biondi et al. (2012). A combination of absolute value error and normalized goodness-of-fit statistics is currently recommended to assess hydrological models (Ritter and Munoz-Carpena 2013). Therefore, in addition to graphical results, we implement root mean square error (RMSE) as an absolute error statistic and Nash-Sutcliffe efficiency (NSE) as a normalized efficiency value in this study. Mathematical expressions of the indices are presented below:

$$NSE = 1 - \frac{\sum_{i=1}^n (X_i^{obs} - X_i^{pre})^2}{\sum_{i=1}^n (X_i^{obs} - X_{mean}^{obs})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (X_i^{obs} - X_i^{pre})^2}{n}} \tag{4}$$

where X_i^{obs} = observed streamflow at the time i , X_i^{pre} = predicted streamflow at the time i , X_{mean}^{obs} = mean observed streamflow, and n is the number of arrays at each vector.

5 Results and discussion

As previously mentioned, effective lags were selected with respect to both linear and nonlinear correlations throughout the correlogram and mutual information analysis. To this end, we calculated AMI measure for the target streamflow series. At first, the joint likelihood between the observed discharge at time t (Q_t) and its preceding 62 steps (i.e., $Q_{t-\tau}$, $\tau = 1, 2, \dots, 62$) was calculated. Then, the associated AMI values were attained using Eq. (2). Figure 5 illustrates the AMI values attained for the observed daily, weekly, and monthly streamflow series. Overall, the figure demonstrates that the AMI values generally decrease by increasing the number of lags. It contains an oscillating pattern at weekly and monthly time scales. Regarding the daily and weekly timeseries, the greatest AMI value

Table 1 Statistical features of the observed streamflow in Palojärvi gauging station

Statistic	data set	Time series type		
		Daily Mean	Weekly Mean	Monthly Mean
Modeling period	Entire	2012 to 2019	2012 to 2019	1983 to 2019
	Training	2012 to 2016	2012 to 2016	1983 to 2007
	Testing	2017 to 2019	2017 to 2019	2008 to 2019
Number of observations	Entire	2861	409	434
	Training	1827	288	300
	Testing	1034	121	134
Max (m^3s^{-1})	Entire	34	30.86	16.35
	Training	34	30.86	16.35
	Testing	28	20.53	15.88
Ave (m^3s^{-1})	Entire	3.53	3.53	3.36
	Training	3.65	3.67	3.38
	Testing	3.33	3.22	3.32
Min (m^3s^{-1})	Entire	0.63	0.69	0.57
	Training	0.63	0.69	0.57
	Testing	0.79	0.83	0.96
Standard deviation	Entire	3.56	3.44	2.66
	Training	3.57	3.62	2.73
	Testing	3.55	2.94	2.51

stood at the first lag; however, at the monthly scale, the maximum AMI was attained at lag #12. This implies that monthly streamflow in the river relies heavily on the past year's value than that of the previous months. Considering the AMI threshold of 0.25, 0.1, and 0.035 as well as the attained correlogram (see Fig. 6), the most effective inputs for one-step-ahead daily, weekly, and monthly univariate streamflow forecasting are shown in Eqs. (5) to (7).

$$Q_{td} = f(Q_{td-1}, Q_{td-2}, Q_{td-3}, Q_{td-4}, Q_{td-5}) \quad (5)$$

$$Q_{tw} = f(Q_{tw-1}, Q_{tw-38}, Q_{tw-39}, Q_{tw-40}, Q_{tw-41}) \quad (6)$$

$$Q_{tm} = f(Q_{tm-1}, Q_{tm-2}, Q_{tm-4}, Q_{tm-12}, Q_{tm-32}, Q_{tm-33}) \quad (7)$$

where Q_{td} , Q_{tw} , and Q_{tm} denote mean daily, mean weekly, and mean monthly streamflow, respectively.

5.1 Results of standalone GP and SARIMA models

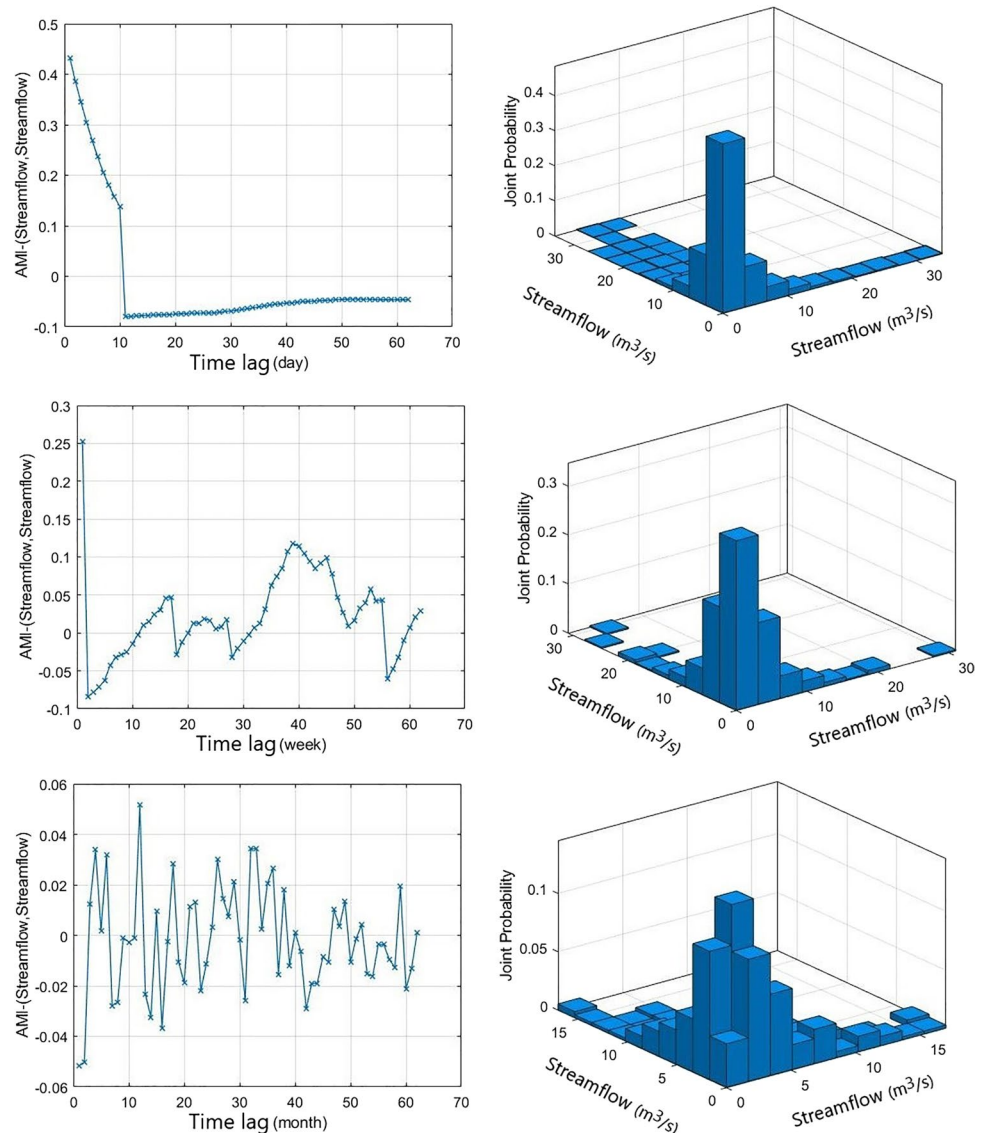
As illustrated in Fig. 2, the efficient streamflow vectors (shown in Eqs. (5) to (7)) were imported as the inputs for the GP engine. Apart from input/target vectors, the modeler needs to define a set of functions, random numbers, aka floating-point, and rates of evolutionary operators to run GP. Considering the given time scales, the evolutionary algorithm can generate various formulae representing the lake-river's streamflow time series. Here, we employed the basic arithmetic (+, /, ×, and -), trigonometric, and exponential

functions. Table 2 summarizes the setup features for GPdotNET v5.0, a non-commercial GP tool. It is worth mentioning that the main evolutionary parameters (i.e., crossover, mutation rate, and reproduction) were optimized through a trial-and-error strategy.

To cope with the overfitting problem in the GP, we ran the GP with lower trees at the first trials and then, linearly increased the maximum tree depth up to six (see Table 2). Meanwhile, the mean fitness value throughout gene productions was checked to end the run. This is a kind of supervised control in which the evolutionary process is ended once either a weaker solution is created, or the number of generations passes a user-defined maximum number of generations.

The mathematical expressions of the best GP solutions are tabulated in Table 3. It is clear from the table that the best model does not necessarily comprise all the predefined effective lags and functions. For instance, the best daily model was attained via a nonlinear combination of the first two lags although up to five lags were considered as potential input vectors. Similarly, the weekly model indicates that the first and 38th lags are more informative inputs among those given in Eq. (6). This is due to interior evolutionary function optimization of the GP algorithm that allows it to optimize its shape by eliminating less efficient inputs/functions existing in the user-specified search space at each time scale. Considering the monthly time scale, GP produces the most complex model (in terms of both numbers of inputs and functional nodes) that utilizes all the predefined input vectors. The flexibility of GP structure against the given inputs

Fig. 5 The AMI (left panels) and joint probability (right panels) amongst the observed streamflow data at the study site



or functions is one of its advantages over SARIMA which has a fixed structure.

To attain the best SARIMA models at each time scale, the first step is to estimate the order of autoregressive, moving average, and integration components. To this end, the ADF test and a visual inspection of correlograms (see Fig. 6) were respectively utilized in this study. The p values of the ADF test (see Table 4) less than 5% implied that the observed streamflow hydrographs could be regarded as stationary series.

In Figs. 6a and b, the sudden drops in the first lag of the partial autocorrelation functions indicate the insignificant correlation after the first lag. Thus, a seasonal autoregressive process of order one and period one could be considered. Figure 5c exhibits the highest strength of the serial correlations at lag 12. Following Danandeh Mehr and Gandomi (2021), multiple combinations

of seasonal (p , d , q) and non-seasonal parameters (P , D , Q) were tested in this study to select the best SARIMA model. The model which shows the smallest corrected Akaike information criterion (AIC_c) in the training period and RMSE in the testing period was selected as the best solution. Table 4 summarizes some of the best SARIMA combination trials.

According to the results, the SARIMA (1,1,1)(1,0,1), SARIMA (1,0,1)(0,0,0), and SARIMA (2,0,2)(2,0,2) are respectively the best autoregressive models for daily, weekly, and monthly streamflow forecasting for the study site. The results of the weekly scale indicated that the best performing SARIMA model has no seasonal component and the order of integration of the non-seasonal portion is equal to zero which means ARMA could effortlessly model weekly mean streamflow series.

Fig. 6 Autocorrelation function (ACF) and partial autocorrelation functions (PACF) of the observed **a** daily, **b** weekly, and **c** monthly streamflow in the study site

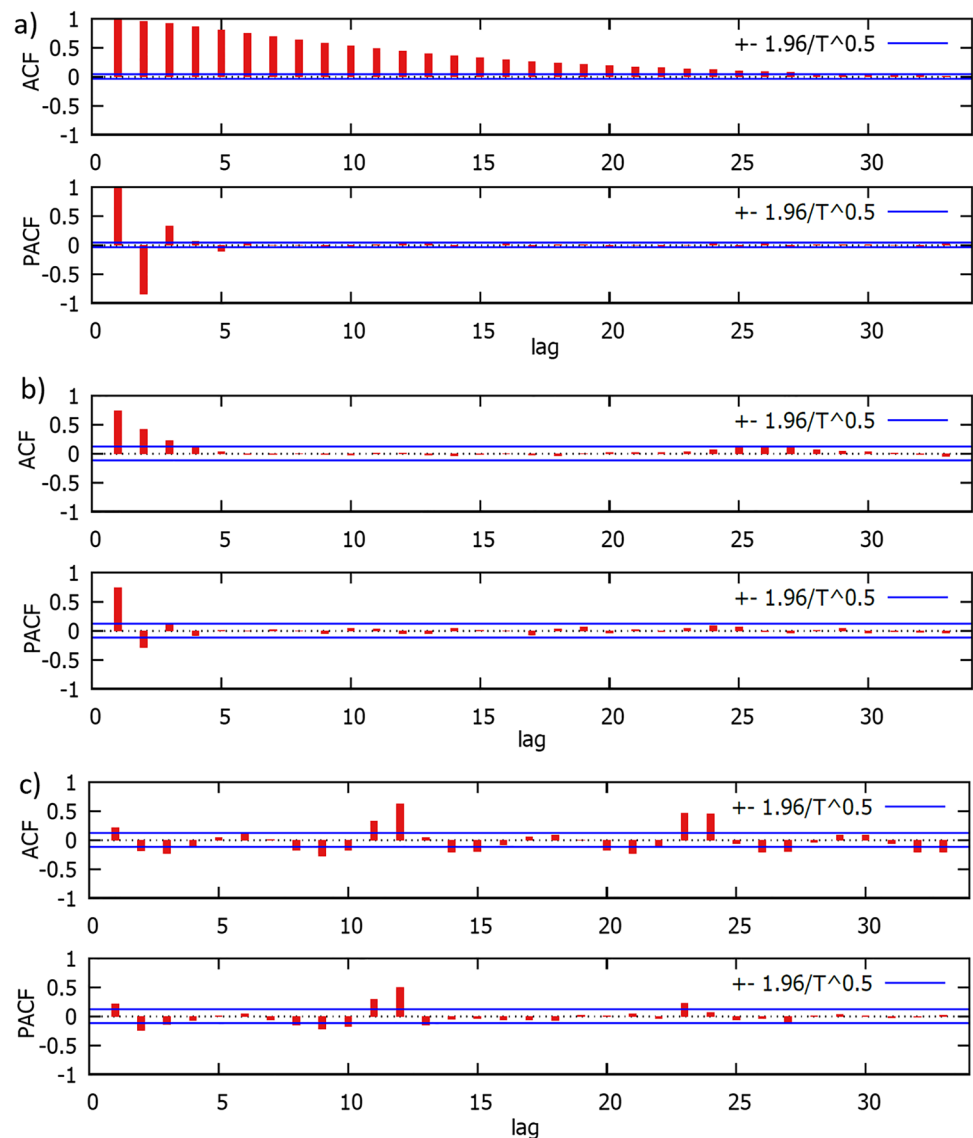


Table 2 Parameters used to run GPdotNET

Parameter	Value
Population number	500
Initialization method	Half-half
Elitism	1
Initial level	3
Tree depth (maximum)	6
Selection method	Rank (0.8)
Crossover	0.9
Mutation	0.05
objective function	RMSE
Data scaling	Min-max normalization
Range of floating points	0 and 1

For performance appraisal, the best GP and SARIMA models' hydrographs are depicted in Fig. 7, and the associated goodness-of-fit values are tabulated in Table 5. From Fig. 7, it is seen that both GP and SARIMA precisely capture the oscillating regime of the observed daily flow in the snow-dominated lake-river system. According to the goodness-of-fit results in the daily model, both the GP and SARIMA offer similar predictive accuracy ($NSE = 0.997$) with the lower error between the model and observed data in the testing period. Comparing to the results of a similar study (see Abdollahi et al. 2017) that applies GP to model daily streamflow in a hot climate ($NSE = 0.94$), our findings indicate that GP (and even SARIMA) exhibits better performance ($NSE = 0.99$) in cold climates.

At the weekly time scale, global peak and local maxima were better forecasted by the SARIMA. In contrast, GP exhibited higher efficiency in tracing peak monthly streamflow values.

Table 3 GP models evolved for one-step-ahead streamflow forecasting for the study site

Time scale	Model
Daily	$Q_{td} = Q_{td-1} / (\exp(\frac{0.6384}{Q_{td-2}}) \times (Q_{td-2} - Q_{td-1}))$
Weekly	$Q_{tw} = Q_{tw-1} / (\cos(\cos(\frac{0.6245}{Q_{tw-38}}))) + Q_{tw-1}$
Monthly	$Q_{tm} = (0.558 / ((0.683 + ((Q_{tm-12} / 0.683) \times (Q_{tm-2} + Q_{tm-1}))) \times (0.683 + ((Q_{tm-12} + Q_{tm-32}) \times (Q_{tm-2} + Q_{tm-4})))) / (((0.683 \times Q_{tm-4}) \times (Q_{tm-1} - Q_{tm-12})) \times ((0.403 - Q_{t-12}) \times (0.558 / Q_{tm-32}))) + (((Q_{tm-33} \times Q_{tm-1}) \times 0.797) - (Q_{tm-1} - (Q_{tm-1} + Q_{tm-12}))))$

Table 4 Performance appraisal of the evolved SARIMA models using observations at the training set

Time scale	ADF test <i>p</i> value	Model structure	AIC _c	Forecasting period*	
				RMSE (m ³ s ⁻¹)	MAE** (m ³ s ⁻¹)
Daily	2.56 × 10 ⁻¹³	SARIM (1,0,1)(0,0,0)	1935.5	0.248	0.101
		SARIMA (1,0,1)(1,0,1)	1883.1	0.243	0.103
		SARIMA (1,1,1)(0,0,0)	881.6	0.156	0.066
		SARIMA (1,1,1)(1,0,1)	879.0	0.155	0.067
		SARIMA (1,1,1)(1,1,1)	912.5	0.156	0.070
Weekly	3.05 × 10 ⁻⁷	SARIMA (1,0,1)(0,0,0)	1108.6	1.752	0.907
		SARIMA (1,0,1)(2,0,2)	1114.6	1.965	0.915
		SARIMA (1,2,1)(0,0,0)	1161.6	1.998	0.830
		SARIMA (1,2,1)(2,0,2)	1166.4	2.517	1.066
		SARIMA (1,0,1)(2,2,2)	876.9	3.875	1.984
monthly	0.0212	SARIMA (1,0,1)(0,0,0)	1342.2	2.473	1.643
		SARIMA (1,0,1)(2,0,2)	1218.2	2.477	1.546
		SARIMA (1,2,1)(2,0,2)	1320.2	2.871	1.699
		SARIMA (0,0,0)(2,0,2)	1190.5	2.755	1.691
		SARIMA (2,0,2)(2,0,2)	1198.5	2.350	1.449

* Error measures were calculated using the last 100 observations of training set

** Mean absolute error

According to the fitness criteria, the GP identifies the weekly and monthly flow process better than SARIMA during the training period; however, they suffer from low efficiency in the testing period. Regarding the higher performance of the SARIMA during the testing period, it must be highlighted that such results could be due to the relatively lower number of testing observations and the higher variance of the training observations that include the global maximum.

5.2 Results of the proposed GP-SARIMA model for long-term streamflow forecasting

Following the methodology flowchart, the SARIMA (1,1,1)(2,1,0), the best GP outputs, and the most effective lag (i.e., Q_{t-12}) were utilized as predictors for the rolling forecast of monthly streamflow using GP-SARIMA. Running this model through the same training (testing) period, we get an NSE of 0.715 (0.483) and RMSE equals 1.437 (1.817) m³/s. In comparison to the goodness-of-fit values given in Table 5, GP-SARIMA is superior to the standalone GP and SARIMA models. Compared to the best SARIMA, the proposed ensemble model yielded

an approximately 25 and 20% reduction in RMSE in the training and testing periods, respectively. Cross-correlation analysis between the new predictors (best GP and SARIMA models) and the target streamflow series showed that they have a higher correlation (0.67 and 0.63 for GP and SARIMA, respectively) than standalone models' inputs (maximum of 0.56 for Q_{t-12}). Since all these models generally utilize autocorrelation of the time series, the mentioned higher correlation could be considered as the origin of the improvement of the performance of the ensemble model compared to the standalone models. The observed and forecasted monthly streamflow series and the associated scatter plots during the training and testing periods were depicted and compared in Fig. 7.

Overall, the forecasts mimic the strong fluctuation of the observed streamflow series even though significant errors are observed in the prediction of some of the peaks throughout the year. It is seen that the SARIMA suffers from low-variance forecasts unable to capture peak flows. This result agrees with that of Danandeh Mehr and Gandomi (2021) in which SARIMA was applied to model the Sedre River flow in Turkey. Such drawback at

Fig. 7 The observed compared with modeled streamflow data during both training and testing periods

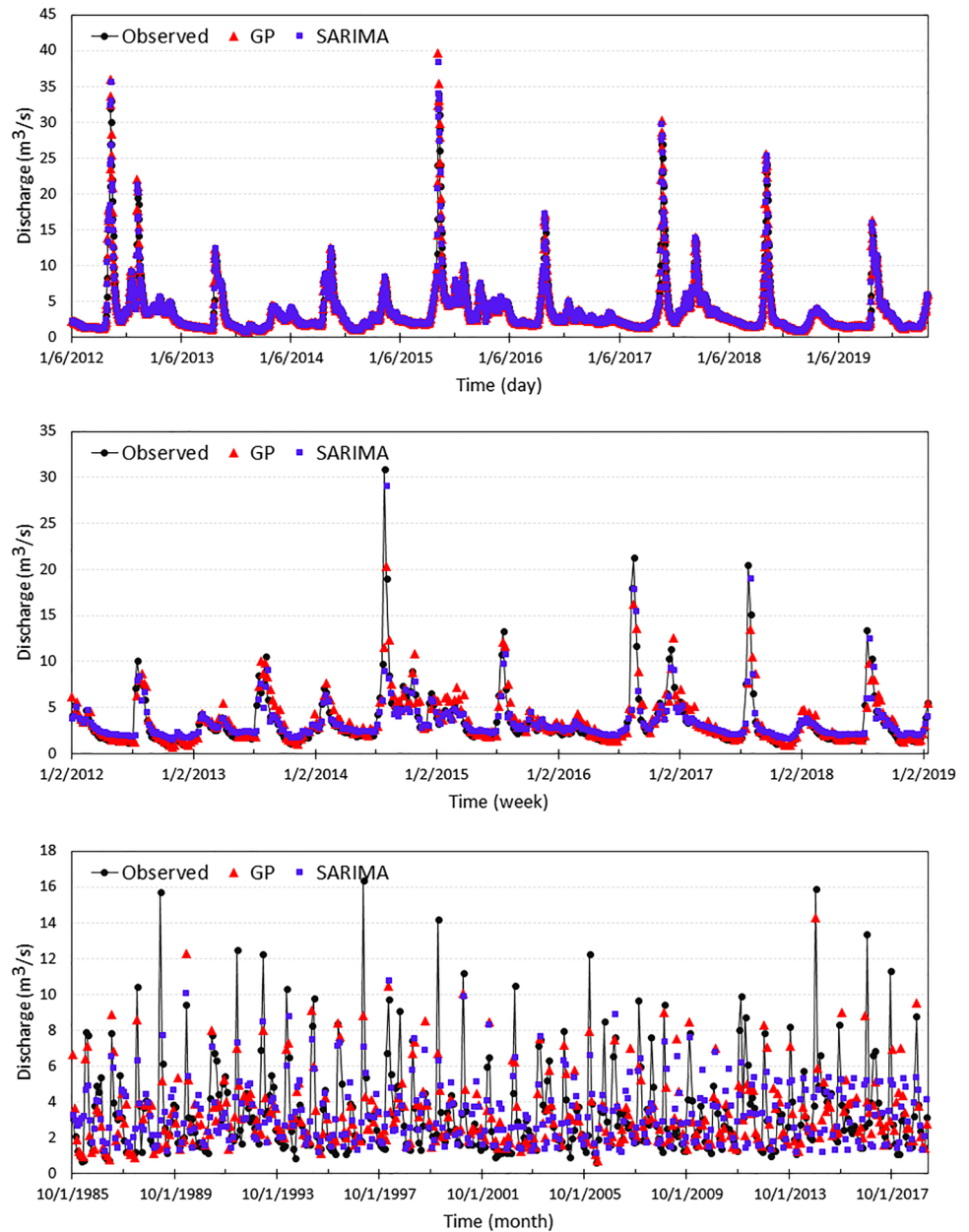


Table 5 Performance values of the best GP and SARIMA models used for univariate streamflow forecasting in the study site

Time scale	Model	RMSE (m^3s^{-1})		NSE	
		Training	Testing	Training	Testing
Daily	GP	0.327	0.188	0.992	0.996
	SARIMA	0.300	0.155	0.994	0.997
Weekly	GP	2.000	1.915	0.622	0.552
	SARIMA	2.038	1.752	0.608	0.625
Monthly	GP	1.859	2.408	0.524	0.093
	SARIMA	1.935	2.288	0.484	0.181

the estimation of high discharge values might be due to the existence of strong deviation during the snow melting months so that the linear SARIMA cannot capture it. For such months during the testing period, Fig. 8c demonstrates that the SARIMA forecasts are converged to a false local maximum of around $6.0 \text{ m}^3/\text{s}$. By contrast, the GP, and in particular, the ensemble GP-SARIMA were not trapped into such maxima. Although GP-SARIMA considerably diminishes the residuals, it still underestimates the observed high flows. Such difference could be due to intramonthly accumulated snow water equivalent that is difficult to be distilled from historical streamflow data using univariate AI models. Therefore, the use of

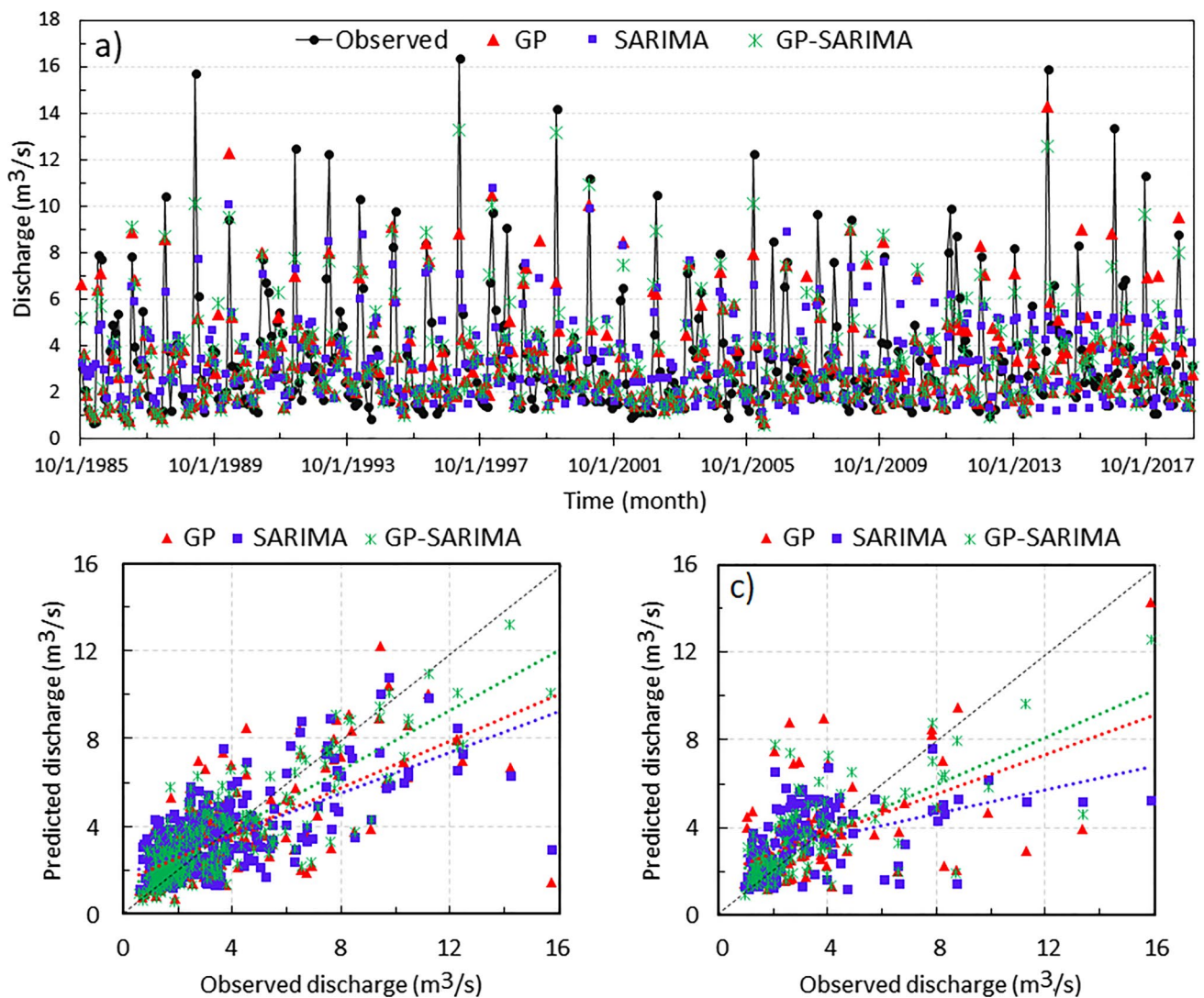


Fig. 8 Monthly streamflow hydrograph (a) and the associated scatter plots in the **b** training and **c** testing periods

exogenous inputs such as snow cover extent or depth for long-term streamflow forecasting in snow-fed rivers is recommended.

To explore the contribution of each input in the best evolved GP-SARIMA model, its tree expression is shown in Fig. 9. In this model, Q_{tm-12} , GP, and SARIMA are the normalized values of the 12-month antecedent observed streamflow, concurrent GP, and SARIMA forecasts, respectively. The constant 1.11 represents the summation of two random floating points (0.47 and 0.64) attained in the terminal nodes of the raw GP-SARIMA model. It is seen that the Q_{tm-12} and GP solution appeared two times in this model. Therefore, they could be counted as the most dominant variables among the pre-specified predictors.

6 Conclusion

Many studies have proved that AI techniques outperform classical time series models for streamflow forecasting. While tackling nonstationary features of a given time series is the utmost important issue in univariate streamflow modeling using classical autoregressive models, selecting a suitable AI technique, finding the more efficient inputs (i.e., lags), and being heedful of the common overfitting challenge are some of the critical concerns that a modeler should contemplate in time series modeling using AI techniques (Thapa et al. 2020). In this study, the abilities of GP and SARIMA for one-step-ahead daily, weekly, and monthly streamflow forecasting in the headwaters of the Oulujoki River were investigated. The comparative performance appraisal of the

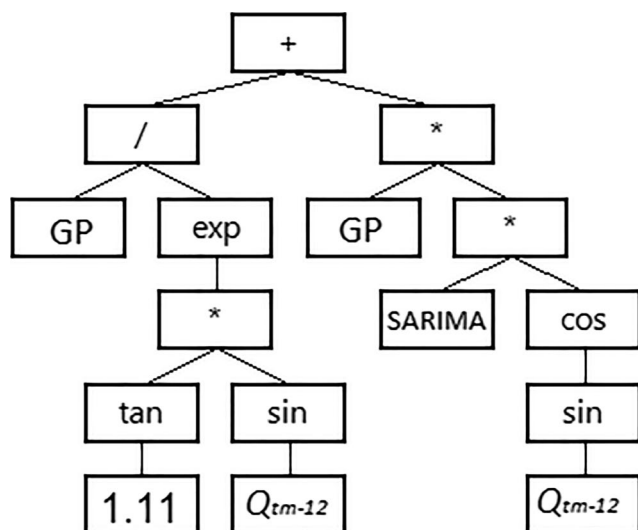


Fig. 9 The GP-SARIMA model evolved for univariate streamflow forecasting in the study site

models showed a good and more or less the same accuracy for both GP and SARIMA models in daily streamflow forecasting. The techniques also showed acceptable performance for weekly streamflow forecasting with a slight superiority of SARIMA over GP during the testing period. Our results also revealed that the standalone techniques are not suitable for monthly streamflow modeling in the case study lake-river system. This drawback was attributed to the effect of snow melting during spring months that creates extreme oscillating structure in the monthly streamflow hydrograph so that the models, particularly SARIMA, cannot model the streamflow series 1 month in advance. Consequently, they underestimate maxima/peak streamflow in the study site. To enhance the prediction accuracy at the monthly time scale, an ensemble univariate GP-SARIMA model was introduced. The associated results demonstrated a significant improvement in the predictive accuracy of the GP and SARIMA. Therefore, we concluded that the ensemble strategy is more robust than standalone methods, being less likely to be trapped into a false local maximum.

From an evolutionary programming perspective, the GPdotNET running experiences indicated that using more complex functions or deep GP trees would not essentially enhance the models' accuracy. Contrariwise, it may lead the algorithm to be trapped in a local optimum (i.e., over-fitted solution) during the initial generations. Benefiting from the function optimization feature, GP yielded in simpler models than SARIMA, particularly for short-term forecasting. Although the GP-SARIMA model was superior to GP and SARIMA, the attained results emphasize the necessity of applying further efforts to improve model accuracy in long-term forecasting.

The present study was limited to the use of (i) univariate modeling strategy (i.e., observed streamflow data is used as both input and target variables) and (ii) classic GP as the nonlinear regression technique. At monthly time-scale, the evolved models generally underestimated the streamflow hydrograph. This highlights the necessity for further investigations either on the modeling strategy or the training approach. For long-term (weekly, monthly, or seasonal) streamflow prediction in boreal rivers, one may investigate the use of multivariate models so that they would better reflect intramonthly snow water equivalent proportion (Yang et al. 2009). In addition, future studies can consider more robust GP variants such as multigene GP or multistage GP that might better handle streamflow series in boreal lake-river systems. It was well-documented that AI-based models are case-sensitive, and their generalization ability depends on the range and type of data used to train/test the desired model. Thus, it could be informative if the efficiency of GP-SARIMA to model intermittent rivers has been checked in a similar study.

Author contribution Ali Danandeh Mehr: conceptualization, methodology, formal analysis, validation, writing (original draft), writing (review and editing), and visualization. Sahand Ghadimi: writing, review and editing, and visualization. Hannu Marttila writing, review and editing; supervision. Ali Torabi Haghighi: conceptualization; writing, review and editing; supervision.

Funding Open Access funding provided by University of Oulu including Oulu University Hospital. The study was supported by the Hydro-RDI project funded by the Academy of Finland (decision number: 337523) and Maa- ja vesitekniiikan tuki ry (Reference Number: 41878).

Data availability The data used in this study is recorded by Finnish Environment Institute (SYKE), and data is openly available at the national OIVA-database.

Code availability The GPdotNET v5.0 package that support the findings of this study is available in <https://github.com/bhrnjica/gpdotnet>

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated

otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdollahi S, Raeisi J, Khalilianpour M, Ahmadi F, Kisi O (2017) Daily mean streamflow prediction in perennial and non-perennial rivers using four data driven techniques. *Water Resour Manage* 31(15):4855-4874
- Al-Juboori AM, Guven A (2016) A stepwise model to predict monthly streamflow. *J Hydrol* 543:283-292
- Babovic V, Keijzer M (2002) Rainfall runoff modelling based on genetic programming. *Hydrol Res* 33(5):331-346
- Bender M, Simonovic S (1994) Time-series modeling for long-range streamflow forecasting. *J Water Resour Plan Manag* 120(6):857-870
- Biondi D, Freni G, Iacobellis V, Mascaro G, Montanari A (2012) Validation of hydrological models: conceptual basis, methodological approaches and a proposal for a code of practice. *Physics and Chemistry of the Earth, Parts a/b/c* 42:70-76
- Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) *Time series analysis: forecasting and control*. John Wiley & Sons
- Boucher, M. A., Quilty, J., & Adamowski, J. (2020). Data assimilation for streamflow forecasting using extreme learning machines and multilayer perceptrons. *Water Resources Research*, 56(6), e2019WR026226.
- Bozorg-Haddad O, Soleimani S, Loáiciga HA (2017) Modeling water-quality parameters using genetic algorithm-least squares support vector regression and genetic programming. *J Environ Eng* 143(7):04017021
- Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., & Pain, C. C. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. *Journal of Hydrology*, 590, 125376.
- Citakoglu H, Babayigit B, Haktanir NA (2020) Solar radiation prediction using multi-gene genetic programming approach. *Theoret Appl Climatol* 142(3):885-897
- Danandeh Mehr A, Demirel MC (2016) On the calibration of multi-gene genetic programming to simulate low flows in the Moselle River. *Uludağ University Journal of the Faculty of Engineering* 21(2):365-376
- Danandeh Mehr A, Gandomi AH (2021) MSGP-LASSO: An improved multi-stage genetic programming model for streamflow prediction. *Inf Sci* 561:181-195. <https://doi.org/10.1016/j.ins.2021.02.011>
- Danandeh Mehr A, Nourani V, Kahya E, Hrnjica B, Sattar AM, Yaseen ZM (2018) Genetic programming in water resources engineering: a state-of-the-art review. *J Hydrol* 566:643-667
- Darudi, A., Rezaeifar, S., & Bayaz, M. H. J. D. (2013, November). Partial mutual information based algorithm for input variable selection For time series forecasting. In 2013 13th International Conference on Environment and Electrical Engineering (EEEIC) (pp. 313-318). IEEE.
- Fallah-Mehdipour E, Haddad OB, Marino MA (2014) Genetic programming in groundwater modeling. *J Hydrol Eng* 19(12):04014031
- Fraser AM, Swinney HL (1986) Independent coordinates for strange attractors from mutual information. *Phys Rev A* 33(2):1134. <https://doi.org/10.1103/PhysRevA.33.1134>
- Giustolisi O (2004) Using genetic programming to determine Chezy resistance coefficient in corrugated channels. *J Hydroinf* 6(3):157-173
- Govindaraju, R. S. (2000). Artificial neural networks in hydrology. II: hydrologic applications. *Journal of Hydrologic Engineering*, 5(2), 124-137.
- Guven A (2009) Linear genetic programming for time-series modeling of daily flow rate. *J Earth Syst Sci* 118(2):137-146
- Havlíček V, Hanel M, Máca P, Kuráz M, Pech P (2013) Incorporating basic hydrological concepts into genetic programming for rainfall-runoff forecasting. *Computing* 95(1):363-380
- Herath HMVV, Chadalawada J, Babovic V (2021) Genetic programming for hydrological applications: to model or forecast that is the question. *J Hydroinf* 23(4):740-763
- Hrnjica, B., Danandeh Mehr, A., 2019. Optimized genetic programming applications: emerging research and opportunities. IGI Global, Hershey, PA.
- Karimi S, Shiri J, Kişi Ö, Shiri AA (2016) Short-term and long-term streamflow prediction by using wavelet-gene expression programming approach. *ISH Journal of Hydraulic Engineering* 22(2):148-162
- Karimi B, Safari M, Mehr AD, Mohammadi M (2019) Monthly rainfall prediction using ARIMA and gene expression programming: a case study in Urmia, Iran. *Online J Eng Sci Technol* 2(3):8-14
- Kisi O, Sanikhani H, Cobaner M (2017) Soil temperature modeling at different depths using neuro-fuzzy, neural network, and genetic programming techniques. *Theoret Appl Climatol* 129(3):833-848
- Kisi O, Shiri J (2011) Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. *Water Resour Manage* 25(13):3135-3152
- Koza JR (1992) *Genetic programming: on the programming of computers by means of natural selection*, vol 1. MIT press, USA
- Mehdizadeh S, Sales AK (2018) A comparative study of autoregressive, autoregressive moving average, gene expression programming and Bayesian networks for estimating monthly streamflow. *Water Resour Manage* 32(9):3001-3022
- Moeeni H, Bonakdari H, Ebtehaj I (2017) Monthly reservoir inflow forecasting using a new hybrid SARIMA genetic programming approach. *J Earth Syst Sci* 126(2):18
- Mohammad-Azari S, Bozorg-Haddad O, Loáiciga HA (2020) State-of-art of genetic programming applications in water-resources systems analysis. *Environ Monit Assess* 192(2):1-17
- Nourani V, Pradhan B, Ghaffari H, Sharifi SS (2014) Landslide susceptibility mapping at Zonouz Plain, Iran using genetic programming and comparison with frequency ratio, logistic regression, and artificial neural network models. *Nat Hazards* 71(1):523-547
- Raghavendra N, Deka PC (2014) Support vector machine applications in the field of hydrology: a review. *Appl Soft Comput* 19:372-386
- Rezaie-Balf M, Fani Nowbandegani S, Samadi SZ, Fallah H, Alaghamand S (2019) An ensemble decomposition-based artificial intelligence approach for daily streamflow prediction. *Water* 11(4):709
- Ritter A, Munoz-Carpena R (2013) Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *J Hydrol* 480:33-45
- Salojärvi K., Auvinen H., Ikonen E. (1982) A plan for fisheries management in the lakes drained by the Oulujoki river. In: Ilmavirta V., Jones R.I., Persson P.E. (eds) *Lakes and Water Management. Developments in Hydrobiology* 7, vol 7. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-8003-7_33
- Sivapragasam C, Maheswaran R, Venkatesh V (2008) Genetic programming approach for flood routing in natural channels. *Hydrol Process* 22(5):623-628
- Tanev I, Brzozowski M, Shimohara K (2005) Evolution, generality and robustness of emerged surrounding behavior in continuous predators-prey pursuit problem. *Genet Program Evolvable Mach* 6(3):301-318

- Terzi Ö, Ergin G (2014) Forecasting of monthly river flow with autoregressive modeling and data-driven techniques. *Neural Comput Appl* 25(1):179-188
- Thapa, S., Zhao, Z., Li, B., Lu, L., Fu, D., Shi, X., ... & Qi, H. (2020). Snowmelt-driven streamflow prediction using machine learning techniques (LSTM, NARX, GPR, and SVR). *Water*, 12(6), 1734.
- Uyumaz A, Danandeh Mehr A, Kahya E, Erdem H (2014) Rectangular side weirs discharge coefficient estimation in circular channels using linear genetic programming approach. *J Hydroinformatics* 16(6):1318–1330. <https://doi.org/10.2166/hydro.2014.112>
- Valipour M (2015) Long-term runoff study using SARIMA and ARIMA models in the United States. *Meteorol Appl* 22(3):592-598
- Wang WC, Chau KW, Cheng CT, Qiu L (2009) A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *J Hydrol* 374(3):294-306
- Yadav B, Ch S, Mathur S, Adamowski J (2016) Discharge forecasting using an online sequential extreme learning machine (OS-ELM) model: a case study in Neckar River, Germany. *Measurement* 92:433-445
- Yan X, Mohammadian A, Khelifa A (2021) Modeling spatial distribution of flow depth in fluvial systems using a hybrid two-dimensional hydraulic-multigene genetic programming approach. *J Hydrol* 600:126517. <https://doi.org/10.1016/j.jhydrol.2021.126517>
- Yaseen ZM, Ebtehaj I, Bonakdari H et al (2017) Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model. *J Hydrol* 554:263-276
- Yaseen ZM, Sulaiman SO, Deo RC, Chau KW (2019) An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *J Hydrol* 569:387-408
- Yang, D., Zhao, Y., Armstrong, R., Robinson, D., & Brodzik, M. J. (2007). Streamflow response to seasonal snow cover mass changes over large Siberian watersheds. *Journal of Geophysical Research: Earth Surface*, 112(F2).
- Yang D, Zhao Y, Armstrong R, Robinson D (2009) Yukon River streamflow response to seasonal snow cover changes. *Hydrological Processes: an International Journal* 23(1):109-121
- Zhang Z, Zhang Q, Singh VP (2018) Univariate streamflow forecasting using commonly used data-driven models: literature review and case study. *Hydrol Sci J* 63(7):1091–1111
- Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: a review. *Journal of Hydrology*, 598, 126266.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.