ACTA HYDROLOGICA
SLOVACA

**Ročník 18, č. 2, 2017, 282 – 289**

# LONG TERM PREDICTION OF INFLOW USING THE SUPERVISED LEARNING

Jadran Berbić, Eva Ocvirk

For the purposes of long-term (on time basis of one month) planning and management of water resources systems, long term prediction of inflow is needed. In the last two decades, usage of machine learning is becoming popular in the field of water resources systems management, whether for real time, short term, mid term or long term predictions of hydrological variables. Especially interesting is the usage of supervised learning, defined as the type of machine learning used for model development based on given data, which enables prediction and extrapolation on so far unseen examples. Supervised learning models are able to use arbitrarily huge amount of variables for model development and forecasting. Mentioned facts make the problematics interesting from the climate change point of view, as also from the view of model development for assistance in water resources systems planning and management. By using the models developed on historical data it is able to predict inflows in conditions of future scenario from climate models and get insight in future hydrological conditions and systems management efficiency. Besides classicaly used rainfall for runoff modelling, other meteorological variables, if are on disposition, could be included. As searching for appropriate way of prediction and supervised learning models architecture is not an easy task, necessary step is reviewing the literature about long-term predictions. Therefore, the insight in previous research of long term prediction of inflow using the supervised learning is given in the paper. In the literature review are also included researches with considerations of climate change influence on water resources systems, based on the predictions using the supervised learning.

KEY WORDS: supervised learning, long-term prediction, water resources systems, climate change

**DLHODOBÁ PREDPOVEĎ PRÍTOKU POUŽITÍM SUPERVÍZOVANÉHO UČENIA**. Na účely dlhodobého plánovania a riadenia systémov vodných zdrojov je potrebná dlhodobá predpoveď prítoku. V posledných dvoch desaťročiach sa využívanie strojového učenia stáva populárnym v oblasti riadenia vodných zdrojov, či už v reálnom čase, krátkodobé, strednodobé alebo dlhodobé predpovede hydrologických premenných. Zvlášť zaujímavé je použitie supervízovaného učenia, definovaného ako typ strojového učenia, ktoré sa používa na vývoj modelov na základe daných údajov, čo umožňuje predpovedanie a extrapoláciu na doteraz neviditeľných príkladoch. Supervízované modely učenia sú schopné ľubovoľne používať obrovské množstvo premenných pre vývoj a predpovedanie modelov. Uvedené fakty sú zaujímavé z hľadiska klimatických zmien, ako aj z hľadiska modelového rozvoja pomoci pri plánovaní a riadení vodných zdrojov. Použitím modelov vyvinutých na základe historických údajov je možné predpovedať prítoky v podmienkach budúceho scenára z klimatických modelov a získať prehľad o budúcich hydrologických podmienkach a efektívnosti riadenia systémov. Okrem klasicky používaných zrážok na modelovanie odtoku môžu byť zahrnuté aj iné meteorologické premenné, ak sú k dispozícií. Keďže hľadanie vhodného spôsobu predpovede a supervízovaných modelov učenia nie je jednoduché, potrebným krokom je preskúmanie literatúry o dlhodobých predpovediach. Z tohto dôvodu je v práci uvedený podrobný pohľad na dlhodobú predpoveď prítoku pomocou supervízovaného učenia. V prehľade literatúry sú zahrnuté aj výskumy zaoberajúce sa vplyvom klimatickej zmeny na systémy vodných zdrojov, založené na predpovediach použitím supervízovaného učenia.

KĽÚČOVÉ SLOVÁ: supervízované učenie, dlhodobá predpoveď, systémy vodných zdrojov, zmena klímy

## Introduction

Machine learning (ML) is the part of computer science, that is, artificial intelligence, created from endeavor of learning computers to recognize and extrapolate data patterns by programming. Supervised learning (SL) is one of the basic parts of machine learning. SL is able to, on the basis of given data (inputs and outputs) and an assumed hypothesis, find such hypothesis (function) parameters which generalize on unseen example as well as possible. It is used for problems of classification and regression. Some examples for SL methods are artificial neural networks (ANN), support vector machines (SVM), nearest neihgbours methods (NNM), fuzzy logics (FL), Bayesian method (BM) etc. As hydrological variables have continuous range of values (they are not classes), the focus is put on the regression part. The nature of SL makes it appropriate for application on hydrologicaly studied basins, with developed network of hydrological stations, as also with meteorological, climatological stations and rain gauges.

The first phase of the research provided conclusion that SL is used for real-time, short term, mid term and long term prediction of hydrological quantities. Generally, real-time and short term predictions are mostly used as they provide greatest amounts of data for model building. And as it can be concluded, for models to learn patterns in data, it is always more appropriate to have greater amount of data. Long term predictions are more rarely done. While the predictions on smaller lead time (1h or 1 day) are of great importance for decision making in the operational purposes, greater lead times are important for long term planning purposes. For example, long term planning and providing of long term policy of water reservoirs is usually done on monthly basis. Providing information about the monthly inflow for reservoir is of great importance for operation strategy Cigizoglu (2005). Here also comes the next important classification about SL: predictions are usually done in the manner: hour-by-hour, day-by-day, month-by-month or any other lead time. That means, i.e., that for the prediction of inflow in the next hour, **observed** flow in the current hour is used (if it is the part of the model configuration). When someone would try to use the built model for next several hours prediction, by using already predicted values of flows, significant amount of the error would be generated. The same thing would happen with any other used lead time. So, when it is about long term prediction in the month-by-month manner, it should be differed from the long term planning. Long term planning of water resources systems is made for several years or even decades ahead, depending on the lifetime and purpose of the system. This is quite important because it leads us to the SL classification by principal way of data usage: using endogeneous or exogeneous variables. Endogeneous means that output variable is predicted using the input variables of the same type (i.e. prediction of flow using flows). Exogeneous means that output variable is predicted using physically different input variables (i.e. prediction of flow by using precipitation and temperature). So, generally it is hard task to implement usage of endogeneous approach for long term planning, but is much easier task to implement it for long term month-by-month prediction. Endogeneous approach can be used for long term planning if timely averaged and periodic variables are used, similar to the approach of synthetic streamflows generation. The focus of the paper is to review the usage of SL for month-by-month prediction and long term planning, either by endogeneous or exogeneous variables, and provide valuable conclusion from literature. The review is the framework for attaining the higher goal – building the models which use climatological data (mainly precipitation and temperature) and which could be used for long term predictions and planning. Also, such models could then be used for modelling of the hydrological conditions under different climatic scenarios as a support for decision making.

## Review of the literature

### *Principles of SL*

The general procedure of applying of SL technique (as also and ML) is consisted of Šmuc (2013): choosing model representation, search algorithm and error estimators, as all of those are covered in the review. Model representation is simply the SL technique as itself – ANN, SVM, NNM, etc. Representations are consisted of function and their parameters. Search algorithm is the optimization algorithm used for searching of the best function parameters, that is, parameters which generalize calibration part of the data with the smallest error value. Error estimators are statistical error measures which describe the accuracy of modelled data by comparing it with observed data.

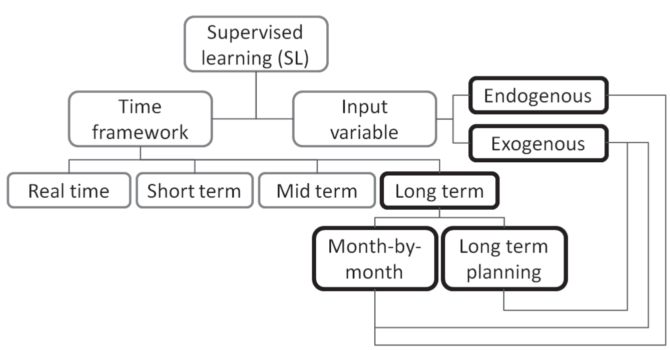


*Fig. 1. Application of supervised learning based on time framework and input variables.*
*Obr. 1. Aplikácia supervízovaného učenia na základe časového rámca a vstupných premenných.*

From different researches and experience it can be concluded that model configuration plays the most important role in achieving accurate results. Searching for the appropriate model configuration can be time exhausting. By model configuration it is meant: model representation with its parts and used input and output variables. I.e. ANNs are consisted of three or more layers with its nodes. Nodes in the first layer are input variables (predictors), while nodes in the last layer are output (predicted) variables (usually 1 output variable). Hidden layers are nodes with its activation functions (sigmoid function, radial basis function etc.). Along with configuration, learning parameters also play important role in the model training process.

Data is usually split into two or three non-overlapping parts: training and testing or training, calibration and verification parts. General recommendation is to split it in the ratio 65:35, that is, 60:20:20, but those values vary as can be seen on figures 2 and 3. Most of the researches use the first principle, while the second principle is more appropriate. It is because satisfactory model performance means that performance on all parts should be similar, based on statistical error measures. A well behaviour of the model on the training part, and a lack of good performance on other parts, probably means that model failed to generalize (due to overfitting or bias). Performance is checked by using various statistical error measures, as can be seen on figures 2, 4 and 5. Among all of those, correlation coefficient, root mean squared error, mean absolute error and particularly coefficient of determination as score percentage are the most popular.

Search algorithms are basically optimization algorithms used for searching of the best function parameters in SL. Optimization algorithms may be used for improving of the SL models performance. For training of the ANN, improvements made by Levenberg-Marquardt algorithm are highly popular (Latifoglu et al., 2015; Okkan and Fistikoglu, 2014; Farajzadeh et al., 2014; Saghafian et al., 2013; Akiner and Akkoyunlu, 2012; Guo et al., 2011; Wu and Chau, 2010), while basically used algorithm is the backpropagation algorithm. The procedure can additionally be improved by using of a genetic algorithm for searching of an appropriate ANN architecture (Wu and Chau, 2010). In fact, any search algorithms can be applied for parameter searching in SL methods or for finding of an appropriate configuration of model.

***Possible improvements for the purpose of long term planning***

As it is already mentioned in the introductory part, for the long term planning, usage of exogeneous input variables is unavoidable. Most of the researches of SL usage are related to month-by-month predictions (Latifoglu et al., 2015; Liu et al., 2015; Pandhiani and Shabri, 2015; Farajzadeh et al., 2014; Terzi, O., 2014; Saghafian et al., 2013; Guo et al., 2011; Wu and Chau, 2010; Nilsson et al., 2006; Cigizoglu, H.K., 2005;Jain et al., 1999). Of those researches, most of them are used for 1 month ahead, while one is used for 1-6 months (Latifoglu et al., 2015) and two for 1-3 months predictions (Liu et al., 2015; Saghafian et al., 2013). Smaller amount of researches is used for the purpose of long term planning. Okkan and Fistikoglu (2014) used ANN for downscaling of precipitation and air temperature from global climate models, while monthly runoff was estimated by using of a conceptual hydrological model GR2M (Okkan and Fistikoglu, 2014). Research provided by Akiner and Akkoyunlu (2012) included usage of ANN for the purpose of missing values interpolation and prediction of rainfall, while runoff was estimate by using of SWAT model (Akiner and Akkoyunlu, 2012). Nilsson et al (2006) estimated an influence of different climatic variables on monthly runoff predictions with ANN, as also with a combination of ANN and a conceptual hydrological model (Nilsson et al., 2006). Also, most of reviewed researches use endogeneous input variables for predicttion or even a combination of endogeneous and exogeneous input variables. For the purpose of long term planning, building of SL models with exogeneous variables is necessary, especially if there is an intention to provide long term policy of water resources systens under different climate scenarios. Thus, providing of SL models with rainfall and air temperature as input variables, and flow (that is, water availability or other hydrological quantity) as output variable is of main concern for future researchers in the field of water resources management. It is also noticed that number of instances for model building varies over different resarches. General recommendation in hydrology is to use at least 30 years of historical data for modeling purposes and analysis. As ML generally provides more accurate results when larger amounts of data are used, one of the main concern is to analyse how the length of historical series influence the SL models ability to generalize. Resarches of this kind should provide answers like: which are the minimum lengths of historical series needed for building of quality models, and do models have ability to generalize outside of used historical series and for how much in the future, which SL models are most stable in dependence of historical series length. After fulfilling this space for improvement, basis for long term planning by using of SL models will be achieved.

Space for additional improvements is the usage of different decomposition methods, where techniques like Fourier analysis, wavelets, phase space reconstruction, singular spectrum analysis etc. can be used. Latifoglu et al. (2015) used singular spectrum analysis combined with ANN. After the usual model building by using of the ANN, they divided flow data into 12 independent subcomponents and used them as inputs (from 1 to 4 previous months) for the prediction of each subcomponent of the month ahead. Predicted subcomponents were summed to create forecasted monthly flow. The procedure resulted in significant enlargement of the coefficient of correlation and decreased mean absolute error and mean squared error. Moreover, the procedure was applied for 1-6 months prediction and resulted in correlations close to one and low values of other two errors (Latifoglu et al., 2015). Pandiani and Shabri (2015) achieved significant improvements by using of a discrete wavelet transformation as pre-processing before the usage of ANN and SVM. While ANN combined with wavelet is very popular method in hydrology (Nourani et al., 2014), researches which use Hilbert-Huang transform for prediction of hydrological variables are rare. Hilbert-Huang transform is a powerful data decomposition method able to describe nonlinear and nonstationary signals (Barnhart, 2011). Its applications may be found for decomposition of climatic data, which could make it appropriate for long term planning and usage of exogeneous variables.

**Literature review**

A review of selected literature is shown on figures 2 and 3 (Latifoglu et al., 2015; Liu et al., 2015; Pandhiani and Shabri, 2015; Okkan and Fistikoglu, 2014; Farajzadeh et al., 2014; Terzi, O., 2014; Saghafian et al., 2013; Akiner and Akkoyunlu, 2012; Guo et al., 2011; Wu and Chau, 2010; Nilsson et al., 2006; Cigizoglu, H.K., 2005; Jain et al., 1999), while belonging abbreviations are in table 1.

**Discussion**

***Inclusion of the climate change impact***

In the terms of long term predictions and planning, inclusion of climatic change for the purpose to decision policy dependent on different climate scenarios is the next step. Some works including SL for that purpose will be mentioned, but also there are few researches of that nature, especially when it comes to more speficific subject of hydraulic structures used for water regime regulation. Anyway, the methodology is similar no matter which water resource system is used. Generally, global climate models (GCM), with spatial resolution of 100-300 km, are used for climate scenarios generation derived from various greenhouse gases emissions and response of the climate system (Baede, 2007 as cited in Winkler et al., 2011). As for experts in water resources field, of the greatest interest is to derive information about hydrological variables on local or eventually regional level, the process of downscaling from GCM is necessary. For water availability estimation based on diferent climate scenarios, the usage of the local information is needed. For that purpose, as precipitation and air temperature are the most reliable outputs from climate models, water availability on specific location should be connected with precipitation and air temperature from surrounding meteorological stations. Therefore, SL can be used as a type called empirical-dynamical downscaling, where downscaling is done directly from GCM or from regional climate models (derived from GCM) to meteorological stations (principles of downscaling can be found in Winkler et al., 2011). Most of researches use SL hydrological modeling of present situations. There is enough space for researches of the estimation of the impact of climate changes by using SL. I.e. Elgaali and Garcia (2007) used ANN for the estimation of climate change impact on accumalated precipitation for water supply (Winkler et al., 2011). Okkan and Fistikoglu (2014) used LM ANN for downscaling of the air temperature and precipitation, but used parametric conceptual hydrological model for monthly runoff estimation from those variables (Jain et al., 1999). Moreover, water availability problems for long term planning propose a need for using of optimization procedures to estimate a degree of satisfactory of users. Thus, simulations made by SL can be used for simulation-optimization procedures for the estimation of long term policy of water resources systems. Principal representation of potential methodlogy is shown on figure 4, together with the order of model building and usage. Usage of SL for water availability estimation in simulation-optimization procedures is not so common. I.e. example of usage of ANN for simulation, and also comparison with dynamic programming optimization method, can be found in (Jain et al., 1999).

**Table 1. Abbreviations related to the literature review**
**Tabuľka 1. Skratky týkajúce sa prehľadu literatúry**

| | |
|---|---|
| SL models | ANN…artificial neural network; SVM…support vector machine; NNM…nearest neighbours method; ANFIS…adaptive neuro-fuzzy inference system; GP…genetic programming; MLR…multiple linear regression; ARMA…autoregressive moving average; ARIMA…autoregressive integrated moving average (ARMA and ARIMA are stochastic models, but often compared with SL) |
| Search algorithms and methods for improvement | PSO…particle swarm optimization; LM…Levenberg-Marquardt algorithm; GD…gradient descent algorithm; GA…genetic algorithm; W…wavelet method; LS…least squares; SSA…singular spectrum analysis; PSR…phase space reconstruction; BMA…Bayesian model averageing; MA…moving average |
| Statistical error measures | *RMSE*…root mean squared error; *MAE*…mean absolute error; *NS*…Nash-Sutcliffe coefficient; *R*…correlation coefficient; $R^2$…coefficient of determination; *MRE*…mean relative error; *E*…coefficient of efficiency; *PI*…persistence index; *QR*…qualified rate; *MSE*…mean square error |
| Other explanations | /…uknown, missing, or remark; b…basin; s…station; t/t…training and testing; t/c/v…training, calibretion and testing; EN…endogeneous; EX…exogeneous; MBM…month-by-month; LTP…long term planning; inp…input variables; out…output variables; tot…totally; mv…missing values; p…prediction |

| Year | Authors | Purpose (predicted quantity) | Location | Data on disposition | Time framework (long term) MBM or LTP | Data manag. t/c/v or t/t | Remarks | Method | Structure (inp/out) | Inp-out relation EN or EX | Num. of instances | Methodology | Improvements (search algorithms, decomposition) | Statistical error measure Used | Conclusion |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2015 | Latifoglu et al. | Flow | river Kizilirmak, Turkey | 32 years of data | MBM | t/t: 68, 32 % | Principally MBM was applied, but forecastings were for 1-6 months ahead. | ANN | 1-4 inp, 1 out. | EN | tot 380 | The first step was to decompose time series into (12) independent components by using SSA. Then 1-4 previous flows were used for forecasting. Each subcomponent was forecasted particularly. Model with 3 inp was best and used for 1-6 months predictions. | SSA- (LM) ANN | *RMSE* *MAE* | SSA significantly improved the resullts, generally for ANN, but also for 1-6 months predictions. |
| 2015 | Liu et al. | Flow | r, Dongjiang, r. Xinfeng | Flows, rainfall, ENSO, IOD, 1989-2011 | MBM | t/t: 87, 13 % | The input configuration is not provided. | SVM ANFIS | 1 out | EN+EX | tot 276 | All predictors were decomposed by using of discrete wavelet transform in 90 different ways. After the decomposition, support vector regression was done. The best models were selected and ensemble forecast was done by using Bayesian model averaging. El Nino-Souther Oscillation (ENSO) and Indian Ocean dipole (IOD) were also used. ANFIS, best W-SVM, BMA-SVM, BMA-SVM without ENSO and IOD were used for prediction of flows 1-3 months ahead. | BMA-W-SVM | *RMSE* *NS* *R* | Approach with BMA provided the most accurate results. Inclusion of ENSO and IOD slightly increased the accuracy. 1 month ahead forecast was more accurate then 2-3 months ahed forecasts. |
| 2015 | Pandhiani and Shabri | Flow | rivers Neelum and Indus, Pakistan | Flows 1983-2012, 1983-2013 | MBM | t/t: 75, 25 % | / | ANN SVM | 1-6 inp, 1 out. | EN | tot 350 | Different configurations of ANN and least square SVM were used (1-6 previous monthly flows). Then the observed flows were decomposed into components using wavelet analysis. Three decomposition levels 2-4-8 months were used. Then those pre-processed data was used for prediction. | W-ANN W-LS-SVM | *RMSE* *MAE* *R* | Inclusion of wavelet decomposition improved the performance of both ANN and LS SVM. |
| 2014 | Okkan and Fistikoglu | Rainfall, temperature (and runoff) | b. Tahtali, Turkey | Output from climatic models, 1950-1999 | LTP | t/c/v: 50, 25, 25% | ANN was used for downscaling of rianfall and temperature. | ANN MLR | 2-12 inp, 2 out | *EX *(see remark)* | tot 600 | The third Generation Coupled Global Climate Model was used to model precipitation and air temperature in the future. 12 ANN models were built for temperature and for precipitation based ond varying predictors: air temperature, air pressure, sea level pressure, prate precipitatio etc. Monthly runoff was obtained by using conceptual GR2M on historical data. | LM (ANN) | $R^2$ *MSE* *RMSE* | Yearly averaged projections of runoff for 2010s, 2020s and 2030s were done. Predicited statistical decreases are, by order, 32, 39 and 38% (for A1B scenario) |
| 2014 | Farajzadeh et al. | Rainfall and runoff | Lake Urmia, Iran | Rainfall from 228 s. runoffs from 18 s., 1973-2011 | MBM | t/c/v: 70, 15, 15% | / | ARIMA ANN | / | EX EN | / | Firstly ANN and ARIMA were used for rainfall prediction. Using the runoff coefficient, runoff was predicted from predicted rainfalls. Those data was compared with runoff directly predicted by ANN and ARIMA. | LM (ANN) | *RMSE* *MAE* | Direct prediction of runoff gave less satisfying results, slightly more accurate with ARIMA. Runoff obtained from predicted rainfall gave good results. |
| 2014 | Terzi. | Inflow | r. Kizilrmak | Rainfall from 3 s., flow on 3 s., 1975-2005 | MBM | t/t 80, 20 % | Rainfall and inflows were used as predictors. | GP MLR | 1-5 inp, 1 out | EN + EX | tot 322 | Genetic programming (GP) is a generalization of genetic algorithms and it treats computer programmes like population and is used to find the best programme and its parameters. Flows from another 2 stations and rainfalls from 3 stations were used and results were compared with MLR. | GP programmes are expression trees | *RMSE* $R^2$ | GP model which used two flows from another two stations provided best results. GP was more accurate then MLR. |

*Fig. 2. Review of SL methods for the purpose of long term predictions and planning, part I.*
*Obr. 2 Prehľad metód SL na účely dlhodobých predpovedí a plánovania, časť I.*

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013 | Saghafian et al. | Flow | r. Karun, Iran | Precipitation from 28 s., temperature from 52 s., flow, SOI, 1974-2003 | MBM | t/t: 86, 14 % | Cross-validation was used in which 4 years subsets of whole data were left out every time. | ANN ANFIS NNM | Up to 15 inp, 1 out | EN+EX | tot 360 | Precipitation was spatially averaged by inverse distance weighted interpolation technique. Mean and maximum temperatures were used and maps were generated by using digital elevation model and elevation-temperature relationship. Input variables were varied: 1-3 previous months flows, mean and minimum temperatures, precipitations were used for 1-3 months ahead forecastings of flow. Later, 1-3 previous southern oscillation index (SOI) was added. Both point and spatial variables were used for prediction. | LM (ANN) | *RMSE* $R^2$ *MAE* | Spatially distributed data improved results of ANN and ANFIS more significantly then of NNM. SOI slightly improved results of flow prediction. ANN was more accurate then other two (probably because of usage of LM). |
| 2012 | Akiner and Akkoyunlu | Rainfall (and runoff) | b. Buyuk Melen, Kucuk Melen, Turkey | Rainfall, 2008-2009 (s. Duzce); rainfall, 1995-2009 (s. Bolu) | LTP | / | Calibration and verification are not egzactly described for ANN (they are for SWAT model). ANN was used for rainfall estimation. | ANN | 1-2-2-1 for m.v.; 5-6-4-1 for pr. | *EN (*see remark*) | / (seems tot. cca 150-200) | ANN is used for mv interpolation and rainfall pr., SWAT for runoff pr. Mv for Duzce station are obtained using data from Bolu. 5 inputs were rainfalls in last 5 years. 2010-2020 predictions for Duzce are obtained from Duzce rainfall. | LM (ANN) GD (ANN) | *NS* *r* $R^2$ | The purpose is obtaining the information about water availability in the future. |
| 2011 | Guo et al. | Inflow | b. Changjiang, China | Inflows from 1890-1990 | MBM | t/t: 80, 20% | / | SVM ANN | 8 inp, 1 out | EN | tot 1212 | Methodology for gaining more quality ML models is given, using noise, chaocity and PSR analysis. SVM improved by adaptable nonsensitive factor and PSO, ANN by LM, are compared to conventional SVM and ANN. | W-PSR-PSO (SVM) LM (ANN) | *MAE* *MRE*, $R^2$ | Proposed methodology obtained more accurate results then conventional ANN and SVM. On the other side, more complex models slow the process. |
| 2010 | Wu and Chau | Inflow | S. Cuntan, b. Xiangjiaba, Manwan, Danjiangko, China | 1893-2007; 1940-1997; 1974-2003; 1930-1981 | MBM | t/t: 66.5, 33.5% | / | ARMA ANN NNM | 6-12 inp, 1 out | EN | / | After preparing inputs with correlation exponent method and false nearest neighbours (PSR methods) ANN was used. Later MA was combined with ANN, LM-GA was used for optimizing ANN. Those are compared to NNM and ARMA. | MA (ANN) PSR (ANN) LM-GA (ANN) | *RMSE* *E, PI* *QR* | Local techniques (NNM) provided more accurate results then global averaging (ANN). ANN was improved by MA and LM-GA, but less accurate then ARMA. |
| 2006 | Nilsson et al. | Runoff | Sub-basins Bulken and Skarsvatn | Runoff: Bulken, 1957-2002, Skarsvatn, 1973-2002 | LTP | t/t: 80, 20% | For b. Bulken rainfall and temperature from 3 s. were used. For s. Skarsvatn from 4 s. | ANN | 6-12 inp, 1 out | EX | tot 548, 348 | Different possibilites for data usage were analyzed. First rainfall and temperature were used, then seasonal characteristics, snow quantity, soil moisture, north Atlantic oscillation index (NAO). ANN results are compared with results from conceptual model and its combination with ANN. | / | $R^2$ *AD* | Including snow and seasonal characteristics improved results. ANN generally was better then conceptual model and could detect sudden changes, except for spring flows. Combination performed best. |
| 2005 | Cigizoglu et al. | Inflow | s. Krahacili, Goksu river | Inflows: 1963-1989 | MBM | t/t: 63, 37 % | / | ANN MLR ARMA | 6-7 inp, 1 out | EN | tot 324, 32525 | Classically used feedforward backpropagation ANN was compared with generalized regression ANN. Synhtetic flows series was built using the ARMA model. Mean monthly flows from synthetic series was added as input in FFBP ANN and GR ANN. The same thing was done with MLR method. Then ANN and MLR were trained with 32400 generated monthly flow values from synthetic generation. All those models were compared with ARMA model. | GR (ANN) ANN+AR | $R^2$ *MSE* | ANN models were improved by incorporating the synthethicaly generated flow series. GR ANN gave best results, but incorporated with generated series overestimated peak flows. FFBP ANN is less stable and builds models quickly then GR ANN. |
| 1999 | Jain et al. | Inflow | Upper Indravati reservoir, India | Inflows: 1951 - 1982 | MBM | t/t: 87.5, 12.5 % | / | ANN ARMA | 3-5 inp, 1 out | EN | tot 384 | The number of input variables was varied and final configuration ended with two previous months and the same month from previous year as inputs. Predicted value was inflow to the reservoir. Additionally, ANN approach was used for reservoir policy estimation and compared to dynamic programming. | / | *RMSE* *MSE* | ANN generally provided more precise results. High flows were better estimated with ANN, while low flows were better estimated with ARIMA. |

*Fig. 3. Review of SL methods for the purpose of long term predictions and planning, part II.*
*Obr. 3. Prehľad metód SL na účely dlhodobých predpovedí a plánovania, časť II.*

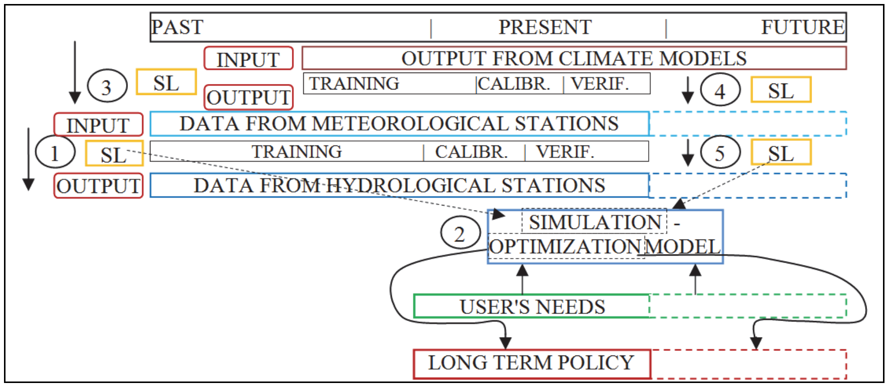


*Fig. 4. Principal representation of the methodology for the estimation of climate change impact on long term planning*
*Obr. 4. Hlavné znázornenie metodiky pre odhad vplyvu zmeny klímy na dlhodobé plánovanie.*

According to figure 4, the first step is building of the SL model for obtaining the relationship between hydrological variable at the station of interest. Therefore, water availability is described with data from meteorological stations. This simulation model can be used for optimization of the user needs and long term policy of water resources system in the present situations (step 2). The third step is obtaining of the relationship between output from climate models (global or regional) and meteorological stations. Then, future variables on meteorological stations can be predicted, as also water availability from predictions on meteorological stations (model built in the first step is used). Finally, on the basis of simulations from the fifth step and projected user's needs, future long term policy of water resources system can be estimated. Moreover, cost-benefit analysis can be done, and appropriate policy, depending on the scenario in the future, can be applied. Obtaining of the first step, building of the SL model able to provide water availability on certain location from surounding meteorological stations is the most important step.

## Conclusion

In the paper basic principles of usage of supervised learning techniques of machine learning, for long term prediction and planning are proposed. Long term planning for the purpose of adequate water resources management by using of the supervised learning predictions is the objective of the research. Basic assumptions for application were introduced as also the literature review was proposed. From the literature review, space for future researches was shown. Methods of machine learning learn from data, so the major question for application is the number of instances in data for long term predictions and planning. So, the influence of historical series lengths on supervised learning methods ability to model hydrological variables should be investigated. Proposed models should be stable in dependence of length if they are going to be used for long term planning. Also, for the purpose of long term planning, models must be built by using of exogeneous input variables, mainly climatological variables. Building of models with mentioned assumptions, able to model relations between climatological variables as input, and hydrological variable as output, is the main step and basis of the water resources systems management with supervised learning.

## Literature

Akiner, M.E., Akkoyunlu, A. (2012): Modeling and forecasting river flow rate from the Melen Watershed, Turkey, Journal of Hydrology 456–457, 121–129.

Barnhart, B.L. (2011): The Hilbert-Huang transform: theory, applications, development, dissertation, University of Iowa.

Cigizoglu, H.K. (2005): Generalized regression neural network in monthly flow forecasting, Civil Engineering and Environmental Systems 22(2), 71–84.

Elgaali, E., Garcia, L.A. (2007): Using Neural Networks to Model the Impacts of Climate Change on Water Supplies, Journal of Water Resources Planning and Management 133(3), 230–243.

Farajzadeh, J., Fard, A.F., Lotfi, S. (2014): 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, 38–48.

Guo, J., Zhou, J., Qin, H., Zou, Q., Li, Q. (2011): Monthly streamflow forecasting based on improved support vector machine model, Expert Systems with Applications 38, 13073–13081.

Jain, S.K., Das, A., Srivastava, D.K. (1999): Application of ANN for reservoir inflow prediction and operation,

Journal of Water Resources Planning and Management 125(5), 263–271.
Latifoglu, L., Kisi, O., Latifoglu, F. (2015): Importance of hybrid models for forecasting of hydrological variable, Neural Computations and Applications 26, 1669–1680.
Liu, Z., Zhou, P., Zhang, Y. (2015): A Probabilistic Wavelet-Support Vector Regression Model for Streamflow Forecasting with Rainfall and Climate Information Input, Journal of Hydrometeorology 16, 2209–2229.
Nilsson, P., Uvo, C.B., Berndtsson, R. (2006): Monthly runoff simulation: Comparing and combining conceptual and neural network models, Journal of Hydrology 321, 344–363.
Nourani, V., Baghanam, A.H., Adamowski, J., Kisi, O. (2014): Applications of hybrid wavelet – Artificial Intelligence models in hydrology: A review, Journal of Hydrology 514, 358–377.
Okkan, U., Fistikoglu, O. (2014): Evaluating climate change effects on runoff by statistical downcaling and hydrological model GR2M, Theorethical and Applied Climatology 117, 343–361.
Pandhiani, S.M., Shabri, A.B. (2015): Time Series Forecasting by Using Hybrid Models for Monthly Streamflow Data, Applied Mathematical Sciences 9(57), 2809–2829.
Šmuc, T. (2013): Strojno učenje, lectures, Faculty of Science, Zagreb.
Saghafian, B., Anvari, S., Morid, S. (2013): Effect of Southern Oscillation Index and spatially distributed climate data on improving the accuracy of Artificial Neural Network, Adaptive Neuro-Fuzzy Inference System and K-Nearest Neighbour streamflow forecasting models, Expert Systems 30(4), 367–380.
Terzi, O. (2014): A genetic programming approach to river flow modeling, Journal of Intelligent & Fuzzy Systems 27, 2211–2219.
Winkler, J.A., Guentchev, G.S., Perdinan, Tan, P-N., Zhong, S., Liszewska, M., Abraham, Z., Niedzwiedz, T., Ustrnul, Z. (2011): Climate Scenario Development and Applications for Local/Regional Climate Change Impact Assessments: An Overview for the Non-Climate Scientist, Part I, Geography Compass 5/6, 275–300.
Wu, C.L., Chau, K.W. (2010): Data-driven models for monthly streamflow time series prediction, Engineering Applications of Artificial Intelligence 23, 1350–1367.

## DLHODOBÁ PREDPOVEĎ PRÍTOKU POUŽITÍM SUPERVÍZOVANÉHO UČENIA

V článku sú navrhnuté základné princípy používania supervízovaných učebných techník strojového učenia, dlhodobej predpovede a plánovania. Cieľom výskumu je dlhodobé plánovanie na účely primeraného manažmentu vodných zdrojov využitím predpovedí supervízovaného učenia. V práci boli uvedené základné predpoklady pre aplikáciu, ako aj návrh literatúry. Z prehľadu literatúry sa ukázal priestor pre budúce výskumy. Metódy strojového učenia sa učia z dát, takže hlavnou otázkou pre aplikáciu je počet prípadov v údajoch pre dlhodobé predpovede a plánovanie. Preto by mal byť skúmaný vplyv historických sérií dĺžok na metódy supervízovaných metód učenia na modelovanie hydrologických premenných. Navrhované modely by mali byť stabilné v závislosti od rozsahu, ak sa budú používať na dlhodobé plánovanie. Na účely dlhodobého plánovania sa modely musia stavať aj pomocou exogénnych vstupných premenných, najmä klimatologických premenných. Vybudovanie modelov s uvedenými predpokladmi, ktoré sú schopné modelovať vzťahy medzi klimatologickými premennými ako vstupom a hydrologickou premennou ako výstupom, je hlavným krokom a základom riadenia systémov vodných zdrojov so supervízovaným vzdelávaním.

Jadran Berbić
Croatian Hydrological and Meteorological Service
Zagreb
Croatia
E-mail: jberbic@hotmail.com

Eva Ocvirk
Faculty of Civil Engineering
University of Zagreb
Zsgreb
Croatia
E-mail: ocvirk@grad.hr