Primljen / Received: 9.5.2017.
Ispravljen / Corrected: 1.3.2018.
Prihvaćen / Accepted: 17.8.2018.
Dostupno online / Available online: 10.9.2018.

Comparison of supervised learning methods for prediction of monthly average flow

Authors:

Jadran Berbić, PhD. CE
Croatian Meteorological and Hydrological Service
jberbic@hotmail.com

Assoc.Prof. Eva Ocvirk, PhD. CE
University of Zagreb
Faculty of Civil Engineering
ocvirk@grad.hr

Assist.Prof. Gordon Gilja, PhD. CE
University of Zagreb
Faculty of Civil Engineering
ggilja@grad.hr

Original scientific paper

Jadran Berbić, Eva Ocvirk, Gordon Gilja

Comparison of supervised learning methods for prediction of monthly average flow

Long-term planning of water resources systems requires knowledge of long-term availability of water, most often in the form of monthly average flow information. Knowledge from stochastic hydrology is most often applied, and possible scenarios also involve generation of synthetic flow. The use of climatic models imposes the possibility of modelling based on future scenarios, and it is assumed in the paper that supervised learning can be applied for this purpose. The paper analyses accuracy of three supervised learning models in three approaches and the autoregressive model in the first approach, for predicting monthly average flow as related to the length of a historic dataset.

Key words:
long-term planning, monthly average flow, autoregressive model, supervised learning

Izvorni znanstveni rad

Jadran Berbić, Eva Ocvirk, Gordon Gilja

Usporedba metoda nadziranog učenja u svrhu predviđanja srednjeg mjesečnog protoka

Dugoročno planiranje hidrotehničkih sustava zahtijeva poznavanje dugoročne dostupnosti vode, najčešće u obliku srednjeg mjesečnog protoka. Uglavnom se koriste znanja iz stohastičke hidrologije, a mogući scenariji dobivaju se generiranjem sintetičkog protoka. Raspolaganje klimatskim modelima nameće mogućnost modeliranja iz budućih scenarija, a pretpostavka u radu je da se za tu svrhu može primjenjivati nadzirano učenje. U radu je analizirana preciznost tri modela nadziranog učenja u tri pristupa i autoregresivnog modela u prvom pristupu, za predviđanje srednjeg mjesečnog protoka, a u ovisnosti o duljini povijesnog niza.

Ključne riječi:
dugoročno planiranje, srednji mjesečni protok, autoregresivni model, nadzirano učenje

Wissenschaftlicher Originalbeitrag

Jadran Berbić, Eva Ocvirk, Gordon Gilja

Vergleich der Methoden des überwachten Lernens zum Zweck der Vorhersage des mittleren monatlichen Durchflusses

Die langfristige Planung von hydrotechnischen Systemen erfordert Kenntnisse über die langfristige Verfügbarkeit von Wasser, meist in Form des mittleren monatlichen Durchflusses. Hauptsächlich werden Kenntnisse aus der stochastischen Hydrologie angewendet, und mögliche Szenarien erhält man durch Erzeugung des synthetischen Durchflusses. Die Verfügung über Klimamodelle drängt die Möglichkeit der Modellierung anhand zukünftiger Szenarien auf, und die Voraussetzung in der Abhandlung ist die, dass zu diesem Zweck das überwachte Lernen angewendet werden kann. In der Abhandlung wurde die Präzision von drei Modellen des überwachten Lernens in drei Ansätzen und des autoregressiven Modells im ersten Ansatz zur Vorhersage des mittleren monatlichen Durchflusses analysiert, abhängig von der Länge der historischen Reihe.

Schlüsselwörter:
langfristige Planung, mittlerer monatlicher Durchfluss, autoregressives Modell, überwachtes Lernen

1. Introduction

Upcoming pressures on water resources like increasing of population, need for energy and food, demand increasing of efficiency and effectivity of production [1, 2]. Significant climatic variations and changes cause more often phenomenons of extremely wet and dry periods and change statistical distribution of hydrological events [3-5]. Stochastic methods and supervised learning methods represent practical tool for simulation of river flow on hydrologically studied basins. The assumption is that it is possible to analyze present and future needs related to water resources systems by using of appropriate simulation models in water resources systems management if the long enough historical time series of measurements is on disposition. For example, building of quality simulation model is necessary for conduction of simulation-optimization procedure for analysis of availability of water for needs dependent on water reservoir [6]. Therefore, predictions of mean monthly river flow month-by-month and long term planning are of great importance for planning and choosing of water reservoir regime.

Analysis of acceptability of usage of historical flow time series in the dependence of length and data on disposition is given in the paper. Possibility of usage of autoregressive model (AR) and three supervised learning (SL) methods for prediction on the basis of flow, and the same three methods for prediction on the basis of precipitation amount and air temperature, is tested. Objectives of the analysis are: give answer on the question what is the minimum length of historical time series at which is acceptable to use mentioned methods and analyze possibility of building a quality model which could be used for prediction of flow from the results of climatic models.

1.1. Overview and conclusions from previous researches

Machine learning is used for finding of patterns in data and their generalization by induction. Supervised learning is the part of machine learning and artificial intelligence used for searching of parameters of hypothesis (function), based on given data (inputs and outputs) and assumed hypothesis, which results with the best predictions on unseen instances, for solving of problems of classification and regression. From the literature review it can be concluded that in hydrology SL is often used for needs of real time prediction (with time step to several hours) and for short term and mid term predictions (1-7 days), but rarely for long term predictions (1 month) and even less often for long term planning. Usage of smaller time step is interesting due to the presence of greater amount of data for model building and is relatively simple to build quality model without usage of external variables (hence, flow is predicted from flow by itself). On the other side, the model built in that way is not able to reliably predict several time steps ahead from the current step (due to error generation with increase of time steps number), except if timely averaged variables and external predictors (air temperature, precipitation amount, etc.) are eventually used as input variables.

The most popular SL model is artificial neural network (ANN) and is present in the vast majority of work in the subject area. Building of models with ANN is consisted of choosing the weights in synapses with objective of minimization of differences between desirable input and real input of ANN on the basis of chosen criterium and by learning from examples [7]. Considering the reasons of continuous improvement of hydrological cycle, hydrologists used to set greater aspiration on physically based models through the history of modelling, which leads to the design of more complex models with time [8]. Main advantages of ANN, for example, avoiding the problem of full understanding of runoff for hydrological modelling, which is complex on real spatial scale, have already been noticed in the last two decades. There is no need for introducing the assumptions of linearity and describing complex relationships of different processes in detail, usage of data is more flexible, models can be built relatively quickly [9]. Similar advantages are present in the application of other SL models. Support vector machine (SVM) is appreciated because of its generalization ability, strict theoretical basis, relatively simple usage, and robustness on the problems of regression and pattern recognition [7]. One of the objectives of the work is comparison of three different SL models: ANN as a popular one, SVM as a robust one, but in hydrological literature less present, and NNM (nearest neighbours method) as very simple compared to the other two.

Cigizoglu et al. (2005) compared ANN with stochastic autoregressive moving averages model (ARMA) and multilinear regression model (MLR) [10]. ANN coupled with generalized regression (GR) gave more accurate results than classically used ANN. Additionally, they built ANN on the series of synthetic flows, which enabled the usage of considerably more data and improved models, and the most accurate was GR ANN. Nilsson et al. (2006) predicted mean monthly inflow in the basin by using ANN with 6-12 external predictors. After the attempt with temperature and precipitation, results are improved by adding the amount of snow and seasonal characteristics, while soil moisture and north Atlantic oscillation index did not improve the prediction accuracy [11]. Wu and Chau (2010) predicted mean monthly flow one month ahead by using the ANN, NNM and ARMA model with 6-12 inputs (flows). Phase space reconstruction (PSR) preceded the model building. NNM and ARMA gave better results than ANN and the PSR-ANN combination, while ANN improved with moving averages gave best results [12]. Guo et al. (2011) introduced improvements of, orderly, ANN and SVM with the wavelet method and PSR, particle swarm optimization with the SVM and Levenberg-Marquartd algorithm with the ANN, by building models with 8 inputs (flows). More accurate results were gained, with more complex procedure for prediction of flow 1 month ahead, compared to ANN and SVM [13]. Akiner and Akkoyounly (2012) used ANN for missing data reconstruction and precipitation prediction, while runoff for the following decade was estimated by using the predicted precipitation in SWAT model [14]. Farajzadeh et al. (2014) compared the accuracy of ARIMA (autoregressive model with integrated moving average) and ANN for prediction of mean monthly flow. After predicting the precipitation by those models, runoff was predicted from the precipitation – by using those models and by using runoff coefficient. ARIMA gave slightly better results, but the approach with the runoff coefficient was more accurate [15]. Terzi (2014) used GP (genetic programming) for prediction of mean monthly flow from precipitation measured at three stations and from flow measured at two stations, and compared the method with MLR [16].

On the subject area (Vinalić, Cetina) ANN was applied for the purpose of short term prediction of inflow in the work by Matić (2014). Through the different approaches in usage of input variables (inflow, precipitation, air temperature) and application of ANN, the problem of model response compared to the real event for prediction from 1 to 10 days was resolved. Time series models (prediction of inflow by inflow), rainfall-runoff models (prediction of inflow by precipitation or by precipitation and inflow) and multivariate models (prediction of inflow from precipitation, temperature, etc.) were compared. A direct and an indirect method were used for the prediction of inflow. While the direct method is used for building a separate model for every different time step, the indirect method is used for building a single model for all of the time steps and due to the error generation is less accurate than the first method. Time series methods were the most accurate, but the response problem had to be resolved. After the following steps: introducing of the precipitation frequency and accumulated precipitation, usage of adaptive neural model with submodels for different seasons and optimization of neural model, introducing the averaged variables, the accuracy was significantly increased. Model building and calibration were done on the data from 2007th to 2011th year (1862 instances of data), and verification was done on the data from 2012th year (365 instances) [17]. As a rule, longer historical time series are used in the literature, 20-40 years [10, 16], 40-60 [11, 12, 15] and even about 100 years [13]. As the quality of SL models directly depends on the amount of data used for model building (probability to build a quality model is increased with the amount of used instances due to a greater possibility of generalization of laws in data patterns), it is interesting to test what length of historical time series is needed for building a model capable to predict outside of the time domain of historical time series, with satisfactory accuracy. According to the planning timeline there is a distinction between models for long term prediction (month-by-month) and for long term planning, while the purpose of this work includes development of the model for both. Models of time series were used (for one step ahead) and multivariate models (direct methods, considering that a single model learns general laws between input and output variables). The amount of instances was about 110-750 (historical time series from 10 to, orderly, 65, 62 and 60 years, years without flow measurements were not accounted), from which 60 % was used for building, 20 % for calibration, 20 % for verification, and the rest (from about 640 to 0 instances for 10 to, orderly, 65, 62 and 60 years) for additional verification of models.

Predictions of mean monthly flow by using SL are less frequently represented than predictions on shorter time basis, especially for long term planning (from the used literature, works [10, 14]). According to the knowledge of authors, there are no works which analyze influence of historical time series length on the accuracy of SL, while the amount of instances in data directly influences the model accuracy.

2. Methodology of the research

2.1. Autoregressive model: Thomas-Fiering AR(1)

Stochastic processes in water resources systems management are often described by Markov processes, and for the application purpose an assumption of historical time series stationarity is introduced. Markov processes are discretized by discrete processes – Markov chains [18]. General form of autoregressive models AR(p) of order ρ is [19]:

$$z_t = \sum_{i=1}^{p} \varphi_i z_{t-i} + \varepsilon_t$$

where: z_t is timely independent, normalized and standardized series, φ_i are autoregressive coefficients, ε_t are timely independent variables. The simplest is the autoregressive process of the first order AR(1). For normally distributed monthly flows with mean μ, variance σ^2, month-by-month correlation ρ Thomas-Fiering model AR(1) can be applied [6, 18]:

$$Q_{i+1} = \mu_{j+1} + \rho_j \frac{\sigma_{j+1}}{\sigma_j}(Q_i - \mu_j) + V_i \sigma_{j+1}(1 - \rho_j^2)^{0.5} \tag{1}$$

where: Q_i, Q_{i+1} are mean monthly flows for i+1st and i–th month, μ_j, μ_{j+1} are yearly averaged mean monthly flows for j-th i j+1st month, σ_j, σ_{j+1} are standard deviations of the j-th and j+1st month (yearly averaged), ρ_j are correlation coefficients of j-th and j+1st month, V_i is randomly chosen variable from normal distribution with mean $E[V_i]=0$ and unit variance $E[V_i]=1$. This model is often used for synthetic flow generation and is able to preserve statistical similarity with historical time series (e.g. [6]). The model is applied in the first approach (chapter 3), and the procedure is written in the programming environment Python (www.python.org, [20]), which is also used for all the other models.

2.2. Artificial neural networks

ANN mimics the learning principle used in the brain, by using the assumption that the process of the learning is taking place through electrochemical activity in networks consisted of neurons [21]. The most often ANN is consisted of three layers: the first one characterized by nodes which are in fact input variables, hidden layer consisted of nodes with an activation function and the layer with output node – predicted value (e.g. flow). Possibility of varying the number of hidden layers and nodes refers to the fact that the process of finding the appropriate ANN architecture is complex task [21, 22]. The type multilayer perceptron, with three layers, for solving the problem of regression, is used in the paper. Differences between real and modelled values are minimized by stochastic optimization algorithm based on the first order gradients (Adaptive Moment Estimation - ADAM). The algorithm is computationally efficient, does not demand much memory and is suitable for great amount of data [20, 23]. Parameters which most significantly affect the quality of model building are number of hidden layers and number of nodes in the layer, activation function, learning rate and learning momentum, maximum number of iterations in error optimization and error tolerance. There are also some other parameters, but generally the choice of the input variables is the key step in applying SL.

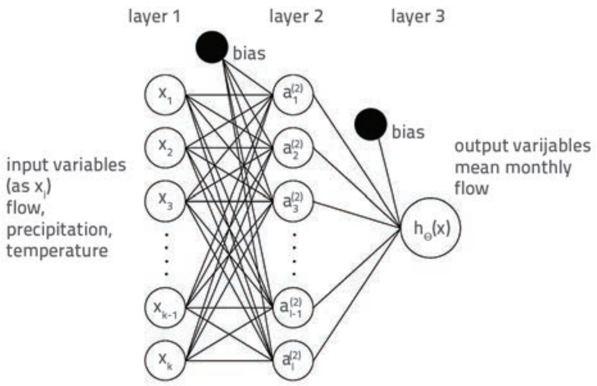

Figure 1. Structure of ANN with three layers

At ANN with three layers an activation $a_j^{(2)}$ in the node j = 1, 2, ..., l of the hidden layer (label 2) is calculated in the following way [24, 25]:

$$a_1^{(2)} = g(\theta_{1,0}^{(1)}x_0 + \theta_{1,1}^{(1)}x_1 + \theta_{1,2}^{(1)}x_2 + \ldots + \theta_{1,k\text{-}1}^{(1)}x_{k\text{-}1} + \theta_{1,k}^{(1)}x_k) = g(z_1^{(1)})$$

$$a_2^{(2)} = g(\theta_{2,0}^{(1)}x_0 + \theta_{2,1}^{(1)}x_1 + \theta_{2,2}^{(1)}x_2 + \ldots + \theta_{2,k\text{-}1}^{(1)}x_{k\text{-}1} + \theta_{2,k}^{(1)}x_k) = g(z_2^{(1)})$$

$$\vdots \tag{2}$$

$$a_l^{(2)} = g(\theta_{l,0}^{(1)}x_0 + \theta_{l,1}^{(1)}x_1 + \theta_{l,2}^{(1)}x_2 + \ldots + \theta_{l,k\text{-}1}^{(1)}x_{k\text{-}1} + \theta_{l,k}^{(1)}x_k) = g(z_l^{(1)})$$

where: g is the activation function, $\theta_{j,i}^{(1)}$ is weighted influence of input variable x_i on the activation $a_j^{(2)}$, i = 1, 2, ..., k. Index k refers to the numbers of nodes in the first layer, index i refers to the number of nodes in the hidden layer, and index 0 refers to the "bias" variable. Predicted value is calculated by the equation:

$$h_\theta(x) = a_1^{(3)} = g(\theta_{1,0}^{(2)}a_0^{(2)} + \theta_{1,1}^{(2)}a_1^{(2)} + \theta_{1,2}^{(2)}a_2^{(2)} + \ldots + \theta_{1,l\text{-}1}^{(2)}x_{l\text{-}1} + \theta_{1,l}^{(2)}x_l) \tag{3}$$

2.3. Support vector machine

In the classification problem SVM for chosen function finds parameters with which the function is optimally distanced from different classes, while in the regression problem the procedure is used for finding the optimal way of describing the data with chosen function. The problem is often multidimensional (it can be seen in the chapter 3 that the mean monthly flow is described as the function of at least 6 different predictors) and complex for graphical representation. SVM considers data as support vectors and approximates them with given hypothesis by minimizing the error of predicted value approximation. Thereat within the defined margin, that is, error ε, there must be as much points as possible. Bias and tolerance of amount of deviation greater than error are estimated by trade-off parameter C, positive constant value which determines the degree of error penalization. Bias and variance are estimated through the minimization of the sum of regularization part and model building error in the equation (4) [26, 27]:

$$\min(\frac{1}{2}\|w\|^2 + C\sum_{j=1}^{l}(\xi_j + \xi_j^*))^2 \tag{4}$$

With conditions:
$$y_j - \langle w, x_j \rangle - b \leq \varepsilon + \xi_j$$
$$\langle w, x_j \rangle + b\text{-}y_j \leq \varepsilon + \xi_j^*$$
$$\xi_j, \xi_j^* \geq 0$$

where: x_j are input variables, y_j predicted variable, w vector from the space of input variables, b bias variable, ξ_j, ξ_j^* slack variables used for estimation of deviation of input variables from margin.

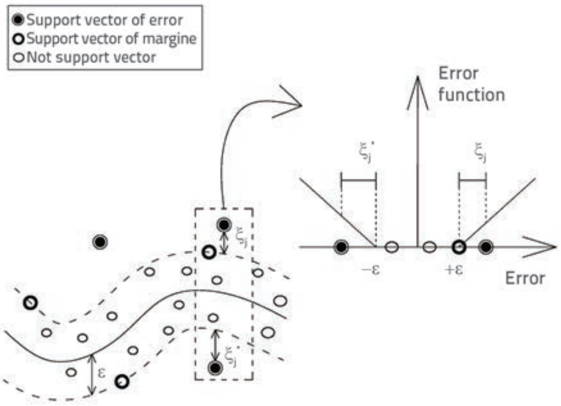

Figure 2. Representation of data (support vectors), hypothesis and margin of SVM (adopted from [28])

Hypothesis used for predicted variable approximation is [28]:

$$g(x) = \sum_{i,j=1}^{i}\left(\underline{\alpha}_i - \overline{\alpha}_i\right)K\left\langle u_i u_j \right\rangle + b$$

where: α_i are variables resulted from the transition to dual optimization problem, and K is the label for kernel. Programming environment enables choosing of function and kernel parameters (linear, polynomial and degree, radial basis function), and parameter C which affect the accuracy of prediction.

2.4. Nearest neighbours method

NNM uses the principle of searching for the set of values (in the part of data for model building) which are most similar to given ones (on the part of data for model prediction). For that purpose is needed to find distances between given and most similar points (k nearest neighbours). Too small amount of neighbours implies that the model is of greater sensitivity, while too large amount implies smaller accuracy due to the influence of distant neighbours. After the nearest neighbours are found, NNM calculates mean of the predicted values for every single neighbour [25, 27]. Defining of distance measures (Euclidian, Miknowski, etc.) in the paper did not have significant influence on the results accuracy. Beside the number of neighbours, weights of the influences of neighbours (uniform or dependent on distance) significantly affect the model accuracy. Programming environment enables choice between four algorithms for searching of the nearest neighbours: *ball tree***,** *kd tree***,** *brute* algorithm and auto choice of the best of those three. They are important because of computationally demanding calculation of distances between neighbours. Brute searches through the all possible options, which can last long for great number of neighbours, while other two use the logic of trees for searching. Kd tree is a binary tree which uses the logic of avoiding the calculation of distances for those points for which is known that they are distant (if the point A is far from the point B, and the point C is close to the B, then the C is far from the A). This algorithm is not efficient when D-dimensional measures for distances are used, if D>20 (the number of predictors is >20). The problem is solved by ball tree algorithm which, instead of using the Cartesian coordinate system, calculates the distances in the spherical coordinate system [27, 29].

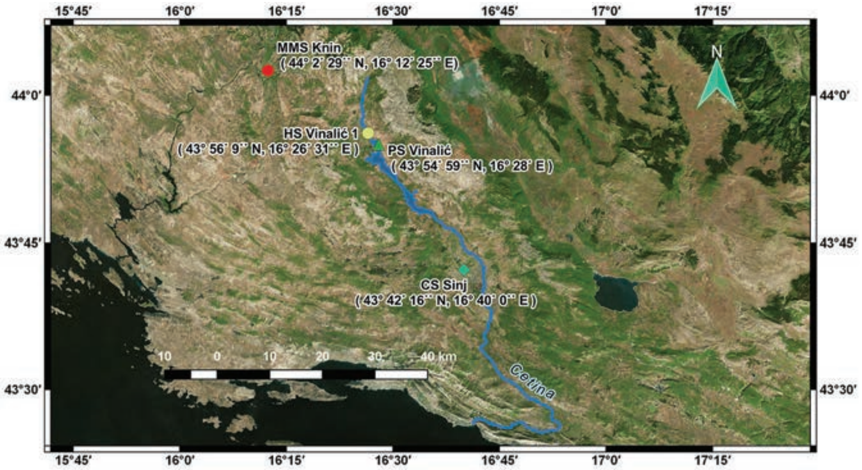
15°45'
16°0'
16°15'
16°30'
16°45'
17°0'
17°15'
44°0'
43°45'
43°30'
MMS Knin
(44° 2' 29'' N, 16° 12' 25'' E)
HS Vinalić 1
(43° 56' 9'' N, 16° 26' 31'' E)
PS Vinalić
(43° 54' 59'' N, 16° 28' E)
CS Sinj
(43° 42' 16'' N, 16° 40' 0'' E)
N
10 0 10 20 30 40 km
Cetina

Figure 3. Overview map with the positions of the stations (Cartographic background: QGIS, © 2007-2018 RDC ScanEx, http://kosmosnimki.ru/)

3. Used basis and forming of the models

3.1. Research area

Methodology and models were applied on the flows measurements of river Cetina from hydrological station Vinalić 1 (HS Vinalić 1). Historical time series of daily flows from 1946 until 2015 was on disposition, with gap in measurements from 1991 until 1997 [30]. Principally, those flows can be understood as inflows in water reservoir Peruća, but with dose of caution because the area is karstic. There are on disposition: accumulated daily precipitation (amount of fallen rainfall) and mean daily air temperature (further temperature) from main meteorological station Knin (MMS Knin, 250 m a. s. l.) in the period from 1949 until 2015, accumulated daily precipitation from precipitation station Vinalić (PS Vinalić, 350 m a. s. l.) in the period from 1951 until 2015 (gap in measurements 1991-1997), and mean monthly temperature from climatological station Sinj (CS Sinj, 308 m a. s. l.) in the period from 1949 until 2015 [31]. The overview map and the position of stations can be seen on the figure 3, while mean, minimum, maximum and averaged monthly flow can be seen on the figure 4. Forming of the water reservoir did not have significant influence on the flows of the HS Vinalić 1. It is important to consider this in every statistical analysis and also at applying SL due to its use of principle of learning from data. SL can cover changes in naturally present flows arose from building if enough number of instances for model building is present. It can be assumed that inflows from HS Vinalić 1 can be used for long term analysis of water availability.

3.2. Forming of the models

The first step is the choice of input variables, that is, predictors. Those are variables from which mean monthly flow is predicted, and three different approaches are used in the work - prediction of flow

- by using flow
- by using precipitation and temperature from one station (MMS Knin)

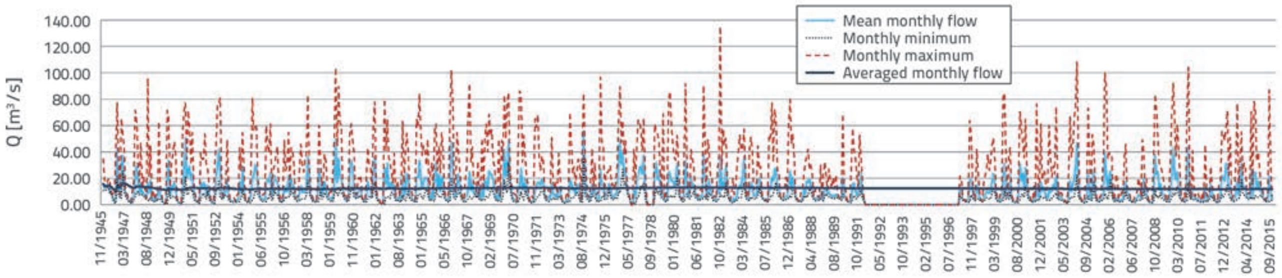

Figure 4. Flows on the hydrological station Vinalić 1

Table 1. Input variables used in analysis

	The first approach	The second and the third approach	
Quantity	Q...flow [m^3/s]	T...air temperature [°C]	P...precipitation [mm]
Index	avm, min, max, yavm, avmmin, avmmax	avm, min, max, yavm, avmin, avmax	avm, acc, max, yavm, avacc, avmax
avm, min, max...mean, minimum and maximum monthly value yavm, avmin, avmax...mean, minumum and maximum monthly value averaged at all years avmmin, avmmax...minimum and maximum mean monthly value at all years acc...accumulated monthly value avacc...accumulated monthly value averaged at all years			

- by using precipitation and temperature from two stations (precipitation from MMS Knin and PS Vinalić, temperature from MMS Knin and CS Sinj).

The first approach can be appropriate for forming the model which generates flow synthetically or for prediction 1 month ahead (eventually 2-3 with certain modifications). The second and the third approach use exclusively external variables and are appropriate for long term planning. Quantities in the table 1 are defined as input data.

Characteristic quantities shown in table 2 represent variables which were on the disposition for selection of the model configuration. E.g., at the first configuration one of the potential input variables is Q_{avmmin} – minumum mean monthly flow averaged at all years. Of all of the monthly values, its minimum for period 1946-2015 is 0.56 m^3/s, mean value is 2.70 m^3/s, and maximum value is 5.51 m^3/s, which is shown in table 2. Analogically, this also applies for other physical quantities at the second and the third configuration.

In the programming environment the procedure for processing and preparation of the data for model building is written. For every approach correlation of potential input variables (table 1) with mean monthly flow is analyzed. In the preliminary choice of input variables only variables with correlation of at least 0.55-0.60 were used. For yearly averaged variables, correlation with mean monthly flows for each particular year was considered, and for using in the preliminary choice it was needed to satisfy threshold in at least 30-40 % of historical time series. The result of procedure is time series for modelling procedure. The second step is preliminary building of the models. With obtained time series possibility of models AR, ANN, SVM and NNM for flow approximation is tested. Model parameters are preliminary

Table 2. Statistics of used characteristic quantities at all configurations

Flows (Vinalić 1) (1946-2015)						
	Q_{min}	Q_{avm}	Q_{max}	Q_{avmmin}	Q_{yavm}	Q_{avmmax}
Min.	0.13	0.56	1.01	0.56	3.38	10.5
Mean	5.52	11.9	28.4	2.70	11.9	35.9
Max.	36.9	55.9	135.0	5.51	19.8	55.9
St. dev.	4.04	9.48	24.2	1.52	5.78	15.3
N	765	765	765	12	12	12

Temperature and precipitation (Knin) (1949-2015)												
	T_{min}	T_{avm}	T_{max}	T_{avmin}	T_{yavm}	T_{avmax}	P_{avm}	P_{acc}	P_{max}	P_{yavm}	P_{avacc}	P_{avmax}
Min.	-12.4	-3.79	4.00	-5.20	3.49	9.9	0.00	0.00	0.00	0.00	0.00	0.00
Mean	7.04	13.1	18.9	7.09	13.2	18.9	2.91	88.3	29.1	2.89	87.8	28.1
Max.	23.2	26.9	31.9	19.3	24.7	28.5	11.5	354	155	8.03	241	63.7
St. dev.	7.48	6.89	6.17	7.06	6.69	5.85	1.94	58.9	18.8	0.94	28.7	6.33
N	731	731	731	732	732	732	732	732	732	732	732	732

Temperature (Sinj); precipitation (Vinalić) (1951-2015)												
	Tw_{min}	Tw_{avm}	Tw_{max}	Tw_{avmin}	Tw_{yavm}	Tw_{avmax}	H_{avm}	H_{acc}	H_{max}	H_{yavm}	H_{avacc}	H_{avmax}
Min.	-16.7	-3.13	3.4	-3.82	2.89	8.03	0.00	0.00	0.00	0.43	8.15	6.1
Mean	7.02	12.7	18.1	6.87	12.6	17.9	2.89	87.9	27.9	2.95	86.9	28.4
Max.	22.0	26.0	30.4	19.2	23.8	27.6	11.9	356	140	6.53	196	51.2
St. dev.	7.51	6.85	6.22	7.07	6.62	5.92	2.07	62.8	17.3	0.87	27.7	5.33
N	719	719	719	720	720	720	701	701	701	713	713	713

Table 3. Chosen parameters of SL models

Approach	ANN				SVM				NNM		
	Act. function	N. of nodes in the hidden layer	Initial learning rate	Tolerance	Kernel	St.	C	γ	N. of neighbours	Weights	Algorithm
1	tanh	25	$5.0*10^{-5}$	$2*10^{-9}$	poly	1	100.0	3.0	7	Jednol.	*brute*
2	tanh	45	$2.1*10^{-3}$	$2*10^{-9}$	rbf	/	56.73	0.009	10	Udalj.	auto
3	relu	30	$2.2*10^{-3}$	$2*10^{-9}$	rbf	/	56.73	0.009	10	Udalj.	auto

chosen for the purpose of statistical error measures minimization. The third step is conduction of sensitivity analysis of model accuracy for different configurations of input variables. From previously obtained time series some variables are removed or added with the goal of model accuracy increase. It has been shown that mostly variables highly correlated with flow have major contribution to the model accuracy. But, too much variables used to decrease accuracy, while some highly correlated variables have not significantly contributed to the accuracy and they were removed. Sometimes some of slightly less correlated variables contributed to the model accuracy significantly (eg. $T_{avmin-2}$). The result of the third step are the selected model configurations:

$$Q_{avm} = f\,(Q_{avm-1}, Q_{min-11}, Q_{min-1}, Q_{max-11}, Q_{yavm}, Q_{avmmin-11})$$
$$Q_{avm} = f\,(T_{avm-11}, T_{avmin-2}, P_{avm-1}, P_{avm}, P_{acc-11}, P_{acc-2}, P_{acc-1}, P_{max}, P_{avacc-2})$$
$$Q_{avm} = f\,(H_{avm-11}, H_{avm}, H_{acc-11}, H_{acc-1}, H_{acc}, H_{max}, H_{yavm}, H_{avacc-11}, H_{avacc}, T_{avm-11}, T_{avmin-2}, P_{avm-1}, P_{avm}, P_{acc-11}, P_{acc-2}, P_{acc-1}, P_{avacc-2}, P_{avacc}, Tw_{avm-11}, Tw_{avm-2}, Tw_{avm})$$

At the second and the third approach *T* and *P* are referred to a temperature and a precipitation from MMS Knin, *Tw* is referred to a temperature from CS Sinj, and *H* is referred to a precipitation from PS Vinalić.

The next step was optimization of model parameters. At AR model values of parameters *t* resulting in the best results of modelled flows were chosen. By the definition, *t* has normal distribution, and describes variation of monthly flows from the mean value. At ANN the influence of different activation functions (hyperbolic tangens - *tanh*, logistic, identity and rectification function - *relu*), number of nodes of hidden layer, initial learning rate and tolerance on accuracy was tested. At SVM the influence of kernel function (linear, polynomial – *poly,* radial basis function - *rbf*) and its degree, and parameters *C* and *γ*, was tested. At NNM, the number of neighbours, distribution of weights to neighbours and algorithm for distance calculation were varied. With parameters chosen in this step (table 3) the analysis of all historical time series of different lengths was conducted.

Historical time series at building of the models were always split chronologically: the first 60 % of the data for a model building, the next 20 % for a model calibration and the last 20 % for a model verification. In the first test maximum amount of data on disposition was used for all approaches: by order, 65, 62 and 60 years. In the second test, data from last years were removed so, by order, 60, 60 and 55 years have been used. In each further test last 5 years were removed until 10 years of data were left. Data from those years which were not used for the procedure building-calibration-verification were used for additional model verification. Therefore, in the second test, 5, 2 and 5 years for additional model verification was left, and in the last test, 55, 52 and 50 years (tables 4-6).

3.3. Statistical error measures

While optimizing the models the most attention was taken for achieving as high correlation as possible, as small root mean squared error as possible and as high coefficient of determination as possible. Correlation coefficient *R* represents interconnection between measured and predicted variable. Range 0-0.25 refers to weak, 0.25-0.6 refers to mid strong, while 0.6-1.0 refers to strong correlation [32]. High values of correlation coefficient do not necessary mean that the built model is capable to generalize well. Therefore, other error measures were also used: root mean squared error (*RMSE*), mean absolute error (*MAE*), relative absolute error (*RAE*), root relative squared error (*RRSE*), coefficient of determination or efficiency (R^2). Due to the limited space in the paper only R^2 and *RMSE* were shown. The used R^2 is the measure of likelihood of predicting the values unseen by the model and is not necessary the squared value of *R* (there are more definitions) and can be negative if model predicts arbitrarily bad. The value 1.0 represents absolutely accurate prediction [24]. Equations of mentioned measures can be found in the researches from the area (e.g. [10, 13, 24, 33]).

4. Results and discussion

4.1. The first approach

In the first approach it was shown that AR(1), with optimization of parameter *t*, can achieve wide range of values of flow. Those results belong to the part of model building and calibration. Parameter *t* by definition has got normal distribution, but with approximation of *t* by normal distribution the accuracy of results is decreased in the verification (table 4). In fact, variability of flow has not always normal distribution (e.g. [18]), and among optimized values of *t* there are discontinuities. SL models yielded strong correlation, but with low amount of precisely described values. Globally they are accordant to the flows, but significantly underestimate peak values (high *RMSE, MAE, RRSE* and *RAE*). They achieve good correlation in the verification and verification outside the historical time series length (additional verification), while AR(1) achieves weaker correlation outside the historical time series length. Coefficients of determination with values of 0.3-0.4 and lower are not satisfying.

Table 4. Coefficient of determination and root mean squared error of models AR(1) and SL, the first approach

Time series length [year]	Model building								Model calibration								Model verification								Verification out. historical time s. l.							
	R^2 (*RMSE*)								R^2 (*RMSE*)								R^2 (*RMSE*)								R^2 (*RMSE*)							
	AR		ANN		SVM		NNM		AR		ANN		SVM		NNM		AR		ANN		SVM		NNM		AR		ANN		SVM		NNM	
65	0.85	(3.96)	0.39	(7.77)	0.38	(7.89)	0.51	(6.97)	0.75	(3.6)	0.35	(5.54)	0.44	(5.14)	0.22	(6.07)	0.49	(7.03)	0.41	(7.51)	0.43	(7.43)	0.33	(8.02)	/	(/)	/	(/)	/	(/)	/	(/)
60	0.84	(4.09)	0.39	(7.88)	0.38	(7.92)	0.52	(6.97)	0.75	(3.71)	0.32	(5.83)	0.39	(5.55)	0.16	(6.5)	0.45	(7.31)	0.45	(7.32)	0.48	(7.11)	0.4	(7.63)	-0.11	(8.48)	0.15	(7.41)	0.19	(7.26)	0.09	(7.67)
55	0.86	(3.72)	0.4	(7.77)	0.37	(7.93)	0.51	(6.99)	0.69	(5.08)	0.43	(6.7)	0.37	(7.04)	0.36	(7.12)	0.22	(7.45)	0.38	(6.51)	0.42	(6.28)	0.4	(6.37)	0	(9.28)	0.4	(7.2)	0.41	(7.13)	0.28	(7.88)
50	0.86	(3.66)	0.41	(7.64)	0.39	(7.76)	0.52	(6.85)	0.72	(5.52)	0.33	(8.2)	0.33	(8.2)	0.35	(8.09)	0.1	(6.89)	0.41	(5.32)	0.47	(5.04)	0.34	(5.63)	0.02	(9.24)	0.39	(7.27)	0.42	(7.11)	0.33	(7.64)
45	0.87	(3.6)	0.45	(7.33)	0.42	(7.47)	0.55	(6.58)	0.75	(5.5)	0.25	(9.15)	0.27	(9.04)	0.27	(9.02)	0.01	(7.41)	0.32	(6.13)	0.37	(5.88)	0.2	(6.64)	-0.01	(9.15)	0.4	(6.98)	0.42	(6.85)	0.37	(7.16)
40	0.86	(3.66)	0.44	(7.32)	0.43	(7.38)	0.52	(6.77)	0.89	(3.66)	0.26	(9.31)	0.27	(9.24)	0.33	(8.28)	0.36	(7.89)	0.36	(7.47)	0.34	(7.6)	0.24	(8.15)	-0.21	(9.53)	0.38	(6.75)	0.41	(6.6)	0.39	(6.7)
35	0.84	(3.95)	0.44	(7.32)	0.43	(7.41)	0.51	(6.88)	0.83	(3.99)	0.37	(7.61)	0.34	(7.8)	0.41	(7.42)	0.35	(9.3)	0.25	(9.6)	0.2	(9.92)	0.23	(9.78)	-0.3	(9.88)	0.36	(6.83)	0.41	(6.6)	0.36	(6.87)
30	0.87	(3.56)	0.4	(7.68)	0.41	(7.63)	0.49	(7.07)	0.83	(3.89)	0.31	(7.88)	0.31	(7.88)	0.31	(7.9)	0.1	(8.97)	0.22	(8.31)	0.21	(8.37)	0.21	(8.38)	-0.09	(9.73)	0.4	(7.14)	0.42	(7.01)	0.39	(7.19)
25	0.86	(3.75)	0.38	(7.93)	0.43	(7.56)	0.5	(7.13)	0.83	(3.83)	0.37	(7.34)	0.39	(7.22)	0.38	(7.26)	0.54	(7.15)	0.27	(8.96)	0.33	(8.59)	0.28	(8.89)	-0.13	(9.83)	0.36	(7.26)	0.39	(7.12)	0.33	(7.47)
20	0.86	(3.7)	0.29	(8.26)	0.43	(7.4)	0.5	(6.93)	0.85	(4.06)	0.24	(9.16)	0.4	(8.11)	0.36	(8.36)	0.58	(6.26)	0.24	(8.37)	0.41	(7.38)	0.37	(7.67)	-0.15	(10.11)	0.26	(8.01)	0.37	(7.37)	0.32	(7.68)
15	0.84	(4.2)	0.39	(8.14)	0.46	(7.65)	0.51	(7.32)	0.65	(3.77)	0.34	(5.19)	0.12	(5.98)	0.16	(5.86)	0.64	(7.23)	0.29	(10.19)	0.45	(8.99)	0.32	(9.96)	-0.22	(10.39)	0.33	(7.57)	0.36	(7.4)	0.31	(7.73)
10	0.75	(5.17)	0.33	(8.45)	0.48	(7.44)	0.42	(7.83)	0.66	(6.27)	0.23	(9.42)	0.23	(9.4)	0.24	(9.34)	0.5	(6.6)	0.36	(7.48)	0.53	(6.41)	0.55	(6.22)	-0.63	(12.09)	0.25	(8.12)	0.35	(7.55)	0.33	(7.65)

For all approaches, modelled and observed flows for historical time series length of 45 years are shown (figures 5, 6 and 7), due to the satisfying accuracy (in the second and the third approach) and possibility of long term planning to 15 years.

Examples where data is missing are represented by value 0. Analysis of results shows that R^2 in the building and calibration part is satisfying only for AR(1) model ($R^2 > 0.65$), while for other models is in the range of medium strength ($R^2 < 0.45$), with exception of slightly greater values at model NNM in the building part ($R^2 < 0.55$). Verification for the most of the models is in the lower range of medium strength of the coefficient of determination ($R^2 < 0.50$), as also the verification outside the used historical time series length for all of the models. Based on mentioned, it can be concluded that this approach is not for recommendation, except with eventual introducing of the improvement by building hybrid models, for example by using singular spectrum analysis [33]. As it is in the work, emphasis is placed on the long term planning, it is needed to use other approaches. Coefficient of determination and the root mean squared error of AR and SL models are given in the table 4. On the Figure 6 the graphical representation of measure

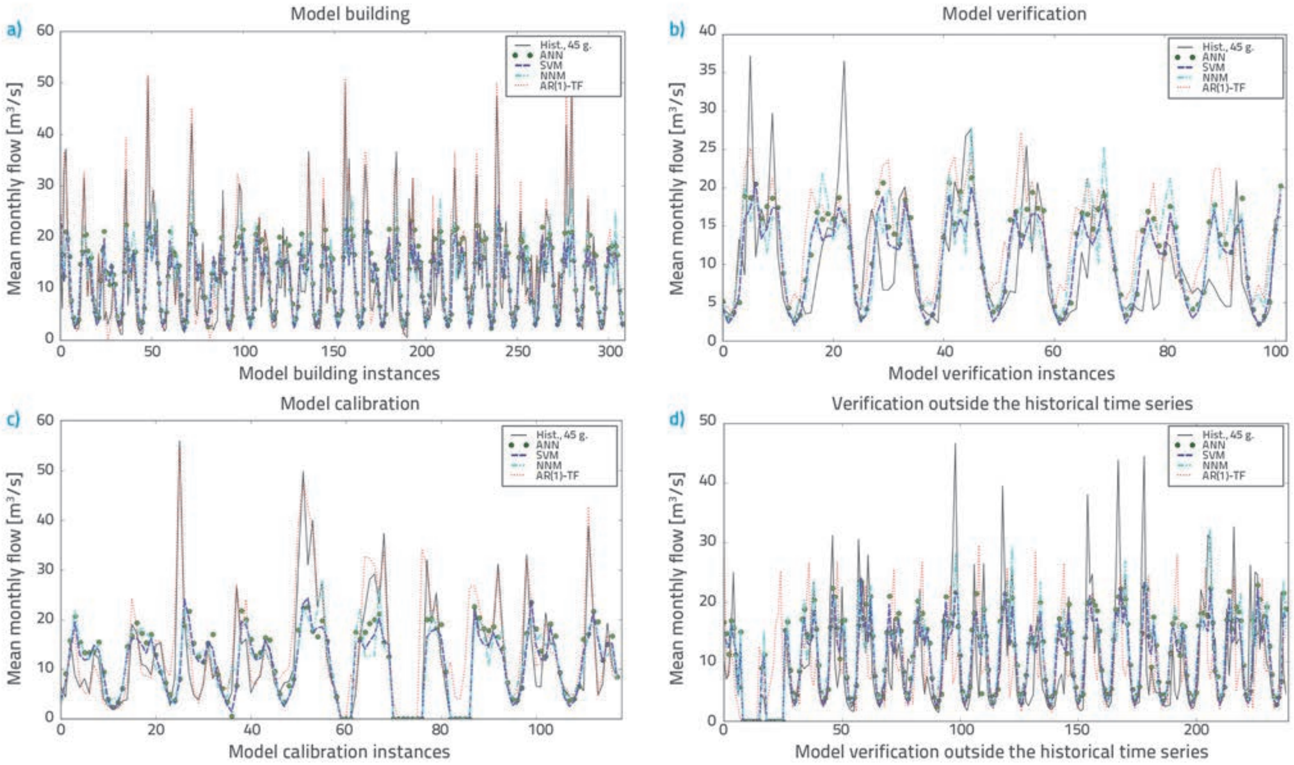

Figure 5. a) building, b) calibration, c) verification and d) verification outside the historical time series length for historical time series length of 45 years, the first approach

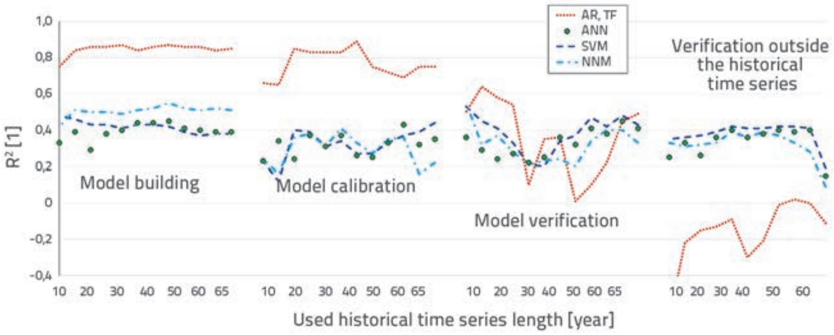

Figure 6. R^2 **on all parts of data in dependence of the used historical time series length, the first approach**

R^2 in dependence of historical time series length for all the approaches is given. At AR(1) care should be payed to parameter *t*, which can be seen on the verification parts. According to the idea of the paper, models should be applied also for long term planning, and it is recommended to pay attention on the error measures on the verification parts. AR(1) does not use external predictors and is not applied in other two approaches. At SL models, NNM describes flows more accurately in the model building part, but accuracy is not preserved in the calibration and verification part. In the second and the third approach, as also in some occasions of the first approach, ANN and NNM give greater accuracy at the model building than SVM. But, at SVM the accuracy is preserved in the calibration and verification part. The most favourable combinations of error measures (greatest values of R^2 and lowest values of *RMSE*) are gained by model SVM, for every time series length. SVM in the second and the third approach shows also the lowest variability of error measures in dependence of time series length. ANN has got great number of options at parameters and architecture choice of the network and it is possible that, by exhaustive research, greater accuracy would be achieved, which can be timely demanding.

4.2. The second approach

In the second approach input variables were changed and models had different parameters than in the first approach. At NNM, by applying the weight distribution depending on the distance between "the neighbours", the flows used for model building are accurately described, while accuracy is reduced for calibration and verification. This is worth to notice, because a model that very closely approximates building data will not necessarily have a good generalization capability on other data. However, in this case, overfitting has not been achieved because equal accuracy has been achieved in calibration and verification (but not also in the model building) with equal weight distribution. The radial basis function kernel gave the highest accuracy of SVM. Correlation of all models, on the model verification and the verification outside the used historical time series length, is in the area of strong correlation ($R^2 > 0.44$), except for an NNM ($R^2 = 0.26$) model that refers to the shortest set of 2-year predictions. It is necessary to emphasize that for time series length from 45 to 55 years R^2 is higher for verification outside the historical time series length than for calibration and verification for all analyzed models. Extreme values are overestimated, or underestimated, for most part of the verification (high *RMSE*, *MAE*, *RRSE*, *RAE*). On the other hand, *RMSE* values are in the range of 4.9-7.05 m³/s, indicating a significant increase in accuracy compared to the first approach where it was in range of 6.6-12.09 m³/s. The ability of all models to globally describe the nature of the flow is high, even outside the used historical time series length. For models with smaller historical time series lengths, not

Table 5. Coefficient of determination and root mean squared error of SL models, the second approach

Time series length [year]	Model building						Model calibration						Model verification						Verification out. historical time s. l.					
	R^2 (*RMSE*)						R^2 (*RMSE*)						R^2 (*RMSE*)						R^2 (*RMSE*)					
	ANN		SVM		NNM		ANN		SVM		NNM		ANN		SVM		NNM		ANN		SVM		NNM	
62	0.76	(4.88)	0.76	(4.88)	1.00	(0)	0.56	(4.58)	0.68	(3.89)	0.61	(4.32)	0.69	(5.53)	0.75	(5)	0.61	(6.25)	/	(/)	/	(/)	/	(/)
60	0.73	(5.17)	0.76	(4.91)	1.00	(0)	0.56	(4.66)	0.70	(3.83)	0.65	(4.12)	0.68	(5.46)	0.75	(4.83)	0.62	(5.96)	0.45	(6.74)	0.57	(5.94)	0.26	(7.78)
55	0.83	(4.14)	0.77	(4.81)	1.00	(0)	0.62	(4.46)	0.64	(4.32)	0.57	(4.72)	0.68	(4.96)	0.70	(4.88)	0.59	(5.65)	0.72	(5.31)	0.76	(4.91)	0.62	(6.21)
50	0.84	(4.02)	0.78	(4.74)	1.00	(0)	0.65	(4.97)	0.66	(4.91)	0.61	(5.24)	0.55	(4.48)	0.64	(3.96)	0.53	(4.55)	0.72	(5.41)	0.75	(5.11)	0.62	(6.28)
45	0.85	(3.86)	0.79	(4.54)	1.00	(0)	0.68	(5.89)	0.66	(6.03)	0.55	(6.97)	0.47	(4.55)	0.63	(3.8)	0.55	(4.21)	0.68	(5.38)	0.73	(4.91)	0.58	(6.09)
40	0.82	(4.17)	0.79	(4.5)	1.00	(0)	0.67	(6.16)	0.71	(5.78)	0.57	(7.09)	0.66	(4.74)	0.66	(4.78)	0.60	(5.17)	0.63	(5.36)	0.69	(4.9)	0.54	(5.96)
35	0.80	(4.4)	0.81	(4.34)	1.00	(0)	0.59	(5.79)	0.65	(5.36)	0.62	(5.62)	0.71	(5.95)	0.70	(6.01)	0.52	(7.58)	0.59	(5.48)	0.66	(5.01)	0.54	(5.8)
30	0.76	(4.96)	0.83	(4.2)	1.00	(0)	0.60	(5.61)	0.63	(5.39)	0.57	(5.82)	0.58	(7.2)	0.70	(6.08)	0.56	(7.43)	0.57	(5.75)	0.64	(5.22)	0.55	(5.89)
25	0.89	(3.28)	0.83	(4.13)	1.00	(0)	0.65	(5.53)	0.67	(5.31)	0.60	(5.92)	0.63	(5.65)	0.67	(5.31)	0.63	(5.62)	0.61	(5.81)	0.67	(5.38)	0.56	(6.19)
20	0.89	(3.16)	0.82	(4.15)	1.00	(0)	0.62	(6.06)	0.69	(5.51)	0.52	(6.79)	0.70	(5.68)	0.76	(5.04)	0.58	(6.68)	0.62	(5.71)	0.67	(5.35)	0.58	(6.06)
15	0.92	(2.87)	0.83	(4.1)	1.00	(0)	0.64	(5.54)	0.72	(4.88)	0.62	(5.68)	0.73	(6.01)	0.75	(5.75)	0.63	(7.02)	0.59	(5.95)	0.67	(5.35)	0.56	(6.22)
10	0.96	(1.82)	0.75	(4.55)	1.00	(0)	0.11	(7.95)	0.44	(6.29)	0.34	(6.87)	0.67	(7.13)	0.77	(6.01)	0.54	(8.44)	0.44	(7.05)	0.67	(5.46)	0.50	(6.67)

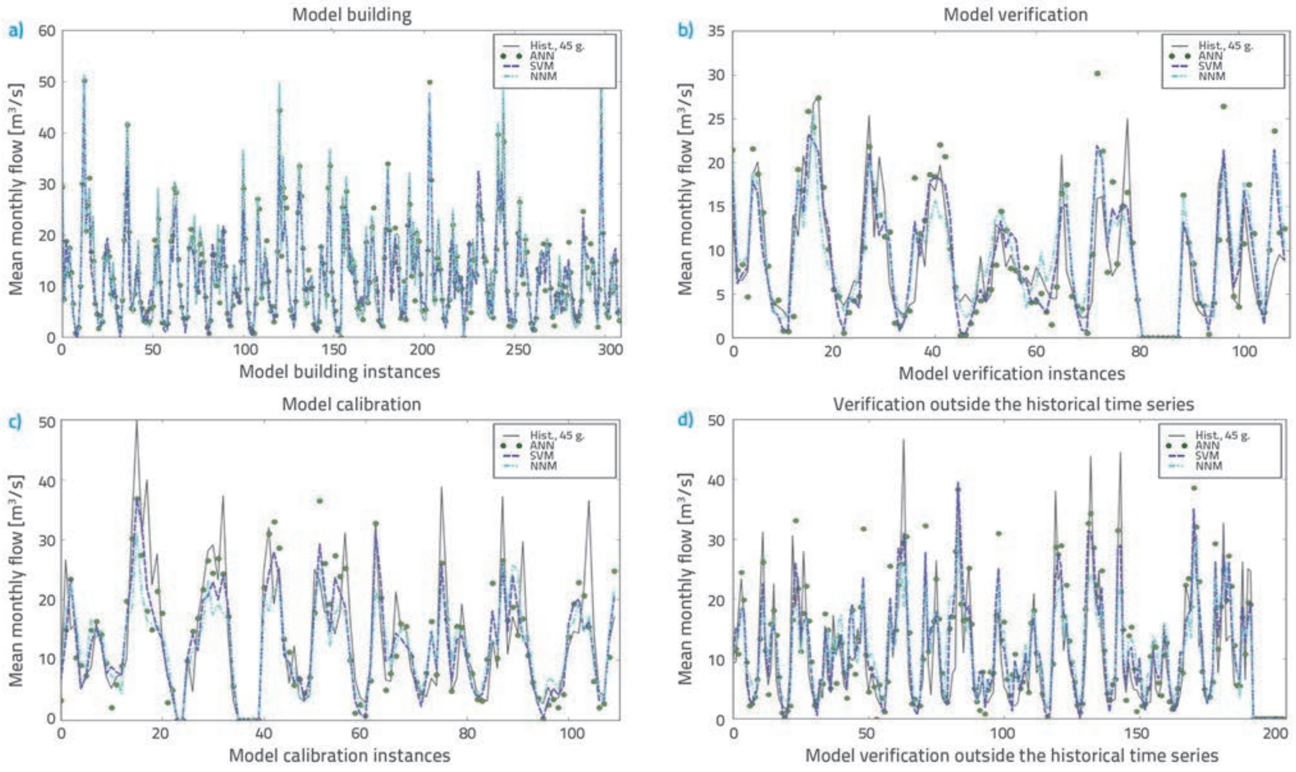

Figure 7. a) building, b) calibration, c) verification and d) verification outside the historical time series length for historical time series length of 45 years, the second approach

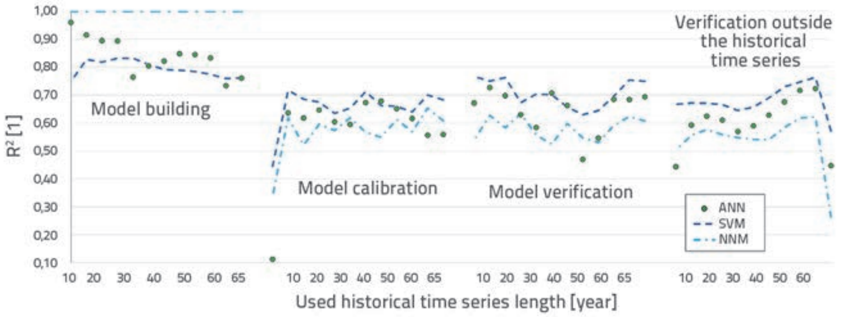

Figure 8. R^2 on all parts of data in dependence of the used historical time series length, the second approach

much data is needed to reconstruct the nature of the flow (65 examples for the model building with a length of 10 years). The most favourable error measures were established with the SVM model for all time series lengths. It is assumed that by adding data to the two remaining stations (in the third approach) higher quality models could be built. Coefficient of determination and the root mean squared error are given in the table 5.

In the second approach, according to the R^2 values, the time series of 40-60 years result with greater accuracy (R^2> 0.7), while the shorter sequences do not result in significantly lower values (0.65 < R^2 < 0.7). Based on the results, the following can be determined: the key part in the use of the SL for forecasting the flow is selection of the predictor, it is possible to determine undocumented flows based on the prediction using the precipitation data (mostly) and temperature data with such accuracy that quantitative availability of water in time can be determined. In the third approach, for verification outside the used historical time series length, R^2 for SVM > 40 years exceeds 0.8, and for 20-40 years it is in the range 0.7-0.8. Modelled and observed flows for the historical time series length of 45 years is given on the figure 7. Graphical representation of measure R^2 in dependence on series length is given on the figure 8.

4.3. The third approach

In the third approach the model precision has been increased according to all statistical measures. When it comes to ANN, the rectification function was shown to be the best activation function. NNM model shows similar functionality as in the second approach. Also, ANN for 10-20 years has produced the perfect accuracy in model building, but significantly reduced in other parts. In the case of ANN, it is a result of overfitting and with the reduction of instances number for model building, additional energy should be used to find the appropriate architecture network. The SVM is the most accurate and shows the ability to maintain error rates low on all parts of the data. Correlation on the verification for SVM outside the time series length is in

Table 6. Coefficient of determination and root mean squared error of SL models, the third approach

Time series length [year]	Model building						Model calibration						Model verification						Verification out. historical time s. l.					
	R^2 (*RMSE*)						R^2 (*RMSE*)						R^2 (*RMSE*)						R^2 (*RMSE*)					
	ANN		SVM		NNM		ANN		SVM		NNM		ANN		SVM		NNM		ANN		SVM		NNM	
60	0.73	(5.08)	0.81	(4.27)	1.00	(0)	0.71	(3.61)	0.75	(3.32)	0.67	(3.82)	0.78	(4.76)	0.83	(4.19)	0.68	(5.7)	/	(/)	/	(/)	/	(/)
55	0.90	(3.16)	0.82	(4.27)	1.00	(0)	0.63	(4.25)	0.73	(3.61)	0.68	(3.96)	0.76	(4.89)	0.83	(4.12)	0.69	(5.56)	0.77	(4.04)	0.80	(3.82)	0.58	(5.52)
50	0.94	(2.5)	0.82	(4.23)	1.00	(0)	0.62	(4.79)	0.71	(4.13)	0.65	(4.59)	0.69	(4.81)	0.75	(4.33)	0.65	(5.08)	0.71	(5.21)	0.83	(3.93)	0.68	(5.45)
45	0.86	(3.77)	0.82	(4.16)	1.00	(0)	0.73	(4.87)	0.72	(4.97)	0.61	(5.84)	0.65	(4)	0.71	(3.6)	0.65	(3.99)	0.78	(4.45)	0.81	(4.19)	0.63	(5.81)
40	0.89	(3.23)	0.81	(4.2)	1.00	(0)	0.72	(5.52)	0.67	(6.02)	0.59	(6.71)	0.64	(4.14)	0.71	(3.7)	0.67	(3.96)	0.74	(4.69)	0.78	(4.33)	0.66	(5.36)
35	0.86	(3.64)	0.81	(4.26)	1.00	(0)	0.73	(5.71)	0.69	(6.06)	0.60	(6.95)	0.69	(4.92)	0.65	(5.18)	0.60	(5.54)	0.76	(4.25)	0.72	(4.59)	0.64	(5.19)
30	0.76	(4.72)	0.85	(3.7)	1.00	(0)	0.53	(7)	0.57	(6.68)	0.54	(6.92)	0.67	(6.09)	0.70	(5.79)	0.59	(6.83)	0.71	(4.65)	0.73	(4.52)	0.63	(5.22)
25	0.99	(0.76)	0.87	(3.53)	1.00	(0)	0.24	(9.25)	0.61	(6.58)	0.51	(7.43)	0.55	(5.69)	0.55	(5.67)	0.52	(5.86)	0.58	(5.99)	0.73	(4.81)	0.66	(5.41)
20	1.00	(0.24)	0.87	(3.5)	1.00	(0)	0.52	(6.84)	0.76	(4.85)	0.63	(6.04)	0.27	(8.54)	0.53	(6.88)	0.47	(7.3)	0.54	(6.21)	0.71	(4.96)	0.63	(5.55)
15	1.00	(0.02)	0.89	(3.35)	1.00	(0)	0.55	(6.06)	0.72	(4.76)	0.71	(4.84)	0.11	(8.53)	0.74	(4.58)	0.74	(4.65)	0.40	(7.29)	0.69	(5.26)	0.61	(5.85)
10	1.00	(0)	0.88	(3.07)	1.00	(0)	0.62	(5.44)	0.70	(4.82)	0.60	(5.59)	0.69	(7.36)	0.76	(6.47)	0.42	(10.15)	0.47	(6.78)	0.67	(5.36)	0.52	(6.5)

all cases equal to or greater than 0,82. *RMSE* and *MAE*, ranging from 3.82-5.36 m^3/s and 2.95-4.02 m/s, respectively; *RRSE* and *RAE*, ranging from 0.45-0.57 and 0.42-0.55, respectively, are the smallest, and R^2 ranging between 0.67-0.83 is the largest. It should also be noted that R^2 for all time series lengths is greater in the range of verification outside the used historical time series length than in the case of calibration and verification for all models (the only exceptions are SVM and NNM models for the shortest 5-year prediction). On the Figure 9 it can be seen that: all models generally follow the observed flow, NNM is the weakest for description of the flow rate variability, and that ANN underestimates the minimums and maximums. SVM shows the greatest tendency to generalize, but all models fail to reach the local maximums.

In the third approach SVM is the most suitable for long-term water availability analysis. Regardless, it is advisable to conduct sensitivity analysis of the model on the time series length. SVM is more stable than ANN and NNM in it. Therefore, with regard to accuracy and

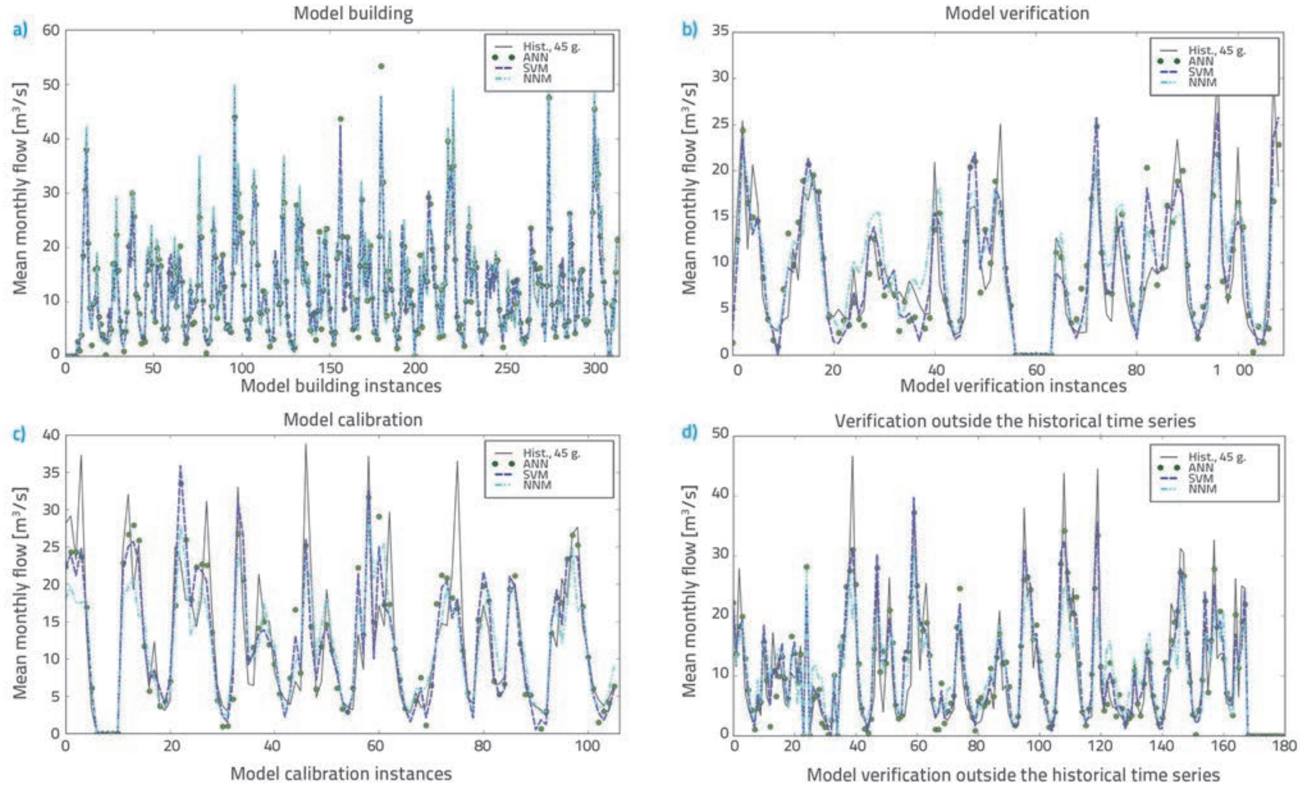

Figure 9. a) building, b) calibration, c) verification and d) verification outside the historical time series length for historical time series length of 45 years, the third approach

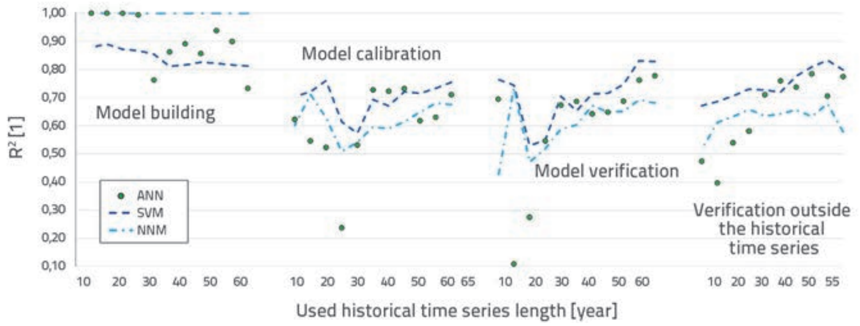

Figure 10. R^2 **on all parts of data in dependence of the used historical time series length, the third approach**

stability, SVM may be recommended for further use. Modelled and observed flows for historical time series length of 45 years are shown on the Figure 9. Graphical representation of measures R^2 in dependence of time series length for the third approach is given on the Figure 10. Additional inclusion of data from nearby stations would certainly increase the precision of the model. However, the goal of machine learning is to build a good model with as few input variables as possible, mimicking realistic situations where significant number of nearby stations is rarely available. Including the number of days in a month with a certain amount of precipitation can also contribute to the precision. After determining the best configuration of the model, improvements can be made by spectral analysis, wavelet based methods, chaos analysis, phase reconstruction of space, etc. (see [12, 13, 29]).

4.4. Statistical analysis of results

Descriptive statistics of the model results are calculated and are given in the table 7. In the first approach significant underestimation of the maximum values of the ANN, SVM and NNM models is noticeable, while AR(1) shows overall minor

Table 7. Descriptive statistics of the model results and observed values

	The first approach					The second approach				The third approach			
Building	Obs.	ANN	SVM	NNM	AR1	Obs.	ANN	SVM	NNM	Obs.	ANN	SVM	NNM
Mean value	12.66	12.61	11.56	12.11	13.89	12.56	12.57	12.00	12.56	12.74	12.79	12.37	12.74
Minimum	0.56	2.20	1.86	2.20	-0.15	0.56	-0.14	-0.94	0.56	0.56	0.23	0.79	0.56
Maximum	51.02	23.38	26.20	29.53	51.53	55.94	51.27	51.12	55.94	55.94	51.81	47.65	55.94
Skewness	1.24	-0.18	-0.07	0.32	1.20	1.44	1.18	1.21	1.44	1.44	1.29	1.32	1.44
Kurtosis	1.62	-1.36	-1.10	-0.67	1.25	2.60	1.90	2.56	2.60	2.48	2.17	2.30	2.48
Calibration													
Mean value	13.87	13.33	12.35	12.77	14.99	14.52	14.21	13.47	13.15	11.31	10.91	11.33	11.39
Minimum	1.84	0.69	1.42	2.65	1.83	2.73	-0.41	0.35	2.02	1.59	-2.61	-0.04	2.38
Maximum	55.94	23.42	24.29	27.78	54.16	49.84	39.92	36.72	31.06	38.81	34.61	35.86	29.89
Skewness	1.52	-0.39	-0.19	0.08	1.30	1.04	0.47	0.51	0.21	1.16	0.54	0.65	0.34
Kurtosis	2.58	-1.01	-0.75	-0.48	1.43	0.55	-0.60	-0.24	-0.62	0.81	-0.47	-0.33	-0.58
Verification													
Mean value	10.30	11.66	10.72	11.64	13.37	9.36	9.98	9.76	10.21	9.35	9.17	9.80	10.33
Minimum	2.22	2.01	2.00	2.63	3.67	2.22	-1.03	0.44	2.19	1.59	-2.61	-0.04	2.52
Maximum	37.14	21.26	20.64	27.78	27.46	27.63	30.53	23.23	25.55	31.20	23.25	26.24	21.45
Skewness	1.38	-0.22	-0.21	0.24	0.26	1.03	0.61	0.31	0.44	1.20	0.34	0.64	0.18
Kurtosis	2.00	-1.46	-1.33	-0.66	-0.93	0.28	-0.51	-0.93	-0.50	0.90	-0.86	-0.38	-0.97
Verification outside the historical time series length													
Mean value	10.70	12.17	11.24	12.17	12.74	10.99	12.51	12.06	11.97	10.94	11.67	11.78	11.43
Minimum	1.43	2.51	2.36	2.25	1.65	1.43	-1.52	-1.22	1.90	1.43	-1.18	0.09	2.25
Maximum	46.65	23.61	24.00	32.31	33.21	46.65	42.31	39.38	31.45	46.65	33.98	39.72	30.84
Skewness	1.51	-0.08	0.00	0.24	0.52	1.44	0.67	0.73	0.46	1.51	0.60	0.88	0.47
Kurtosis	2.24	-1.39	-1.15	-0.48	-0.22	1.84	-0.26	0.09	-0.43	2.07	-0.47	0.24	-0.43

deviation and a higher average flow rate. In the second and the third approach deviations are significantly reduced. In the second approach ANN shows the smallest deviations, while in the third approach SVM shows the smallest deviations. Attention should be also paid to the possible occurrence of negative flow values in ANN and SVM, although they are negligible in the most accurate model (SVM, third approach). For future research solving these problems by optimizing model parameters is suggested.

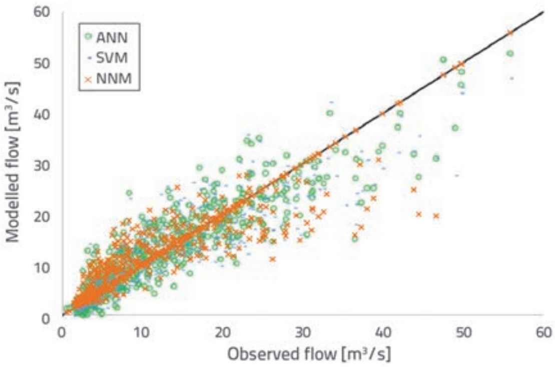

Figure 11. Scatter plot of modelled values, the third approach

Scatter plot of modelled values related to the observed values for the third approach is shown on the Figure 11. Values from all parts of data for ANN, SVM and NNM models are separately presented with unique markers. The largest scatter is observed for the NNM model, except for the model building, whose values completely align to the line of the perfect agreement. By increasing the measured flow rate, NNM significantly underestimates the predicted values. Less dispersion can be seen at the SVM and ANN models, but these models tend to underestimate higher flow rates. Therefore, for future research, it is recommended to calculate the model's reliability intervals and to integrate these values into model results, for example by using quantile regression (e.g. [34]).

5. Conclusion

The paper analyzes the possibility of predicting the mean monthly flow for the purpose of long-term prediction and planning in solving problems related to water availability. Three different approaches were used in which three SL methods were compared, with addition of the stochastic method in the first approach. In the first approach, SL was able to describe the general nature of the flow but with significant deviations in the area of extreme values. AR is capable to replicate full variability of the flow if proper attention is payed to methods used for quantifying flow variability. In order to use SL, more complex models and/or more informative input data must be selected. In the second and the third approach causality of input and predicted variables was described better with SL. The application of precipitation and temperatures for flow forecasting is favourable because of possibility to use projections from climatic models, which cannot be implemented in the first approach. It is generally valid that with a larger amount of data used to build an SL model, greater accuracy and precision are achieved. But the length of the time series does not necessarily reflect quality of built model. The precision of the SVM with the determination coefficient in range of 0.7-0.8 for the 20-40 years length of the time series is satisfactory, whereas for 10 years (the determination coefficient 0.67) the precision is not significantly lower. The recommendation for further research is to focus on the additional elaboration of the input selection variables methodology so that the available data is used more efficiently.

REFERENCES

[1] Parry, M.: Food and Energy Security: Exploring The Challenges of Attaining Secure and Sustainable Supplies of Food and Energy, FOOD AND ENERGY SECURITY, (2012) 1, pp. 1-2

[2] UN (United Nations): World Population Prospects, 2015 Revision Population Database, 2015

[3] Marton, D., Menšik, P.P., Stary, M.: Using Predictive Model for Strategic Control of Multi-reservoir System Storage Capacity, 13th Computer Control for Water Industry Conference, PROCEDIA ENGINEERING, (2015) 112, pp. 994-1002

[4] IPCC (Intergovernmental Panel on Climate Change): Climate Change 2014, Synthesis Report, Geneva, Švicarska, 2015,

[5] Simonović, S.P.P.: Floods in a Changing Climate, International Hydrology Series, UNESCO, Cambridge, SAD, 2012.

[6] Rubinić, J., Margeta, J.: Dimenzioniranje akumulacija primjenom generiranih protoka, GRAĐEVINAR, 53 (2001) 1, pp. 17-23

[7] Haykin, S.: Neural Networks and Learning Machines, 3, izdanje, Upper Saddle River, New Jersey, SAD, 2013,

[8] Govindaraju, R.S., Rao, A.R.: Artificial Neural Networks in Hydrology, Springer Science & Business Media, 2000.

[9] Abrahart, R., Kneale, P.P.E., See, L.M.: Neural Networks for Hydrological Modelling, CRC Press, 2004

[10] Cigizoglu, H.K.: Generalized regression neural network in monthly flow forecasting, CIV. ENG. ENVIRON, SYST, 22 (2005) 2, pp. 71-84

[11] Nillson, P.P., Uvo, C.B., Berndtsson, R.: Monthly Runoff Simulation: Comparing and combining conceptual and neural network models, J HYDROL, 321 (2006), pp. 344-363

[12] Wu, C.L., Chau, K.W., Li, Y.S.; River stage prediction based on a distributed support vector regression, J HYDROL, 358 (2008), pp. 96-111

[13] Guo, J., Zhou, J., Qin, H., Zou, Q., Li, Q.: Monthly streamflow forecasting based on improved support vector machine model,EXPERT SYS APPL, 38 (2011), pp.13073-13081

[14] Akiner, M.E., Akkoyunlu, A.: Modeling and forecasting river flow rate from the Melen Watershed, Turkey, J HYDROL, 456-457 (2012), pp. 121-129

[15] Farajzadeh, J., Fard, A.F., Lotfi, S.: Modeling of monthly rainfall and runoff of Urmia lake basin using "feed-forward neural network" and "time series analysis" model, WATER RESOURCES AND INDUSTRY, 7-8 (2014), pp. 38-48

[16] Terzi, O,: A genetic programming approach to river flow modeling, J INTELL FUZZY SYST, 27 (2014), pp. 2211-2219

[17] Matić, P.P.: Kratkoročno predviđanje hidrološkog dotoka pomoću umjetne neuronske mreže, Fakultet elektrotehnike, strojarstva i brodogradnje, Sveučilište u Splitu, Hrvatska, 2014

[18] Loucks, D.P.P., Van Beek, E.: Water Resources Systems Planning and Management, UNESCO, Paris, Francuska, 2005.

[19] Karamouz, M., Szidarovszky, F., Zahraie, B.: Water Resources Systems Analysis with emphasis on Conflict Resolution, Lewis Publishers, Boca Raton, SAD, 2003.

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.P., Weiss, R., Duborg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python, J MACH LEARN RES, 12 (2011), pp. 2825-2830

[21] Russel, R., Norvig, P.P.: Artificial Intelligence: A Modern Approach, 3rd Edition, Prentice Hall, SAD, 2010.

[22] Mitchell, T.M.: Machine learning, McGraw Hill Inc,, New York, SAD, 1997.

[23] Kingma, D.P.P., Ba, J.L.: Adam: A Method For Stochastic Optimization, 3rd International Conference for Learning Representations, San Diego, 2015.

[24] Ng, A,: Lectures: Machine Learning, Stanford University, https://www,coursera,org/learn/machine-learning

[25] MacKay, D.J.C.: Information Theory, Inference and Learning Algorithms, Cambridge University Press, Cambridge.

[26] Raghavendra, N.S., Deka, P.P.C.: Support vector machine application in the field of hydrology: A review, APPL SOFT COMPUT, 19 (2014), pp. 372-386.

[27] Marsland, S.: Machine Learning, An Algorithmic Perspective, Chapman & Hall, Boca Raton, SAD, 2015,

[28] Smola, A.J., Schölkopf, B.: A tutorial on support vector regression, STAT COMPUT, 14 (2004), pp. 199-222, https://alex,smola,org/papers/2004/SmoSch04,pdf [dostupno 07.04.2017,]

[29] Python: Scikit-learn user guide, release 0,17, 2015.

[30] Državni hidrometeorološki zavod: Hidrološka baza podataka, HIS 2000, 2017.

[31] Državni hidrometeorološki zavod: Relacijska meteorološka baza podataka, 2017.

[32] Šošić, I., Serdar, V.: Uvod u statistiku, Školska knjiga, Zagreb, 1992,

[33] Latifoğlu, L., Kişi, Ö., Latifoğlu, F.: Importance of hybrid models for forecasting of hydrological variable, NEURAL COMPUT APPLIC, 16 (2015), pp. 1669-1680.

[34] Dogulu, N., López López, P., Solomatine, D.P., Weerts, A.H., Shrestha, D.L.: Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments, HYDROL. EARTH SYST. SCI, 19 (2015), pp. 3181-3201.