CHAPTER

3 HYDROLOGICAL TIME SERIES FORECASTING USING THREE DIFFERENT HEURISTIC REGRESSION TECHNIQUES

Ozgur Kisi*, **Jalal Shiri**†, **Vahdettin Demir**‡

**Faculty of Natural Sciences and Engineering, Ilia State University, Tbilisi, Georgia* †*Faculty of Agriculture, University of Tabriz, Tabriz, Iran* ‡*Engineering Faculty, KTO Karatay University, Konya, Turkey*

3.1 INTRODUCTION

Accurate prediction of streamflow is very important in hydrology, hydraulic, and water resources engineering as it can directly affect the dams operation and performance, groundwater recharge/exploitation, sediment conveyance capability of river, watershed management, etc. There are many parameters which affect the next day runoff e.g. precipitation, evaporation, groundwater level, soil antecedent moisture content, etc. [8]. Although it would be possible to identify some sophisticated models taking into account the hydrological and hydro-meteorological parameters, it would be preferable to have a model to simulate the streamflow using previously recorded flow magnitudes [8]. So far, numerous studies have reported streamflow prediction [4,18,1,9].

Nayak et al. [11] applied neuro-fuzzy models for forecasting streamflow. Also, Huang et al. [5], Wang et al. [23], Jain and Kumar [6], and Kisi [8] employed neural networks (NN) for predicting streamflows. Smith et al. [18] used a discrete wavelet transform for quantifying streamflow variability. Coulibaly and Burn [2] used wavelet analysis for identifying variability in annual streamflows. Solomatine and Xue [14] applied M5 model trees (M5Tree) and neural network techniques for flood forecasting. Zhou et al. [24] developed a wavelet-based model for predicting monthly river flow. Kisi [9] introduced the wavelet-ANN conjunction model for forecasting intermittent streamflow. Shiri and Kisi [16] developed and tested a wavelet-neuro-fuzzy model for forecasting short-term and long-term streamflow. Shiri et al. [17] compared different heuristic data driven approaches for predicting daily streamflows. Shabri and Suhartono [15] applied least-square support vector regression (LSSVR) for streamflow forecasting. Pandhiani and Shabri [12] applied wavelet-least-square support vector machines and wavelet regression models for monthly streamflow data. Kisi [10] applied LSSVR and neuro-fuzzy embedded fuzzy c-means clustering for streamflow forecasting and estimation. Karimi et al. [7] introduced a wavelet-genetic programming based model for predicting streamflows. To the knowledge of the authors, there is no published work indicating the input–output mapping capability of LSSVR, multivariate adaptive regression spline (MARS), or M5Tree methods in forecasting daily streamflows.

Handbook of Neural Computation. DOI: 10.1016/B978-0-12-811318-9.00003-X

This paper is concerned with the implementation of three different heuristic methods, LSSVR, MARS, and M5Tree, for forecasting daily streamflow.

3.2 METHODS

3.2.1 LEAST-SQUARE SUPPORT VECTOR REGRESSION

Any function $f(x)$ in support vector regression (SVR) can be considered as [20]:

$$f(x) = w^T \varphi(x) + b \tag{3.1}$$

where w^T and $\varphi(x)$ denote the transposed output layer vector and the kernel function, respectively, and b is the bias. Here, x stands for the model input which has a dimension of $N \times n$, where N and n show the number of data patterns and number of input parameters, respectively. Using the minimization of the following cost function, Vapnik et al. [22] calculate the w and b as follows:

$$\text{Cost function} = \frac{1}{2} w^T + c \sum_{k=1}^{N} \left(\xi_k - \xi_k^* \right) \tag{3.2}$$

subject to the following constraints:

$$\begin{cases} y_k - w^T \varphi(x_k) - b \le \varepsilon + \xi_k, & k = 1, 2, 3, \ldots, N \\ w^T \varphi(x_k) + b - y_k \le \varepsilon + \xi_k^*, & k = 1, 2, 3, \ldots, N \\ \xi_k, \xi_k^* \ge 0, & k = 1, 2, 3, \ldots, N \end{cases} \tag{3.3}$$

where x_k, y_k and ε denote the kth data input, kth data output, and fixed precision of the function approximation, respectively. As well, ξ_k and ξ_k^* show slack variables, which should be used to determine the allowed error margin. From Eq. (3.2), c is considered the tuning parameter of the SVR for minimizing the cost function. The Lagrangian theory might be applied:

$$L\left(a, a^*\right) = -\frac{1}{2} \sum_{k,l=1}^{N} \left(a_k - a_k^*\right)\left(a_l - a_l^*\right) K(x_k - x_l) - \varepsilon \sum_{k=1}^{N} \left(a_k - a_k^*\right) + \sum_{k=1}^{N} y_k \left(a_k - a_k^*\right) \tag{3.4a}$$

$$\sum_{k=1}^{N} \left(a_k - a_k^*\right) = 0, \quad a_k, a_k^* \in [0, c] \tag{3.4b}$$

$$K(x_k - x_l) = \varphi(x_k)^T \varphi(x_l), \quad k = 1, 2, \ldots, N \tag{3.4c}$$

where α_k and α_k^* are Lagrangian multipliers. Consequently, the final form of the SVR is obtained as:

$$f(x) = \sum_{k,l=1}^{N} \left(a_k - a_k^*\right) K(x - x_k) + b \tag{3.5}$$

SVR-based models are usually solved through searching solutions for quadratic programming issues with linear inequality constraints. Suykens and Vandewalle [21] proposed a least-square modification to the original SVR (which will be referred to as LSSVR) for improving the SVR-based models. In LSSVR, solutions can be obtained through solving a set of linear equations. The cost function in LSSVR is determined as:

$$\text{Cost function} = \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum_{k=1}^{N} e_k^2 \tag{3.6}$$

which is subjected to the following constraint:

$$y_k = w^T \varphi(x_k) + b + e_k \tag{3.7}$$

where γ and e_k are tuning parameters in LSSVR method and the variable error, respectively. The Lagrangian for this problem is written as:

$$L(w, b, e, a) = \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum_{k=1}^{N} e_k^2 - \sum_{k=1}^{N} a_k\left(w^T \varphi(x_k) + b + e_k - y_k\right) \tag{3.8}$$

where a_k are Lagrangian multipliers. For solving the problem, the derivatives of Eq. (3.8) should be equated to zero. Thus, the following equations are obtained:

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \quad \Rightarrow \quad w = \sum_{k=1}^{N} a_k \varphi(x_k) \\ \frac{\partial L}{\partial b} = 0 \quad \Rightarrow \quad \sum_{k=1}^{N} a_k = 0 \\ \frac{\partial L}{\partial e_k} = 0 \quad \Rightarrow \quad a_k = \gamma e_k, \quad k = 1, 2, \ldots, N \\ \frac{\partial L}{\partial a_k} = 0 \quad \Rightarrow \quad w^T \varphi(x_k) + b + e_k - y_k = 0 \quad k = 1, 2, \ldots, N \end{cases} \tag{3.9}$$

where γ is a tuning parameter of LSSVR. From Eq. (3.9), it is seen that there are $2N+2$ equations and $2N+2$ unknown parameters (a_k, e_k, w and b). Therefore, the parameters of LSSVR can be obtained by solving the system of equations as presented in Eq. (3.9). In the present research, the radial basis function (RBF) kernel was employed, which can be presented as:

$$K(x, x_k) = \exp\left(-\|x_k - x\|^2/\sigma^2\right) \tag{3.10}$$

where σ^2 is the other tuning parameter. So, there are two tuning parameters in LSSVR with RBF kernel function, which can be obtained through minimizing deviation of the experimental data from the simulated values. Normally, minimization of the mean square error (MSE) is considered as the objective function to find the tuning parameters of LSSVR:

$$\text{MSE} = \frac{\sum_{i=1}^{n} (O_{\text{rep/pred}_i} - O_{\text{exp}_i})^2}{n} \tag{3.11}$$

Table 3.1 The Daily Statistical Parameters of Data Set for Karabuk-1314, Derecikviran-1335 stations

Stations	Data Set	x_{mean} (m^3/s)	S_x (m^3/s)	C_{s_x} (m^3/s)	x_{min} (m^3/s)	x_{max} (m^3/s)	r1	r2	r3
Karabuk	1964–1975	27.96	35.44	3.00	2.73	436	0.932	0.846	0.785
	1976–1986	27.67	37.38	3.68	1.69	474	0.928	0.832	0.771
	1987–1997	22.17	31.28	3.69	0.9	413	0.946	0.863	0.799
	1998–2008	21.53	34.78	5.21	0.07	595	0.940	0.843	0.762
Derecikviran	1964–1975	106.67	107.26	2.32	4.5	1340	0.933	0.844	0.785
	1976–1986	101.00	121.79	4.57	0.1	1816	0.890	0.760	0.682
	1987–1997	90.79	97.07	3.02	5	1373	0.931	0.836	0.764
	1998–2008	93.14	122.63	5.23	6.5	2204	0.946	0.853	0.773

Table 3.2 Regularization Constant and Width of RBF Kernel Parameters of the Optimal LSSVR Models for Each Combination of Karabuk and Derecikviran stations

			Input combination		
Cross Validation	**Training Data Set**	**Test Data Set**	**(i)**	**(ii)**	**(iii)**
Karabuk					
M1	1964–1986	1991–1999	(90, 30)	(90, 40)	(10, 90)
M2	1964–1975 and 1998–2008	1982–1990	(10, 90)	(40, 90)	(20, 50)
M3	1964–1975 and 1987–2008	1973–1981	(10, 20)	(100, 100)	(10, 90)
M4	1976–2008	1964–1972	(10, 60)	(80, 10)	(1, 1)
Derecikviran					
M1	1964–1986	1991–1999	(10, 90)	(10, 90)	(10, 90)
M2	1964–1975 and 1998–2008	1982–1990	(10, 20)	(90, 20)	(90, 20)
M3	1964–1975 and 1987–2008	1973–1981	(40, 10)	(10, 20)	(50, 20)
M4	1976–2008	1964–1972	(10, 10)	(10, 90)	(1, 1)

where O shows the output, and subscripts rep/pred and exp denote, respectively, the represented/predicted and experimental values. Here, n stands for the number of data patterns.

3.2.2 MULTIVARIATE ADAPTIVE REGRESSION SPLINE

Multivariate adaptive regression splines (MARS) is a non-parametric regression technique which can be considered as an extension of linear models that automatically models nonlinearities and interactions between variables [3]. MARS considers models of the form:

$$\widehat{f}(x) = \sum_{i}^{k} c_i B_i(x) \tag{3.12}$$

which is a weighted sum of function $B_i(x)$; c_i stands for the constant coefficients.

Each basis function $B_i(x)$ can take one of the following three forms:

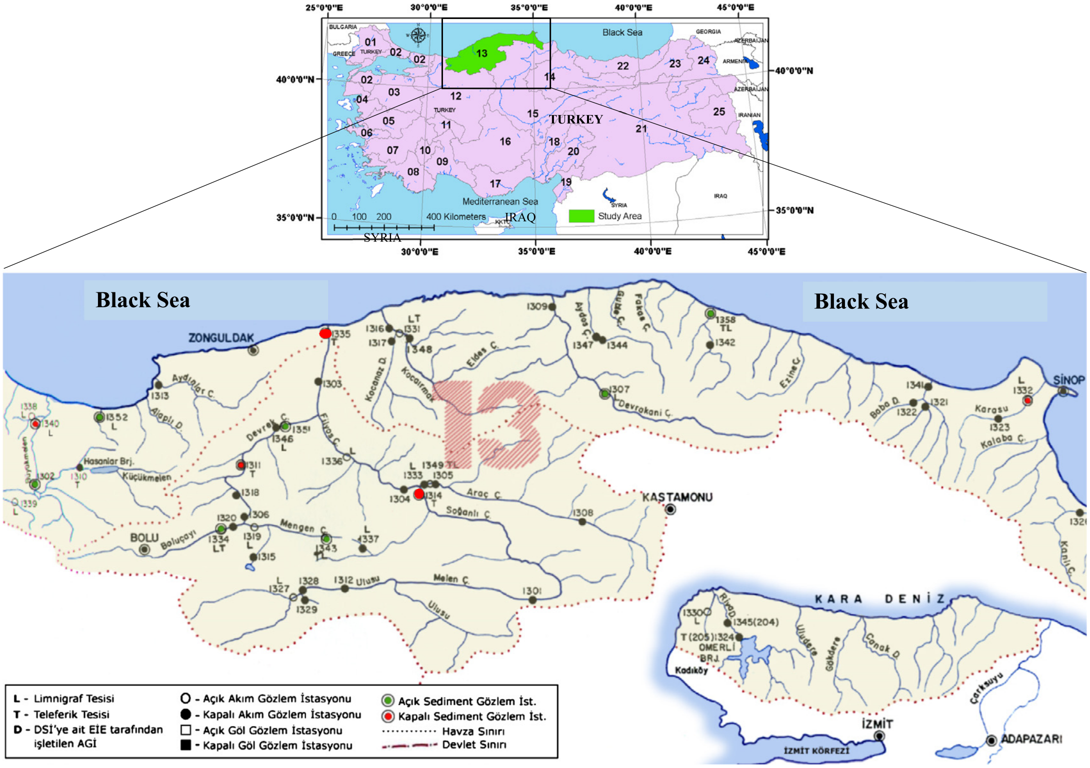

FIGURE 3.1

The basins of Turkey and the Karabuk-1314 and Derecikviran-1335 stations in the western Black Sea Basin (Basin No. 13).

1) a constant 1. There is just one such term, the intercept.
2) *hinge* function. A hinge function is of the form $\max(o, x - \text{constant})$ or $\max(o, \text{constant} - x)$. MARS selects the variables and their values for knots of the hinge functions automatically.

A product of two or more hinge functions: Interactions between two or more variables might be simulated by these functions.

3.2.3 M5 MODEL TREE

M5 model tree is a decision tree learner for regression task which is used to predict values of numerical response variable Y [13], which is a binary decision tree having linear regression functions at the terminal (leaf) nodes, which can predict continuous numerical attributes. M5 model tree can simulate the phenomena with very high dimensionality up to hundreds of attributes [13]. This ability sets M5 apart from regression tree learners at the time (like MARS), whose computational cost grows very quickly when the number of features increases.

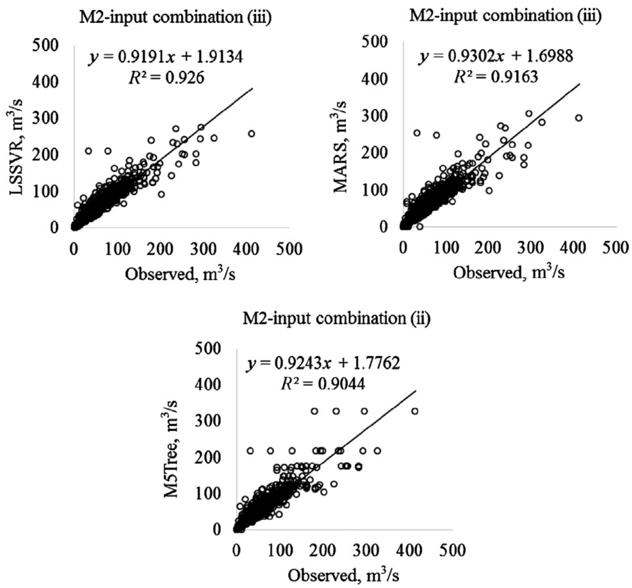

FIGURE 3.2

Observed and estimated streamflows by the optimal LSSVR, MARS, and M5Tree models in Karabuk Station.

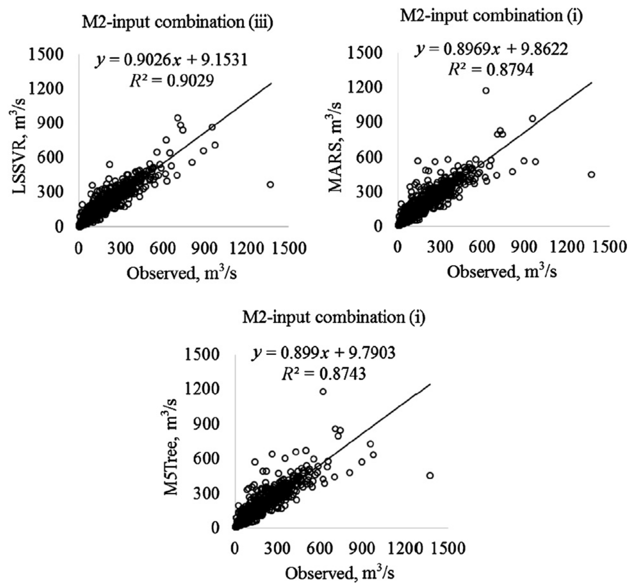

FIGURE 3.3

Observed and estimated streamflows by the optimal LSSVR, MARS, and M5Tree models in Derecikviran Station.

Table 3.3 Comparison of the LSSVR Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

			Input Combinations			
Statistics	**Cross Validation**	**Test Data Set**	**(i)**	**(ii)**	**(iii)**	**Mean**
Karabuk						
RMSE	M1	1998–2008	12.02	12.36	11.94	**12.11**
	M2	1987–1997	9.69	8.76	8.51	**8.99**
	M3	1976–1986	13.30	12.60	12.57	**12.83**
	M4	1964–1975	12.17	11.26	10.94	**11.46**
		Mean	**11.80**	**11.25**	**10.99**	**11.34**
MAE	M1	1998–2008	3.29	3.26	2.95	**3.17**
	M2	1987–1997	3.20	2.77	2.78	**2.92**
	M3	1976–1986	4.33	3.99	4.00	**4.11**
	M4	1964–1975	4.34	3.92	3.91	**4.06**
		Mean	**3.79**	**3.48**	**3.41**	**3.56**
R^2	M1	1998–2008	0.881	0.875	0.884	**0.880**
	M2	1987–1997	0.881	0.875	0.884	**0.880**
	M3	1976–1986	0.873	0.886	0.887	**0.882**
	M4	1964–1975	0.883	0.900	0.905	**0.896**
		Mean	**0.880**	**0.884**	**0.890**	**0.885**
Derecikviran						
RMSE	M1	1998–2008	41.34	41.34	38.35	**40.34**
	M2	1987–1997	35.06	30.71	30.24	**32.01**
	M3	1976–1986	56.64	51.51	49.68	**52.61**
	M4	1964–1975	37.88	36.67	36.82	**37.13**
		Mean	**42.73**	**40.06**	**38.77**	**40.52**
MAE	M1	1998–2008	12.73	12.73	11.97	**12.48**
	M2	1987–1997	12.41	11.19	11.01	**11.54**
	M3	1976–1986	16.97	16.19	15.42	**16.19**
	M4	1964–1975	16.34	15.92	14.99	**15.75**
		Mean	**14.61**	**14.01**	**13.35**	**13.99**
R^2	M1	1998–2008	0.890	0.890	0.906	**0.896**
	M2	1987–1997	0.870	0.900	0.903	**0.891**
	M3	1976–1986	0.791	0.821	0.834	**0.815**
	M4	1964–1975	0.876	0.884	0.884	**0.881**
		Mean	**0.857**	**0.874**	**0.882**	**0.871**

A model tree generation includes two different steps. The first step involves using a splitting criterion to make a decision tree. The splitting criterion of M5 model tree algorithm is based on treating the standard deviation of the class values which reach a node as an error measure at that node, and calculating the expected reduction in this error as a result of testing each attribute at that node. By splitting process, the data in child nodes get lower standard deviation values compared to the par-

Table 3.4 Comparison of the MARS Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

Statistics	Cross Validation	Test Data Set	Input Combinations (i)	(ii)	(iii)	Mean
Karabuk						
RMSE	M1	1998–2008	12.94	13.04	13.14	**13.04**
	M2	1987–1997	10.38	9.54	9.06	**9.66**
	M3	1976–1986	13.20	12.91	12.44	**12.85**
	M4	1964–1975	12.38	11.87	11.77	**12.01**
		Mean	**12.22**	**11.84**	**11.60**	**11.89**
MAE	M1	1998–2008	3.36	3.06	3.07	**3.16**
	M2	1987–1997	3.20	2.90	2.90	**3.00**
	M3	1976–1986	4.29	4.04	4.04	**4.12**
	M4	1964–1975	4.35	3.96	3.99	**4.10**
		Mean	**3.80**	**3.49**	**3.50**	**3.60**
R^2	M1	1998–2008	0.862	0.860	0.857	**0.860**
	M2	1987–1997	0.891	0.908	0.916	**0.905**
	M3	1976–1986	0.875	0.881	0.889	**0.882**
	M4	1964–1975	0.879	0.889	0.891	**0.886**
		Mean	**0.877**	**0.885**	**0.888**	**0.883**
Derecikviran						
RMSE	M1	1998–2008	45.61	42.47	40.98	**43.02**
	M2	1987–1997	33.75	34.47	35.15	**34.46**
	M3	1976–1986	53.31	52.37	52.25	**52.64**
	M4	1964–1975	38.00	36.84	36.62	**37.15**
		Mean	**42.67**	**41.53**	**41.25**	**41.82**
MAE	M1	1998–2008	13.01	12.53	12.18	**12.57**
	M2	1987–1997	12.05	11.35	11.63	**11.68**
	M3	1976–1986	16.60	15.87	16.11	**16.19**
	M4	1964–1975	16.51	15.33	15.52	**15.79**
		Mean	**14.54**	**13.77**	**13.86**	**14.06**
R^2	M1	1998–2008	0.862	0.882	0.893	**0.879**
	M2	1987–1997	0.879	0.876	0.872	**0.876**
	M3	1976–1986	0.809	0.817	0.818	**0.814**
	M4	1964–1975	0.876	0.884	0.885	**0.881**
		Mean	**0.856**	**0.865**	**0.867**	**0.863**

ent node. After evaluating all possible splits, M5 selects the split which maximizes the expected error reduction. This division often produces a large tree-like structure that may cause overfitting. Consequently, the tree must be pruned back. So, the second stage would involve pruning the overgrown tree and replacing the subtrees with linear regression functions. This technique of generating the model tree splits the parameter space into subspaces and builds a linear regression model in each of them.

Table 3.5 Comparison of the M5Tree Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

Statistics	Cross Validation	Test Data Set	Input Combinations			
			(i)	(ii)	(iii)	Mean
Karabuk						
RMSE	M1	1998–2008	13.53	13.83	14.40	**13.92**
	M2	1987–1997	11.00	9.69	10.07	**10.25**
	M3	1976–1986	13.52	14.97	15.65	**14.71**
	M4	1964–1975	12.41	12.87	13.09	**12.79**
		Mean	**12.61**	**12.84**	**13.30**	**12.92**
MAE	M1	1998–2008	3.35	3.20	3.44	**3.33**
	M2	1987–1997	3.22	3.06	3.28	**3.19**
	M3	1976–1986	4.39	4.46	4.72	**4.52**
	M4	1964–1975	4.35	4.25	4.47	**4.36**
		Mean	**3.83**	**3.74**	**3.98**	**3.85**
R^2	M1	1998–2008	0.849	0.842	0.829	**0.840**
	M2	1987–1997	0.879	0.904	0.897	**0.894**
	M3	1976–1986	0.869	0.840	0.826	**0.845**
	M4	1964–1975	0.878	0.871	0.867	**0.872**
		Mean	**0.869**	**0.865**	**0.855**	**0.863**
Derecikviran						
RMSE	M1	1998–2008	40.66	40.91	52.62	**44.73**
	M2	1987–1997	34.51	38.33	38.76	**37.20**
	M3	1976–1986	53.71	14.97	52.55	**40.41**
	M4	1964–1975	38.08	37.86	37.92	**37.95**
		Mean	**41.74**	**33.02**	**45.46**	**40.07**
MAE	M1	1998–2008	12.70	12.24	13.47	**12.80**
	M2	1987–1997	12.27	12.18	12.61	**12.35**
	M3	1976–1986	16.82	4.46	16.31	**12.53**
	M4	1964–1975	16.27	15.69	16.17	**16.04**
		Mean	**14.52**	**11.14**	**14.64**	**13.43**
R^2	M1	1998–2008	0.894	0.892	0.816	**0.867**
	M2	1987–1997	0.874	0.852	0.848	**0.858**
	M3	1976–1986	0.806	0.840	0.816	**0.821**
	M4	1964–1975	0.875	0.877	0.878	**0.876**
		Mean	**0.862**	**0.865**	**0.839**	**0.856**

3.3 APPLICATIONS AND RESULTS

The daily stream flow time series data obtained from two stations, Karabuk (Station No. 1314, Latitude 32°38′32″, Longitude 41°10′11″) and Derecikviran (Station No. 1335, Latitude 32°04′44″, Longitude 41°32′49″), operated by the Directorate General of Renewable Energy were used in the study (Fig. 3.1). The stations (Station Nos. 1314 and 1335) have drainage areas of 5087 km^2 and 13,300 km^2 and gauge

Table 3.6 Comparison of the P-LSSVR Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

Statistics	Cross Validation	Test Data Set	Input Combinations (i)	(ii)	(iii)	Mean
Karabuk						
RMSE	M1	1998–2008	11.72	11.27	11.83	**11.61**
	M2	1987–1997	9.63	8.58	8.48	**8.90**
	M3	1976–1986	13.24	12.41	12.50	**12.72**
	M4	1964–1975	11.87	11.08	10.97	**11.30**
		Mean	**11.61**	**10.83**	**10.95**	**11.13**
MAE	M1	1998–2008	3.25	2.92	2.94	**3.03**
	M2	1987–1997	3.11	2.79	2.80	**2.90**
	M3	1976–1986	4.29	3.97	3.99	**4.08**
	M4	1964–1975	4.38	3.91	3.93	**4.07**
		Mean	**3.76**	**3.40**	**3.42**	**3.52**
R^2	M1	1998–2008	0.887	0.896	0.886	**0.890**
	M2	1987–1997	0.905	0.925	0.927	**0.919**
	M3	1976–1986	0.875	0.890	0.888	**0.884**
	M4	1964–1975	0.888	0.903	0.905	**0.899**
		Mean	**0.889**	**0.903**	**0.901**	**0.898**
Derecikviran						
RMSE	M1	1998–2008	40.98	39.54	38.17	**39.56**
	M2	1987–1997	33.96	29.93	29.91	**31.27**
	M3	1976–1986	52.66	51.34	49.28	**51.09**
	M4	1964–1975	37.50	36.38	34.82	**36.23**
		Mean	**41.27**	**39.30**	**38.04**	**39.54**
MAE	M1	1998–2008	12.66	12.12	11.92	**12.23**
	M2	1987–1997	13.02	11.12	10.91	**11.69**
	M3	1976–1986	16.95	16.03	15.25	**16.08**
	M4	1964–1975	16.16	15.27	15.03	**15.49**
		Mean	**14.70**	**13.64**	**13.28**	**13.87**
R^2	M1	1998–2008	0.891	0.901	0.907	**0.900**
	M2	1987–1997	0.878	0.905	0.905	**0.896**
	M3	1976–1986	0.813	0.822	0.836	**0.824**
	M4	1964–1975	0.878	0.887	0.896	**0.887**
		Mean	**0.865**	**0.879**	**0.886**	**0.877**

data of 271 m and 2 m above the sea level, respectively. Cross validation method was used in the applications and data were divided into four equal parts. For each application, three data sets were used for training of the applied models while the remaining data set was used for testing.

The statistical parameters of the streamflow data of Karabuk and Derecikviran stations are reported in Table 3.1. In the table, the x_{mean}, S_x, C_{s_x}, $x_{\min}$ and $x_{\max}$ show the mean, standard deviation, skewness

Table 3.7 Comparison of the P-MARS Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

Statistics	Cross Validation	Test Data Set	Input Combinations (i)	(ii)	(iii)	Mean
Karabuk						
RMSE	M1	1998–2008	12.97	13.09	13.14	**13.06**
	M2	1987–1997	10.14	9.46	9.06	**9.55**
	M3	1976–1986	13.17	12.68	12.44	**12.76**
	M4	1964–1975	12.18	11.87	11.76	**11.94**
		Mean	**12.12**	**11.78**	**11.60**	**11.83**
MAE	M1	1998–2008	3.52	3.28	3.07	**3.29**
	M2	1987–1997	3.35	3.11	2.90	**3.12**
	M3	1976–1986	4.35	4.11	4.04	**4.17**
	M4	1964–1975	4.35	3.96	4.06	**4.12**
		Mean	**3.89**	**3.62**	**3.52**	**3.68**
R^2	M1	1998–2008	0.862	0.859	0.857	**0.859**
	M2	1987–1997	0.895	0.909	0.916	**0.907**
	M3	1976–1986	0.876	0.885	0.889	**0.883**
	M4	1964–1975	0.882	0.889	0.891	**0.888**
		Mean	**0.879**	**0.886**	**0.889**	**0.884**
Derecikviran						
RMSE	M1	1998–2008	44.36	42.49	40.07	**42.31**
	M2	1987–1997	34.27	34.37	35.76	**34.80**
	M3	1976–1986	53.32	52.39	52.08	**52.60**
	M4	1964–1975	37.84	36.87	36.62	**37.11**
		Mean	**42.45**	**41.53**	**41.13**	**41.70**
MAE	M1	1998–2008	13.76	13.37	12.77	**13.30**
	M2	1987–1997	12.65	11.80	11.83	**12.09**
	M3	1976–1986	16.86	16.32	16.35	**16.51**
	M4	1964–1975	16.45	15.73	15.69	**15.96**
		Mean	**14.93**	**14.30**	**14.16**	**14.47**
R^2	M1	1998–2008	0.870	0.882	0.899	**0.884**
	M2	1987–1997	0.876	0.876	0.868	**0.873**
	M3	1976–1986	0.809	0.816	0.819	**0.815**
	M4	1964–1975	0.877	0.884	0.885	**0.882**
		Mean	**0.858**	**0.865**	**0.868**	**0.863**

coefficients, minimum and maximum, respectively. It is apparent from the table that the flow data show a significantly high skewed distribution.

The ability of LSSVR, MARS, and M5Tree models was compared in forecasting daily streamflows. Models were compared with each other with respect to root-mean-square error (RMSE), mean absolute

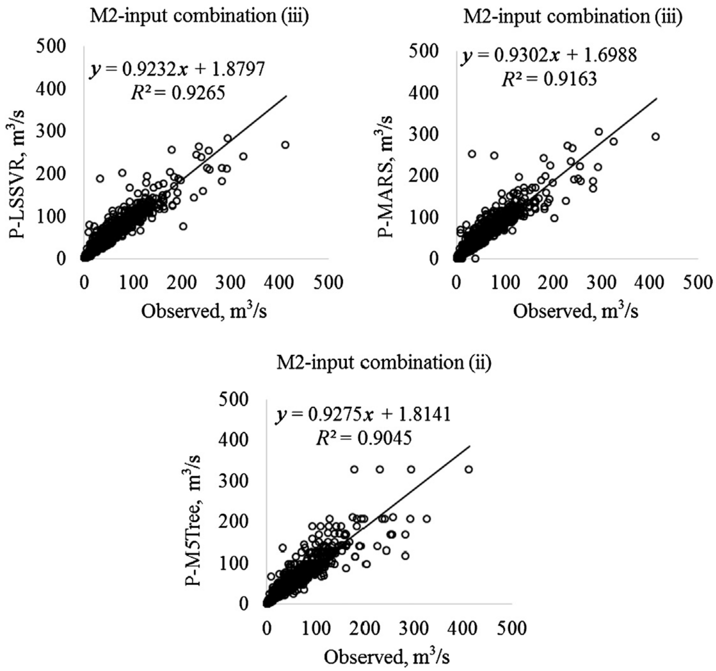

FIGURE 3.4

Observed and estimated streamflows by the optimal P-LSSVR, P-MARS, and P-M5Tree models in Karabuk station.

error, and determination coefficient (R^2). The RMSE and MAE can be given as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{\text{observed},i} - Q_{\text{forecast},i})^2} \tag{3.13}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Q_{\text{observed},i} - Q_{\text{forecast},i}| \tag{3.14}$$

where Q_{observed} and Q_{forecast} are the observed and forecasted streamflow values, n is the number of data.

Different regularization constant and RBF kernel parameter values were tried for the LSSVR models. Three input combinations were tried as: (i) Q_t, (ii) Q_t, Q_{t-1}, (iii) Q_t, Q_{t-1}, Q_{t-2}. The optimal control parameters of the LSSVR models are provided in Table 3.2 for Karabuk and Derecikviran stations. In the table, M1 indicates the first model calibrated using the data covering period of 1964–1986 and tested using the data between 1991 and 1999. Table 3.3 reports the test results of the LSSVR models for the Karabuk and Derecikviran stations. Each model shows different accuracy and M2 gen-

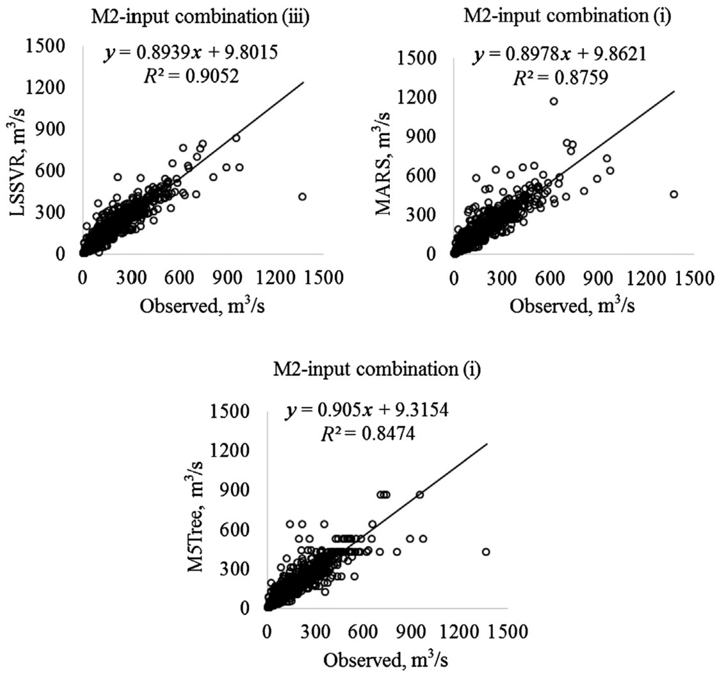

FIGURE 3.5

Observed and estimated streamflows by the optimal P-LSSVR, P-MARS, and P-M5Tree models in Derecikviran station.

erally provides the best forecasting accuracy for the both stations. However, M3 is the worst model with respect to RMSE and MAE criteria. Input combination (iii) including current and two previous streamflow data generally give better accuracy than the other two input combinations. The RMSE of the LSSVR ranges from 8.99 m^3/s to 12.83 m^3/s and from 32.01 m^3/s to 52.61 m^3/s for the Karabuk and Derecikviran stations, respectively. Test RMSE, MAE, and R^2 results of the MARS models are shown in Table 3.4 for the Karabuk and Derecikviran stations. Similarly to the LSSVR, here also M2 provides the best forecasting accuracy. MARS model comprising input combination (iii) has a better accuracy than the other two input combinations. However, there is a slight difference between input combinations (ii) and (iii) for the both stations. The RMSE of the MARS ranges from 9.66 m^3/s to 13.04 m^3/s and from 34.46 m^3/s to 52.64 m^3/s for the Karabuk and Derecikviran stations, respectively. Test results of the M5Tree models are given in Table 3.5 for the Karabuk and Derecikviran stations. For this model, also M2 provides better accuracy than the other models. Differently from the LSSVR and MARS models, the input combination (i) and the input combination (ii) give the best accuracy in predicting daily streamflows of Karabuk and Derecikviran stations, respectively. The RMSE of the M5Tree models ranges from 10.25 m^3/s to 14.71 m^3/s and from 37.20 m^3/s to 44.73 m^3/s for the Karabuk and Derecikviran stations, respectively. Comparison of Tables 3.3–3.5 shows that the LSSVR model outperforms the MARS and M5Tree models in predicting daily streamflows. For the M3 of

Table 3.8 Comparison of the P-M5Tree Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

Statistics	Cross Validation	Test Data Set	Input Combinations (i)	(ii)	(iii)	Mean
Karabuk						
RMSE	M1	1998–2008	14.37	13.21	13.37	**13.65**
	M2	1987–1997	10.59	9.70	10.01	**10.10**
	M3	1976–1986	13.83	14.96	15.06	**14.62**
	M4	1964–1975	12.72	13.37	12.97	**13.02**
		Mean	**12.88**	**12.81**	**12.85**	**12.85**
MAE	M1	1998–2008	3.52	3.31	3.44	**3.42**
	M2	1987–1997	3.31	3.16	3.30	**3.26**
	M3	1976–1986	4.56	4.76	4.75	**4.69**
	M4	1964–1975	4.50	4.42	4.39	**4.44**
		Mean	**3.97**	**3.91**	**3.97**	**3.95**
R^2	M1	1998–2008	0.830	0.858	0.855	**0.848**
	M2	1987–1997	0.889	0.905	0.901	**0.898**
	M3	1976–1986	0.863	0.840	0.838	**0.847**
	M4	1964–1975	0.872	0.862	0.869	**0.868**
		Mean	**0.864**	**0.866**	**0.866**	**0.865**
Derecikviran						
RMSE	M1	1998–2008	58.32	61.16	63.14	**60.87**
	M2	1987–1997	38.41	40.08	40.83	**39.77**
	M3	1976–1986	65.43	65.30	66.97	**65.90**
	M4	1964–1975	40.55	38.44	40.02	**39.67**
		Mean	**50.67**	**51.24**	**52.74**	**51.55**
MAE	M1	1998–2008	15.22	15.42	15.69	**15.45**
	M2	1987–1997	12.56	12.75	13.20	**12.84**
	M3	1976–1986	18.01	17.68	18.49	**18.06**
	M4	1964–1975	16.94	16.52	16.86	**16.77**
		Mean	**15.68**	**15.59**	**16.06**	**15.78**
R^2	M1	1998–2008	0.774	0.753	0.747	**0.758**
	M2	1987–1997	0.847	0.839	0.834	**0.840**
	M3	1976–1986	0.713	0.716	0.702	**0.710**
	M4	1964–1975	0.859	0.874	0.864	**0.865**
		Mean	**0.798**	**0.796**	**0.787**	**0.794**

Derecikviran station, however, M5Tree model seems to be better than the other models (LSSVR and MARS). Figs. 3.2 and 3.3 illustrate the observed and estimated streamflows by the optimal models for Karabuk and Derecikviran stations, respectively. Less scattered estimates of the LSSVR model relative to the other models are clearly seen from the figures for the both stations. The models seem to be more successful in Karabuk than in the Derecikviran station. This may be due to the higher autocorrelations of the Karabuk's steamflow data (see Table 3.1).

Table 3.9 Comparison of the Log-LSSVR Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

Statistics	Cross Validation	Test Data Set	Input Combinations (i)	(ii)	(iii)	Mean
Karabuk						
RMSE	M1	1998–2008	11.67	10.93	10.95	**11.18**
	M2	1987–1997	9.84	9.11	9.08	**9.34**
	M3	1976–1986	13.49	12.64	12.58	**12.90**
	M4	1964–1975	11.82	11.13	10.44	**11.13**
		Mean	**11.71**	**10.95**	**10.76**	**11.14**
MAE	M1	1998–2008	3.29	2.99	2.95	**3.07**
	M2	1987–1997	3.18	2.88	2.90	**2.98**
	M3	1976–1986	4.42	4.00	4.05	**4.16**
	M4	1964–1975	4.34	3.94	3.75	**4.01**
		Mean	**3.80**	**3.45**	**3.41**	**3.56**
R^2	M1	1998–2008	0.891	0.908	0.908	**0.902**
	M2	1987–1997	0.903	0.917	0.919	**0.913**
	M3	1976–1986	0.871	0.887	0.888	**0.882**
	M4	1964–1975	0.890	0.902	0.915	**0.903**
		Mean	**0.889**	**0.904**	**0.907**	**0.900**
Derecikviran						
RMSE	M1	1998–2008	40.02	40.94	41.36	**40.78**
	M2	1987–1997	33.60	30.59	33.21	**32.47**
	M3	1976–1986	53.28	51.05	49.10	**51.14**
	M4	1964–1975	37.71	35.44	34.43	**35.86**
		Mean	**41.15**	**39.51**	**39.53**	**40.06**
MAE	M1	1998–2008	13.09	12.59	12.49	**12.72**
	M2	1987–1997	12.43	11.10	11.39	**11.64**
	M3	1976–1986	16.92	15.95	15.35	**16.07**
	M4	1964–1975	16.48	15.18	14.53	**15.40**
		Mean	**14.73**	**13.70**	**13.44**	**13.96**
R^2	M1	1998–2008	0.898	0.906	0.904	**0.903**
	M2	1987–1997	0.881	0.902	0.883	**0.889**
	M3	1976–1986	0.810	0.826	0.840	**0.825**
	M4	1964–1975	0.877	0.891	0.897	**0.889**
		Mean	**0.867**	**0.881**	**0.881**	**0.876**

Kisi [10] investigated the effect of periodicity in forecasting monthly streamflows by adding month number as input to the models (e.g., LSSVR) and he found that it significantly increased the models' accuracy. Therefore, here, also the effect of periodicity was investigated by adding day number of each streamflow data as input to the applied models. Test results of the periodic LSSVR (P-LSSVR) models are provided in Table 3.6 for the Karabuk and Derecikviran stations, respectively. M2 model seems to have the best accuracy in daily streamflow forecasting. The RMSE of the P-LSSVR ranges from

Table 3.10 Comparison of the Log-MARS Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

Statistics	Cross Validation	Test Data Set	Input Combinations (i)	(ii)	(iii)	Mean
Karabuk						
RMSE	M1	1998–2008	12.88	14.81	12.62	**13.44**
	M2	1987–1997	9.83	9.96	8.71	**9.50**
	M3	1976–1986	13.51	13.20	13.47	**13.39**
	M4	1964–1975	12.24	12.66	12.13	**12.34**
		Mean	**12.12**	**12.66**	**11.73**	**12.17**
MAE	M1	1998–2008	3.34	3.31	3.00	**3.22**
	M2	1987–1997	3.16	3.01	2.85	**3.01**
	M3	1976–1986	4.38	4.17	4.18	**4.24**
	M4	1964–1975	4.36	4.14	4.08	**4.19**
		Mean	**3.81**	**3.66**	**3.53**	**3.67**
R^2	M1	1998–2008	0.871	0.831	0.874	**0.859**
	M2	1987–1997	0.904	0.902	0.925	**0.910**
	M3	1976–1986	0.872	0.876	0.872	**0.873**
	M4	1964–1975	0.882	0.873	0.884	**0.880**
		Mean	**0.882**	**0.871**	**0.889**	**0.881**
Derecikviran						
RMSE	M1	1998–2008	46.90	45.75	44.52	**45.72**
	M2	1987–1997	34.45	33.40	33.04	**33.63**
	M3	1976–1986	53.76	53.71	52.72	**53.40**
	M4	1964–1975	37.67	36.81	36.29	**36.92**
		Mean	**43.19**	**42.42**	**41.64**	**42.42**
MAE	M1	1998–2008	13.83	12.89	12.77	**13.16**
	M2	1987–1997	12.56	11.62	11.55	**11.91**
	M3	1976–1986	17.03	16.34	16.18	**16.52**
	M4	1964–1975	16.42	15.51	15.26	**15.73**
		Mean	**14.96**	**14.09**	**13.94**	**14.33**
R^2	M1	1998–2008	0.871	0.883	0.888	**0.881**
	M2	1987–1997	0.874	0.882	0.885	**0.880**
	M3	1976–1986	0.806	0.808	0.815	**0.809**
	M4	1964–1975	0.877	0.883	0.886	**0.882**
		Mean	**0.857**	**0.864**	**0.868**	**0.863**

8.90 m^3/s to 12.72 m^3/s and from 31.27 m^3/s to 51.09 m^3/s for the Karabuk and Derecikviran stations, respectively. Table 3.7 gives the test accuracy of the periodic MARS (P-MARS) models in predicting daily streamflows of Karabuk and Derecikviran stations. Similarly to the P-LSSVR, M2 model provides the best accuracy. The RMSE of the P-MARS ranges from 9.55 m^3/s to 13.06 m^3/s and from 34.80 m^3/s to 52.60 m^3/s for the Karabuk and Derecikviran stations, respectively. Test accuracy of the periodic M5Tree (P-M5Tree) models is reported in Table 3.8 in respect of RMSE, MAE, and R^2. The

Table 3.11 Comparison of the Log-M5Tree Models in Predicting Daily Streamflows of the Karabuk and Derecikviran Stations

Statistics	Cross Validation	Test Data Set	Input Combinations (i)	(ii)	(iii)	Mean
Karabuk						
RMSE	M1	1998–2008	13.60	14.39	14.56	**14.18**
	M2	1987–1997	10.94	9.44	9.73	**10.04**
	M3	1976–1986	13.58	14.81	15.18	**14.52**
	M4	1964–1975	12.36	12.85	12.38	**12.53**
		Mean	**12.62**	**12.87**	**12.96**	**12.82**
MAE	M1	1998–2008	3.44	3.31	3.39	**3.38**
	M2	1987–1997	3.28	3.09	3.23	**3.20**
	M3	1976–1986	4.44	4.43	4.55	**4.47**
	M4	1964–1975	4.38	4.29	4.39	**4.35**
		Mean	**3.89**	**3.78**	**3.89**	**3.85**
R^2	M1	1998–2008	0.849	0.829	0.825	**0.834**
	M2	1987–1997	0.878	0.909	0.903	**0.897**
	M3	1976–1986	0.869	0.845	0.836	**0.850**
	M4	1964–1975	0.879	0.869	0.879	**0.876**
		Mean	**0.869**	**0.863**	**0.861**	**0.864**
Derecikviran						
RMSE	M1	1998–2008	41.58	46.82	59.42	**49.27**
	M2	1987–1997	34.52	38.22	40.36	**37.70**
	M3	1976–1986	53.54	54.81	54.20	**54.18**
	M4	1964–1975	38.44	37.89	38.62	**38.32**
		Mean	**42.02**	**44.43**	**48.15**	**44.87**
MAE	M1	1998–2008	13.49	13.06	14.16	**13.57**
	M2	1987–1997	12.60	12.21	12.85	**12.55**
	M3	1976–1986	17.11	16.81	16.84	**16.92**
	M4	1964–1975	16.80	16.02	16.32	**16.38**
		Mean	**15.00**	**14.52**	**15.04**	**14.86**
R^2	M1	1998–2008	0.895	0.865	0.767	**0.842**
	M2	1987–1997	0.874	0.849	0.832	**0.852**
	M3	1976–1986	0.809	0.799	0.804	**0.804**
	M4	1964–1975	0.872	0.876	0.871	**0.873**
		Mean	**0.863**	**0.847**	**0.819**	**0.843**

RMSE of the P-M5Tree ranges from 10.10 m^3/s to 14.62 m^3/s and from 39.67 m^3/s to 65.90 m^3/s for the Karabuk and Derecikviran stations, respectively. Comparison of Tables 3.3–3.8 shows that adding periodicity slightly increases the accuracy of LSSVR models in both stations while it significantly decreases M5Tree accuracy in predicting streamflows of Derecikviran station. As seen from Tables 3.6–3.8, P-LSSVR outperforms the P-MARS and P-M5Tree models. Figs. 3.4–3.5 demonstrate the observed and estimated streamflows by the optimal periodic models for the Karabuk and Dere-

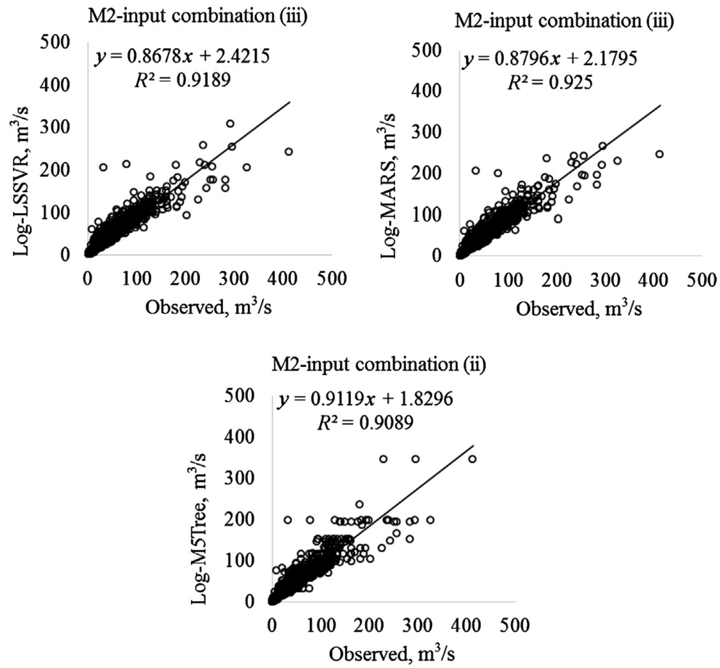

FIGURE 3.6

Observed and estimated streamflows by the optimal Log-LSSVR, Log-MARS, and Log-M5Tree models in Karabuk station.

cikviran stations, respectively. Here also the P-LSSVR model has less scattered estimates than the other models. Similarly to the previous application, the differences within each model are much bigger in Derecikviran than the Karabuk station.

Sudheer et al. [19] used log transformation in streamflow forecasting and he found that the ANN model built on the log-transformed series performed better. Therefore, here also the effect of log transformation on models' accuracy was investigated by using transformed data as inputs to the applied models. Test results of the log-transformed LSSVR (Log-LSSVR) models are provided in Table 3.9 for the Karabuk and Derecikviran stations. Similarly to the previous applications, M2 model has the best performance in daily streamflow forecasting while the M3 provides the worst accuracy. The RMSE of the Log-LSSVR ranges from 9.34 m^3/s to 12.90 m^3/s and from 32.47 m^3/s to 51.14 m^3/s for the Karabuk and Derecikviran stations, respectively. Test accuracy of the log-transformed MARS (Log-MARS) models in predicting daily streamflows of Karabuk and Derecikviran stations is provided in Table 3.10. Here also the M2 model gives the best accuracy. The RMSE of the Log-MARS ranges from 9.50 m^3/s to 13.44 m^3/s and from 33.63 m^3/s to 53.40 m^3/s for the Karabuk and Derecikviran stations, respectively. Table 3.11 demonstrates the test accuracy of the log-transformed M5Tree (Log-M5Tree) models. The RMSE of the P-M5Tree ranges from 10.04 m^3/s to 14.52 m^3/s and from 37.70 m^3/s to 49.27 m^3/s for the Karabuk and Derecikviran stations, respectively. Comparison of Tables 3.3–3.5 and Tables 3.9–3.11 indicates that the log transformation of the streamflow data does not increase models'

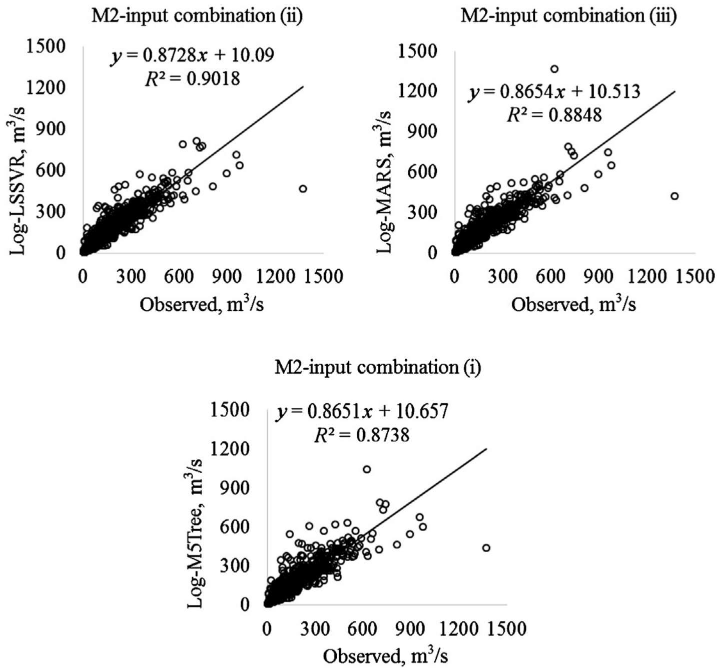

FIGURE 3.7

Observed and estimated streamflows by the optimal Log-LSSVR, Log-MARS, and Log-M5Tree models in Derecikviran station.

accuracy in daily forecasting. Apparent from Tables 3.9–3.11 is that the Log-LSSVR outperforms the Log-MARS and Log-M5Tree models in daily streamflow forecasting. The estimates of the optimal models whose inputs are log-transformed streamflow data are visually compared in Figs. 3.6–3.7. The Log-MARS model seems to be better than the other models in Karabuk station while the Log-LSSVR model has the least scattered estimates in Derecikviran station.

3.4 CONCLUSION

The study compared the accuracy of three different heuristic methods, LSSVR, MARS, and M5Tree models, in forecasting hydrological time series. Daily streamflow time series data from two stations Karabuk and Derecikviran in Turkey were used. Cross validation method was used in the applications by dividing data into four equal parts. Comparison results with respect to RMSE, MAE, and R^2 indicated that the LSSVR model generally outperformed the MARS and M5Tree models in forecasting daily streamflows. The models were found to have better accuracy in Karabuk than in the Derecikviran station due to the higher autocorrelations of Karabuk streamflow data. The effects of periodicity and log transformation were also examined in forecasting daily streamflows. Results showed that by adding periodicity component, a slight increment was found in accuracy of LSSVR models in both stations while the M5Tree accuracy was decreased in forecasting streamflows of Derecikviran station. Log transformation of the streamflow data did not increase the accuracy of the applied models in daily forecasting. LSSVR model with periodicity and log-transformed data performed better than the corresponding MARS and M5Tree models.

REFERENCES

[1] J.F. Adamowski, Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis, J. Hydrol. 353 (2008) 247–266.
[2] P. Coulibaly, H.D. Burn, Wavelet analysis of variability in annual Canadian streamflows, Water Resour. Res. 40 (2004) W03105.
[3] J.H. Friedman, Multivariate adaptive regression splines, Ann. Statist. 19 (1991) 1.
[4] M.J. Hayes, D.A. Wilhite, M.D. Svoboda, K.H. Smith, Drought management: crisis vs management, in: Proceedings of the 1996 International Conference and Exposition on Natural Disaster Reduction, American Society of Civil Engineering, Washington, DC, 1996, pp. 371–372.
[5] W. Huang, B. Xu, A. Chan-Hilton, Forecasting flows in Apalachicola river using neural networks, Hydrol. Process. 18 (2004) 2545–2564.
[6] A. Jain, A.M. Kumar, Hybrid neural network models for hydrologic tome series forecasting, Appl. Soft Comput. 7 (2007) 585–592.
[7] S. Karimi, J. Shiri, O. Kisi, A.A. Shiri, Short-term and long-term streamflow prediction by using 'wavelet–gene expression' programming approach, ISH J. Hydraul. Eng. 22 (2) (2016) 148–162.
[8] O. Kisi, River flow forecasting and estimation using different artificial neural network techniques, Hydrol. Res. 39 (1) (2008) 27–40.
[9] O. Kisi, Neural networks and wavelet conjunction model for intermittent streamflow forecasting, ASCE J. Hydrol. Eng. 14 (8) (2009) 773–782.
[10] O. Kisi, Streamflow forecasting and estimation using least square support vector regression and adaptive neuro-fuzzy embedded fuzzy c-means clustering, Water Resour. Manage. 29 (14) (2015) 5109–5127.
[11] P.C. Nayak, K.P. Sudheer, D.M. Rangan, K.S. Ramasastri, A neuro-fuzzy computing technique for modeling hydrologic time series, J. Hydrol. 291 (2004) 52–66.
[12] S.M. Pandhiani, A.B. Shabri, Time series forecasting suing wavelet-least squares support vector machines and wavelet regression models for monthly stream flow data, Open J. Stat. 3 (3) (2013) 183–194.
[13] J.R. Quinlan, Learning with continuous classes, in: A. Adams, L. Sterling (Eds.), Proc. AI'92 (Fifth Australian Joint Conf. on Artificial Intelligence), World Scientific, Singapore, 1992, pp. 343–348.
[14] D. Solomatine, Y. Xue, M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China, ASCE J. Hydrol. Eng. 9 (6) (2004) 491–501.
[15] A. Shabri, S. Suhartono, Streamflow forecasting using least-squares support vector machines, Hydrol. Sci. J. 57 (7) (2012) 1–19.
[16] J. Shiri, Ö. Kişi, Short term and long term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model, J. Hydrol. 394 (2010) 486–493.
[17] J. Shiri, O. Kisi, O. Makarynskyy, A.A. Shiri, B. Nikoofar, Forecasting daily stream flow using artificial intelligence approaches, ISH J. Hydraul. Eng. 18 (3) (2012) 204–214.
[18] L.C. Smith, D.L. Turcotte, B. Isacks, Stream flow characterization and feature detection using a discrete wavelet transform, Hydrol. Process. 12 (1998) 233–249.
[19] K.P. Sudheer, P.C. Nayak, K.S. Ramasastri, Improving peak flow estimates in artificial neural network river flow models, Hydrol. Process. 17 (3) (2003) 677–686.
[20] J.A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, J. Suykens, et al., Least Squares Support Vector Machines, World Scientific, 2002.
[21] J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300.
[22] V. Vapnik, S.E. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, Adv. Neural Inf. Process. Syst. (1997) 281–287.

[23] W. Wang, P.H.A.J.M. Van Gelder, J.K. Vrijling, J. Ma, Forecasting daily streamflow using hybrid ANN models, J. Hydrol. 324 (2006) 383–399.
[24] H.C. Zhou, Y. Peng, G-H. Liang, The research of monthly discharge predictor–corrector model based on wavelet decomposition, Water Resour. Manage. 22 (2) (2008) 217–227.