Global NEST Journal, Vol 19, No X, pp XX-XX.

A multi-stage methodology for selecting input variables in ANN forecasting of river flows

Panagoulia D.[1,*], Tsekouras G.J.[2], Kousiouris G.[3]

[1]School of Civil Engineering, Department of Water Resources & Environmental Engineering, National Technical University of Athens, 5 Heroon Polytechniou Str., 15780 Zografou, Athens, Greece,

[2]Electric Circuits Laboratory, Department of Electrical Engineering & Computer Science, Hellenic Naval Academy, Terma Hatzikyriakou, Piraeus, Greece

[3]Department of Electrical and Computer Engineering, National Technical University of Athens, 9, Heroon Polytechniou Str, 15773 Athens, Greece

Received: 05/07/2016, Accepted: 05/02/2017, Available online: 15/02/2017
*to whom all correspondence should be addressed:
e-mail: dpanag@hydro.ntua.gr

Abstract

The scientific community has recognized the necessity for more efficiently selected inputs in artificial neural network models (ANNs) in river flows and has worked on this despite some shortcomings. Moreover, there is none or limited inclusion of ANN inputs coupled with atmospheric circulation under various patterns arising from the need of data downscaling for climate change predictions in hydrology domain. This paper presents the results of a novel multi-stage methodology for selecting input variables used in artificial neural network (ANN) models for river flow forecasting. The proposed methodology makes use of data correlations together with a set of crucial statistical indices for optimizing model performance, both in terms of ANN structure (e.g. neurons, momentum rate, learning rate, activation functions, etc), but also in terms of inputs selection. The latter include various previous time steps of daily areal precipitation and temperature data coupled with atmospheric circulation in the form of circulation patterns, observed river flow data and time expressed via functions of sine and cosine. Additionally, the no-linear behavior between river flow and the respective inputs is investigated by the ANN configuration itself and not only by correlation indices (or other equivalent contingency tools). The proposed methodology revealed the river flow of past four days, the precipitation of past three days and the seasonality as robust input variables. However, the temperature of three past days should be considered as an alternative against the seasonality. The produced models forecasting ability was validated by comparing its one-step ahead flow prediction ability to two other approaches (an auto regressive model and a genetic algorithm (GA)-optimized single input ANN).

Keywords: Artificial neural networks, optimization, input variables selection, flow forecasting, atmospheric circulation, seasonality, hydrological evaluation indices

1. Introduction

Hydro-meteorological and water resources systems are extremely complex, nonlinear and dynamic in nature, involving a variety of physical variables. Such systems can be modeled via ANNs or various hybrid schemes of them (Maier and Dandy, 2001; May et al., 2008; Maier et al., 2010). One of the most important steps of ANN model development process is the determination of an appropriate set of input variables (Maier and Dandy, 2001; May et al., 2008) along with the determination of the parameters of the network.

Considering the development of a forecasting model including lags, the minimum number of variables should be used as inputs to the ANN model in order to increase computational efficiency, minimize redundancy, reduce noise and increase the interpretability of the model (May et al., 2008). The significance of input variables selection has been assessed by a number of specific techniques embodied into two categories, namely model-based (Maier et al., 2010) and model-free approaches (May et al., 2008; Fernando et al., 2009). The computational efficiency of the input selection approach is of great significance especially when large data sets have to be modeled as it usually occurs in hydrological modeling. Indicatively, criterions for input selections can be the correlation index, the mutual information (Fernando et al., 2009), the partial mutual

information algorithm of Sharma (Sharma, 2000; Bowden et al., 2005), the average shifted histograms (Fernando et al., 2009], the Hampel distance outlier stopping criterion for large sample sizes (Davies and Gather, 1993), the ensemble empirical mode decomposition (Wang et al, 2015) etc. Most modelers use cross-, auto-, and partial auto-correlations between the input-output variables assuming linear relationship (even though it is not correct) with significant lags (Nayak et al., 2013). Not often the input selection is connected to the forecasting model with a retrospective way based on binary-coded particle swarm optimization (Taormina et al, 2015), with exotic data-preprocessing techniques, such as spectrum analysis (Wu et al, 2009; Chau et al, 2010) or with different kind of algorithms, such as differential evolution, artificial bee colony, ant colony optimization (Chen et al, 2015). From a climate standpoint, numerous studies have established links between large-scale atmospheric circulation in the mode of North Atlantic Oscillation (Lintner and Chiang, 2007), El Niño Southern Oscillation (Philander, 1990) and circulation patterns (Panagoulia et al., 2006a; Panagoulia et al., 2006b; Panagoulia et al., 2008) with hydro-meteorological variables in various time scales. In almost all cases the input variables are the precipitation, the temperature, the groundwater level, the load flow of previous time steps, while rarely other variables are found, such as tree-ring diameters by dendrochronology (Gholami et al, 2015). In this paper, the precipitation and temperature coupled with atmospheric circulation are included for inputs selection in ANN models river flow forecasting with the aim to explore the selection process of such variables.

The objective of this study is to obtain via a systematic manner an efficient selection of input variables for flow forecasting of the next day. As input data the time series of precipitation and temperature formulated through circulation patterns, the observed river flow and the time expressed via functions of sine and cosine (seasonality) are considered. Following, a stepwise multi-stage methodology is developed, tested and applied for selecting which hydro-meteorological input variables will be used. In first stage it makes use of cross-, auto-, and partial auto-correlations, which is similar to previous input selection methodologies and can be replaced by Fourier spectrum analysis etc. In second stage the no-linear ANN behavior is explored by the ANN construction with one input (precipitation, temperature, river flow of past time points) and one output (current river flow). In the third stage the investigation of the respective behavior is extended by the ANN construction with more inputs of the same kind and at the last stage all possible inputs are combined. The inputs selection for each stage (2^{nd} to 4^{th}) is based on a set of crucial statistical indices which is able to exploit nonlinear input variables and through an appropriate process can also optimize the parameters of the same network. The approach is compared against two alternatives, an auto-regressive model with only linear characteristics but very fast creation time, and a GA-based optimized ANN architecture (Kousiouris et al, 2012), investigating a number of different parameters and trade-offs in network design (training functions, different types of neuron activation functions, size and number of layers etc.), but depending only on the previous values of the forecasted metric, in order to showcase the trade-off of not selecting multiple hydrological criteria but dealing only with the data as a time series. Numerous alternatives exist (e.g. SVMs); however the major benefit of the latter (optimal training in the given dataset) exists only for the usage of linear kernels in their structure. In our case the existence of linear features is represented in the AR model and in the pure linear combinations of activation functions in the GA-ANN compared method.

2. Methodology

A stepwise multi-stage methodology selection of ANNs input variables is proposed in Fig.1, based in Fortran, which includes the next steps:

A. 1st stage – Input variables pre-selection

In this stage the goal is to find out the time range of the input –output variables, which influence the output variable, that is the river flow-discharge *Q(t)*. The time k_Q which involves discharges from previous days is determined by the auto-correlation and partial auto-correlation coefficients, while through the cross-correlation coefficients the time intervals k_P and k_T for the precipitation *P(t)* and temperature *T(t)* are calculated. However, this statistical analysis does not identify which variables are not needed to be taken into account due to data overlapping.

B. 2nd stage – Input variables selection based on ANNs with one input

The first step of selecting input variables based on ANNs that have only one input is introduced. Particularly, for each input kind (discharge, temperature and precipitation) an ANN with one input and one output variable is constructed. I.e. since the involved discharges from the first stage resulted in k_Q, then the following k_Q ANNs are constructed with the form of {(input)→(output)}:{($Q(t-1)$) →($Q(t)$)}, {($Q(t-2)$)→($Q(t)$)}, ..., {($Q(t-k_Q)$)→($Q(t)$)}. 11 statistical indices are used to decide which one input - one output ANN per variable kind expresses at the best possible way the non-linear relationship between the discharge $Q(t)$ and each input kind (more details in Section 2.E). This way the time period k_Q/ which involves the discharges of past days is defined, namely the $Q(t-1)$, ..., $Q(t-k_Q/)$, the time period k_P/ which involves the precipitation of current day $P(t)$ and past days $P(t-1)$, ..., $P(t-k_P/)$, as well as the time period k_T/ involving the temperature of current day $T(t)$ and past days $T(t-1)$, ..., $T(t- k_T/)$. From this step, the number of past variables of the same kind which are involved to $Q(t)$ is determined.

C. 3rd stage – Input variables selection based on ANNs with one kind of variables

The second selection step of ANN-based input variables with more than one input of the same kind is performed. More specifically, taking into account the k_Q/ involved discharges resulted from the second stage, the physical structure of the problem and the user's desire to involve

unified sets of same kind of variables in the forming model, the following ANNs of shape {(inputs) → (outputs)} are constructed: {($Q(t\text{-}1)$, $Q(t\text{-}2)$)→($Q(t)$)}, {($Q(t\text{-}1)$, $Q(t\text{-}2)$, $Q(t\text{-}3)$)→($Q(t)$)}, ..., {($Q(t\text{-}1)$, $Q(t\text{-}2)$, $Q(t\text{-}3)$, …, $Q(t\text{-}k_Q)$)→($Q(t)$)}, {($Q(t\text{-}2)$, $Q(t\text{-}3)$)→($Q(t)$)}, ..., {($Q(t\text{-}2)$, $Q(t\text{-}3)$, …, $Q(t\text{-}k_Q)$)→($Q(t)$)}, with a total number of new ANNs equal to (k_Q/·(k_Q/-1))/2.

Similarly, ANNs are constructed for the precipitation and for the temperature. Subsequently, the ANN whose trained parameters lead to the relative majority of better values of statistical evaluation indices is selected against other settings selected (more details in Section 2.E). The ANNs with multiple inputs of the same kind - one output that express at the best possible way the non-linear relationship between the discharge $Q(t)$ and each input kind are selected defining the specific time period [k_{Q1}, k_{Q2}] for the discharges, [k_{P1}, k_{P2}] for precipitation and [k_{T1}, k_{T2}] for temperature. Additionally, the effect of seasonality is separately examined as a 4th kind of input variables through the pair of periodic functions $\cos(2\cdot\pi\cdot d/days_{year})$ and $\sin(2\cdot\pi\cdot d/days_{year})$ (Hippert et al, 2001), where d is the serial number of current day in the current year (from the start of the year) and $days_{year}$ is the number of days of the current year.

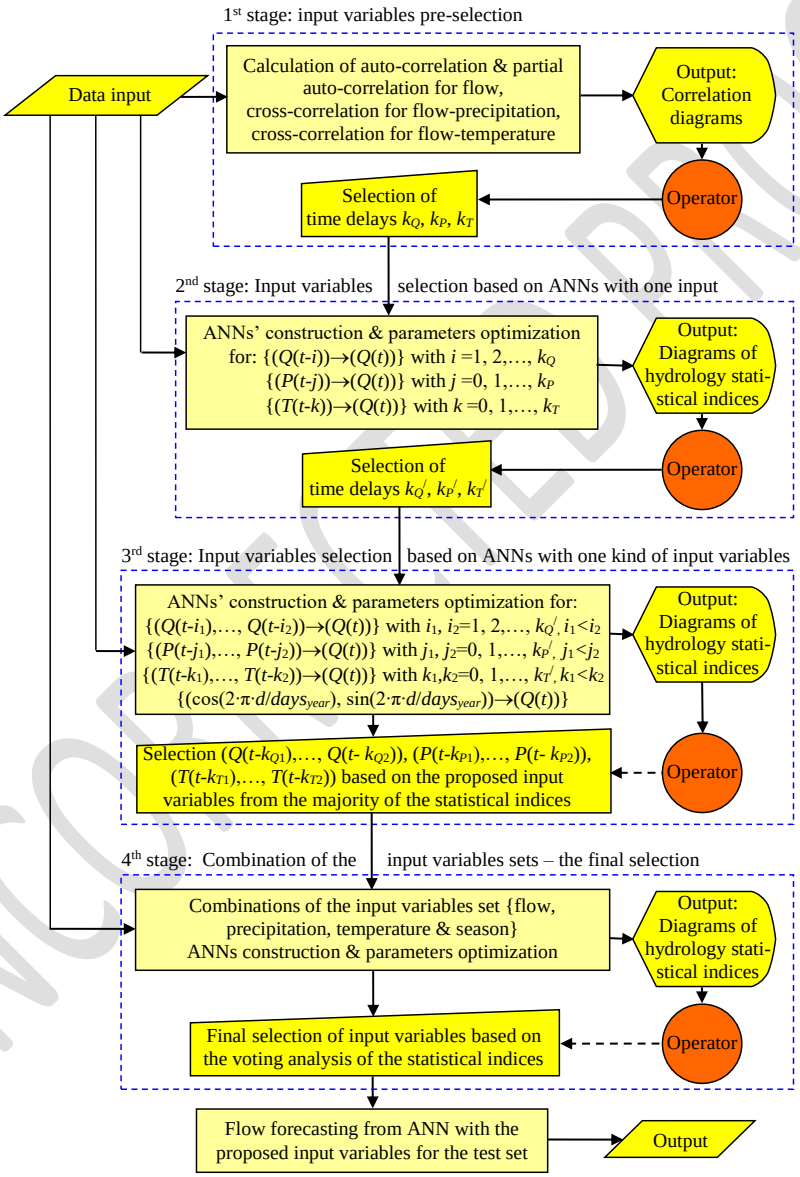

Figure 1. Flow chart of the proposed input variables methodology

D. 4th stage – Combination of input variables sets

The fourth step of input variables selection by using ANNs is performed, by the combination of the sets of same kind variables determined by the previous stage, which are the sets of discharge {$Q(t\text{-}k_{Q1})$, $Q(t\text{-}k_{Q1}\text{-}1)$, ..., $Q(t\text{-}k_{Q2})$}, precipitation {$P(t\text{-}k_{P1})$, $P(t\text{-}k_{P1}\text{-}1)$, ..., $P(t\text{-}k_{P2})$}, temperature {$T(t\text{-}k_{T1})$, $T(t\text{-}k_{T1}\text{-}1)$, ..., $T(t\text{-}k_{T2})$} and seasonality {$\cos(2\cdot\pi\cdot d/days_{year})$, $\sin(2\cdot\pi\cdot d/days_{year})$}.

E. *Optimized ANN Construction and rating*

In stages 2, 3, 4 of the proposed methodology a number of ANNs are constructed, trained and rated on their effectiveness. The basic steps of a typical ANN optimization method are presented in Fig. 2. More specifically, the ANN's training algorithm is the stochastic training back-propagation process with decreasing functions of learning rate and momentum term, for which an optimization process is conducted regarding the crucial parameters values gradually, such as the number of neurons, the kind of activation functions, the initial values and time parameters of learning rate and momentum term, the kind of activation functions and their parameters, the training process uses the training data set, while the optimization ANN parameters process the evaluation set. The specific method was selected since it is the most commonly available in various ANN implementation toolboxes. The performance of each ANN structure is evaluated using the evaluation set based on eleven criteria (Jain and Prased, 2003), which are the root mean square error (RMSE), the correlation index (R), the mean absolute percentage error (MAPE), the mean percentage error (MPE), the mean percentage error (ME), the percentage volume in errors (VE), the percentage error in peak (MF), the normalized mean bias error (NMBE), the normalized root mean bias error (NRMSE), the Nash-Sutcliffe model efficiency coefficient (E) and the modified Nash-Sutcliffe model efficiency coefficient (E1). In an effort to have a generic optimization approach, including various hydrological criteria, the use of all these indices was used through voting analysis. Each criterion may have one specific interest for an adopter of the model, as indicated in the analysis in Section 4. Having all criterions helps to select a model based on a given approach that may be preferred over others (e.g. RMSE for high flows etc.), having all 11 searches for a more generic solution. Finally, the generalization of ANNs is checked by the discharge forecasting of the "unknown" test set.

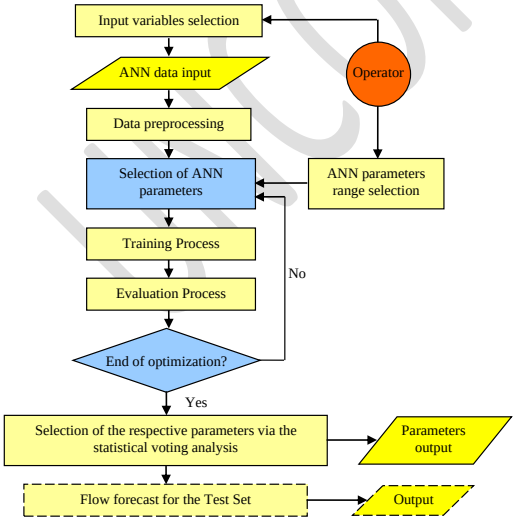

Figure 2. Flow chart of the ANN optimization method

3. Case Study

A. *Study catchment and observed data*

The Mesochora catchment drained by Acheloos' river in central-western Greece was selected for this study due to the partial diversion of the river flow in order to irrigate the arid Thessaly plain and boost hydropower generation in the surrounding region. The catchment has an area of 633 km² and extends nearly 32 km from north (39º42') to south (39º 25') with an average width of about 20 km. Daily precipitation was available at 12 stations for the period of 1972-1992, while mean daily temperature was collected from 4 stations for the period of 1972-1992. The precipitation and temperature variability at the stations was determined by conditioning on circulation patterns (CP) types (Panagoulia et al., 2006a; Panagoulia et al., 2006b; Panagoulia et al., 2008). For ANNs' training process three sets are formed: training set: 80% vectors of time period 1972-86, evaluation set: 20% vectors of time period 1972-86 and test set: 100% vectors of time period 1987-92.

B. *1st stage – Input variables pre-selection*

The auto-correlation and partial auto-correlation coefficients of flow, the correlation coefficients between flow and precipitation, and flow and temperature are defined for the use set data of 1972-86 period and the check set data of 1987-92 period (see Figs. 3 to 6). In this context, the typically selected number of parameters in a first estimation could be: $k_Q = 9$, $k_P = 5$, $k_T = 5$~20.

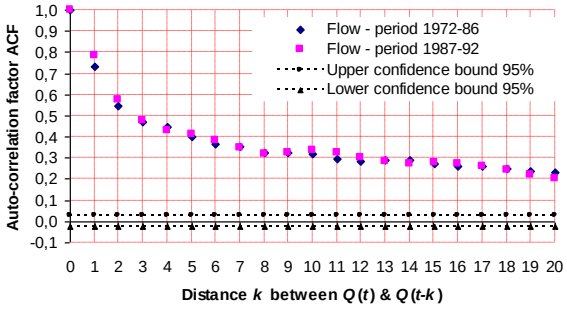

Figure 3. Auto-correlation of flow for the periods 1972-1986 and 1987-1992

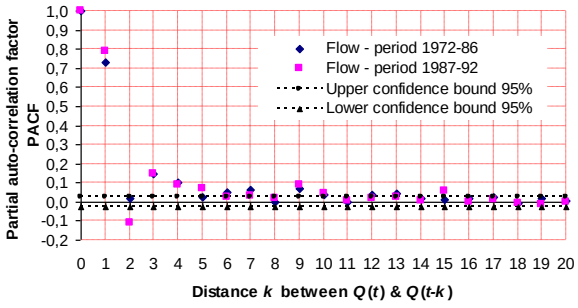

Figure 4. Partial auto-correlation of flow for the periods 1972-1986 and 1987-1992

C. 2nd stage – Input variables selection based on ANNs with one input

For reasons of certainty and comparison, we selected for investigation the flows of 20 past days and the precipitation and temperature both of current day and 20 past days. In this case, 62 ANNs are constructed, i.e. ANN model with serial number 1 (No.1) has formed input vector by the flow of the past day $Q(t\text{-}1)$ and the output vector by the current flow $Q(t)$.

Several parameters have to be selected:

- the number of neurons of the hidden layer { 1 to 20 with incremental step 1},
- the initial value {0.05, 0.01, 0.15, …, 1.00} and time parameter {100, 200, …, 1600} of momentum term and training rate,
- the type of activation functions of the layers (linear, hyperbolic tangent, hyperbolic sigmoid) with multiplicative factor {0..05, 0.10, …, 1.0} for the hidden layer and {0..05, 0.10, …, 0.7} for the output layer.

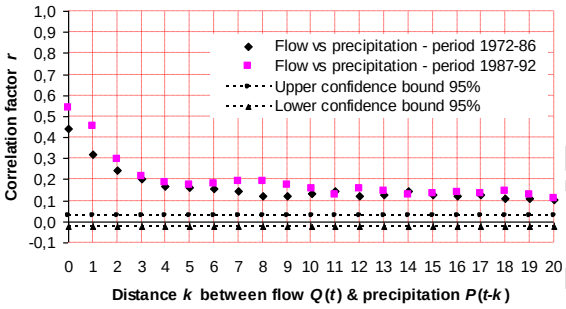

Figure 5. Correlation of flow and precipitation for the periods 1972-86 and 1987-92

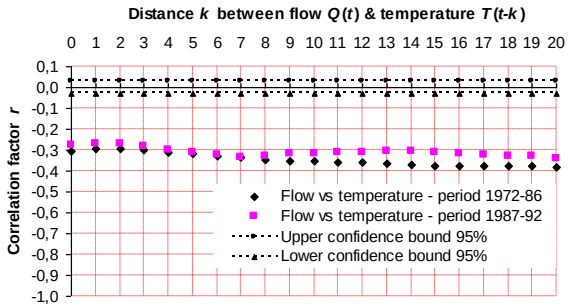

Figure 6. Correlation of flow and temperature for the periods 1972-86 and 1987-92

For each ANN 949 scenarios of different ANN parameter calibrations have been examined based on optimization process of Fig. 2 instead of 5,160,960,000 cases. The graphs between the current flow and past days flow essentially substitute the graphs of auto-correlation coefficients of the 1st stage. From Figure 7 a systematic improvement of correlation values is noted as well as a stabilization of it after 5-6 past days. Similar conclusions are drawn from the study of the statistical indices *RMSE, MAPE, MPE, MF, NRMBE, E,* and E_1. In total, k_Q'=6, k_P' = 5, k_T'=3 have been selected.

D. 3rd stage – Input variables selection based on ANNs with one kind of variables

62 ANNs are constructed which are presented in Table 1.

Each ANN is stepwise optimized as described previously, with the same parameters as the previous stage, except:

- the number of neurons of the hidden layer { 1 to 25 with incremental step 1},
- the initial value {0.05, 0.01, 0.15, …, 1.00} and time parameter {100, 200, …, 2000} of momentum term and training rate

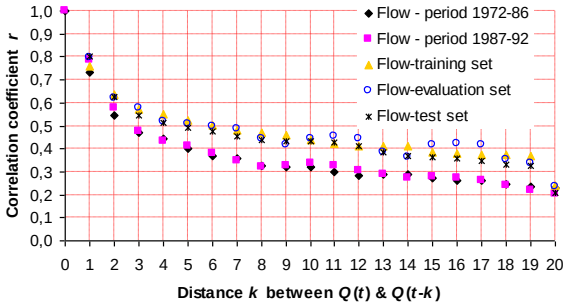

Figure 7. Correlation coefficient r between $Q(t)$ & $Q(t\text{-}k)$ for recorded data sets 1972-1986 and 1987-1992, for ANNs with training set, for ANNs with evaluation set, and for ANNs with test set

Table 1. ANNs Construction for 3rd stage of proposed methodology - Input Variables Selection based on ANNs with Multi-Inputs of the same variable kind

No.	Input variables		No.	Input variables		No.	Input variables	
ANN	From	To	ANN	From	To	ANN	From	To
63	$Q(t\text{-}1)$	$Q(t\text{-}2)$	78	$P(t)$	$P(t\text{-}1)$	93	$T(t)$	$T(t\text{-}1)$
64	$Q(t\text{-}1)$	$Q(t\text{-}3)$	79	$P(t)$	$P(t\text{-}2)$	94	$T(t)$	$T(t\text{-}2)$
65	$Q(t\text{-}1)$	$Q(t\text{-}4)$	80	$P(t)$	$P(t\text{-}3)$	95	$T(t)$	$T(t\text{-}3)$
66	$Q(t\text{-}1)$	$Q(t\text{-}5)$	81	$P(t)$	$P(t\text{-}4)$	96	$T(t\text{-}1)$	$T(t\text{-}2)$
67	$Q(t\text{-}1)$	$Q(t\text{-}6)$	82	$P(t)$	$P(t\text{-}5)$	97	$T(t\text{-}1)$	$T(t\text{-}3)$
68	$Q(t\text{-}2)$	$Q(t\text{-}3)$	83	$P(t\text{-}1)$	$P(t\text{-}2)$	98	$T(t\text{-}2)$	$T(t\text{-}3)$
69	$Q(t\text{-}2)$	$Q(t\text{-}4)$	84	$P(t\text{-}1)$	$P(t\text{-}3)$	99	$\cos(2\cdot\pi\cdot d/days_{year})$,	

70	$Q(t\text{-}2)$	$Q(t\text{-}5)$	85	$P(t\text{-}1)$	$P(t\text{-}4)$	$\sin(2\cdot\pi\cdot d/days_{year})$
71	$Q(t\text{-}2)$	$Q(t\text{-}6)$	86	$P(t\text{-}1)$	$P(t\text{-}5)$	
72	$Q(t\text{-}3)$	$Q(t\text{-}4)$	87	$P(t\text{-}2)$	$P(t\text{-}3)$	
73	$Q(t\text{-}3)$	$Q(t\text{-}5)$	88	$P(t\text{-}2)$	$P(t\text{-}4)$	
74	$Q(t\text{-}3)$	$Q(t\text{-}6)$	89	$P(t\text{-}2)$	$P(t\text{-}5)$	
75	$Q(t\text{-}4)$	$Q(t\text{-}5)$	90	$P(t\text{-}3)$	$P(t\text{-}4)$	
76	$Q(t\text{-}4)$	$Q(t\text{-}6)$	91	$P(t\text{-}3)$	$P(t\text{-}5)$	
77	$Q(t\text{-}5)$	$Q(t\text{-}6)$	92	$P(t\text{-}4)$	$P(t\text{-}5)$	

For each statistical evaluation index the ANN with different inputs that presents the better behavior is identified. The corresponding data are summarized into Table 2. Additionally, Table 3 portrays how many times an input has been proposed for use by the various statistical indices if absolute majority (at least 6 from 11 criteria) has been achieved. Combining the results of the training and evaluation sets, the use of daily flow values *Q(t-1)* to *Q(t-4)*, daily precipitation values conditioned on CPs data *P(t)* to *P(t-3)*, and temperature data ,conditioned on CPs, *T(t)* to *T(t-3)* is suggested.

Table 2. Study of ANNs inputs variables scenarios per input kind with the best behavior per statistical index and set

Statistical index	Daily flow						Daily precipitation						Daily mean temperature					
	Train-ing set		Evalua-tion set		Test set		Train-ing set		Evalua-tion set		Test set		Train-ing set		Evalua-tion set		Test set	
	From	To	From	To	From	To	From	To	From	To	From	To	From	To	From	To	From	To
RMSE	t-1	t-4	t-1	t-6	t-1	t-4	t	t-5	t	t-1	t	t-4	t	t	t	t-1	t	t-3
r	t-1	t-4	t-1	t-6	t-1	t-4	t	t-5	t	t-1	t	t-4	t	t-3	t	t-1	t	t-3
MAPE	t-1	t-3	t-1	t-3	t-1	t-3	t	t-1	t	t-5	t	t-1	t-2	t-3	t-2	t-3	t-3	t-3
MPE	t-1	t-3	t-1	t-3	t-1	t-3	t	t-1	t	t-1	t	t-1	t-2	t-3	t-2	t-3	t-3	t-3
RME	t-4	t-4	t-2	t-5	t-2	t-3	t-2	t-2	t-1	t-4	t-3	t-3	t	t	t	t-3	t	t
VE	t-4	t-4	t-4	t-6	t-2	t-3	t-2	t-2	t-1	t-4	t-3	t-3	t	t	t	t-3	t	t-3
MF	t-1	t-2	t-1	t-2	t-1	t-5	t	t-1	t	t-1	t	t-1	t	t-1	t	t-2	t-2	t-3
NMBE	t-4	t-4	t-2	t-2	t-2	t-3	t-2	t-2	t	t	t-3	t-3	t	t	t	t-3	t-3	t-3
NRMBE	t-1	t-4	t-1	t-2	t-1	t-4	t	t-5	t	t-5	t	t-5	t	t-3	t	t-2	t	t-2
E	t-1	t-4	t-1	t-6	t-1	t-4	t	t-5	t	t-3	t	t-4	t	t-3	t-1	t-1	t	t-3
E1	t-1	t-3	t-1	t-3	t-1	t-3	t	t-5	t	t-5	t	t-5	t-2	t-3	t-2	t-3	t	t-3

Table 3. Determination of ANNs inputs sets per kind of input variable based on activation of statistical indexes for training set, evaluation set and test set

Daily flow	Inputs	*Q*(*t*-1)	*Q*(*t*-2)	*Q*(*t*-3)	*Q*(*t*-4)	*Q*(*t*-5)	*Q*(*t*-6)
Set	Training	8	8	7	7	0	0
	Evaluation	8	10	7	5	5	4
	Test	8	11	11	5	1	0
Daily precipitation	**Inputs**	**P(t)**	***P*(*t*-1)**	***P*(*t*-2)**	***P*(*t*-3)**	***P*(*t*-4)**	***P*(*t*-5)**
Set	Training	8	8	8	5	5	5
	Evaluation	9	10	6	6	5	3
	Test	8	8	5	8	5	2
Daily mean temperature	**Inputs**	**T(t)**	***T*(*t*-1)**	***T*(*t*-2)**	***T*(*t*-3)**		
Set	Training	10	7	7	6		
	Evaluation	7	8	8	6		
	Test	7	6	7	8		

Additionally, the seasonality influence is separately investigated via the two simple functions of $\cos(2\cdot\pi\cdot d/days_{year})$ and $\sin(2\cdot\pi\cdot d/days_{year})$ (Hippert et al.,2001), which constitute the 99th ANN in Table 1. These functions appear to be important as input variables since they reflect a correlation coefficient much greater than any other associated to temperature input that reaches the value of 0.448. The annual periodicity can be described easily by the two aforementioned functions. In case of the week periodicity it can be used either $\cos(2\cdot\pi\cdot d/7)$ and $\sin(2\cdot\pi\cdot d/7)$, or 7 binary digits (1000000 for Monday, 0100000 for Tuesday, etc.). The binary code cannot be used for a year practically.

E. 4th stage – Combination of input variables sets

From the previous step the sets of flows {$Q(t\text{-}1)$, $Q(t\text{-}2)$, $Q(t\text{-}3)$, $Q(t\text{-}4)$}, precipitation {$P(t)$, $P(t\text{-}1)$, $P(t\text{-}2)$, $P(t\text{-}3)$}, temperature {$T(t)$, $T(t\text{-}1)$, $T(t\text{-}2)$, $T(t\text{-}3)$}, and seasonality {$\cos(2\cdot\pi\cdot d/days_{year})$, $\sin(2\cdot\pi\cdot d/days_{year})$} have been defined. Thus 11 new ANNs are constructed which are presented in Table 4 (with serial number from 100 to 110). The available sets and optimization processes are repeated. In the present stage, the crucial parameters of the ANNs are same with those of the 3nd stage with the difference that the number of neurons of the hidden layer ranges from 1 to 40 with incremental step 1. After the training and calibration of the parameters of ANNs the various statistical indices are determined. Table 5 portrays key metrics of the finally created model 110.

In order to investigate the effectiveness of the proposed methodology and the input selection process, the produced best ANN versions (108, 109 and 110) were compared on the same datasets with two other methods. Initially, a simple auto-regressive (AR) model was selected for comparison purposes, since this method represents a candidate that creates simple and fast models. Due to the fact that autocorrelation appears to be significant in the available data, as it appears in Step 1 of the methodology, 9 previous values of Q were used to construct the model. Matlab's *ar* command was used to create the model with default settings (Forward-backward approach that minimizes the sum of a least- squares criterion for a forward model, and the analogous criterion for a time-reversed model).

A(q)y(t) = e(t) where A(q) = 1 - 0.7112 q(t-1) + 0.08531q(t-2) - 0.07419 q(t-3) - 0.09355q(t-4) + 0.009828 q(t-5) - 0.01135 q(t-6) - 0.07566 q(t-7) + 0.04437 q(t-8) - 0.08749 q(t-9)

Table 4. ANNs combinations and construction for 4th stage of proposed methodology – Final Selection

No. combination	No. ANN	Participation of flow set	Participation of precipitation set	Participation of temperature set	Participation of seasonality functions set	Population of input variables
1	99	No	No	No	Yes	2
2	95	No	No	Yes	No	4
3	100	No	No	Yes	Yes	6
4	80	No	Yes	No	No	4
5	101	No	Yes	No	Yes	6
6	102	No	Yes	Yes	No	8
7	103	No	Yes	Yes	Yes	10
8	65	Yes	No	No	No	4
9	104	Yes	No	No	Yes	6
10	105	Yes	No	Yes	No	8
11	106	Yes	No	Yes	Yes	10
12	107	Yes	Yes	No	No	8
13	108	Yes	Yes	No	Yes	10
14	109	Yes	Yes	Yes	No	12
15	110	Yes	Yes	Yes	Yes	14

Table 5. Statistical Indices Of The Proposed ANN With Input Variables {Q(T-1) , Q(T-2), Q(T-3), Q(T-4), P(T) , P(T-1), P(T-2), P(T-3), T(T) , T(T-1), T(T-2), T(T-3), Cos(2·Π·D/Daysyear), Sin(2·Π·D/Daysyear)} –Model 110

Statistical index	Training set	Evaluation set	Test set
RMSE	17,731	19,579	13,075
r	0,845	0,820	0,884
MAPE	33,751	33,493	39,499
MPE	0,456	4,747	2,421
VE	-0,117	0,667	-0,964
ME	0,381	-2,197	3,634
MF	-41,954	-51,213	-42,869
NMBE	-0,004	0,022	-0,036
NRMBE	0,756	0,765	0,924
E	0,714	0,67	0,772
E_1	0,676	0,67	0,683
$TS_{1\%}$	3,79	3,56	2,33
$TS_{2\%}$	15,89	15,53	13,62
$TS_{5\%}$	29,91	28,31	25,96
$TS_{25\%}$	55,80	55,16	49,91
$TS_{50\%}$	79,61	80,27	77,79
$TS_{100\%}$	95,32	95,80	92,78

The second method used to compare the presented approach is the creation of time series specialized ANNs, based on series-parallel implementation. The models of this kind take as input the previous values (potentially multiple time steps) of the flow Q, and they have the capability to take also exogenous inputs. In the specific case, only the time delay of the output was used as predictor, assuming that the influence of the other factors is already depicted in the Q value or in past historical patterns of the latter. This method was selected since it represents a more complex one than AR, but less-complex than the proposed one, exploiting only a single aspect of the data in the tradeoff of simplicity. For the implementation, the newnarxsp function of Matlab NN Toolbox was used. The dataset followed the already mentioned division of the previous two approaches, in order to compare the methods on the same final test set. The model parameters (types and number of neurons, training functions etc.) were optimized based on a Genetic Algorithm, following the defined methodology in (Kousiouris et al., 2012), adapted in the given dataset. A number of training functions were investigated from the available Matlab implementations and was concluded that the trainbr case (based on Bayesian regularization training) was the most beneficial case. This option was used in the main execution of the GA optimization process of the compared method. Options for the activation functions of the neurons included "tansig','logsig','purelin' and 'radbas' functions. The number of generations was set to 50 and the population size to 30, resulting in 1500 different ANNs being examined. Elite count was set to 2 and crossover at 0.8, with the GA returning the MSE on the training error as the evolutionary fitness criterion, which had proved more efficient in past experiments. During each loop, an ANN is saved only if its MSE on the intermediate evaluation set error is less than the current best one. The ANN with this best metric is selected directly in the end and it's cross-validated in the final 30% of the initial dataset that has not been used up to this point (test set).

The compared results in the training and evaluation set of the three approaches appear in Table 6, based on the metrics identified in (Dawson, 2002), on the final test set. The key metrics for comparison were RMSE (Root Mean Square Error) for insights in the high flows, MSRE (mean squared relative error) for insights in the low level flows, MAPE (Mean Absolute Percentage Error) as a general indicator, E and the Akaike information criterion (AIC) for investigation of model complexity. Any combination of the 11 indices used in the study may be used, of course affecting the final selection of the best model. The particular ones were selected given that they cover a range of features, from complexity to extremes identification (low or high flows), which is considered more critical from a hydrological point of view (e.g. predicting low flows that can lead to water shortage or optimal hydroplant operation to high flows indicating imminent flooding). From the comparison it is obvious that the proposed input selection method produces enhanced models in comparison to the other approaches (AR and GA-ANN model), especially in critical criteria such as the RMSE, which indicates the high level influence. This is considered more critical since it is able to better detect cases of increased flow, where corrective actions may be more necessary for mitigating flow peaks. On the other hand, the AR model portrays an enhanced overall behavior (as indicated by the enhanced MAPE), which however can be attributed to its superior performance in the low flow prediction, as also indicated by the reduced MSRE. Furthermore it has significantly poorer performance in the cases of RMSE and E. It also portrays the best performance in AIC, however this is more of an indication that it is simpler to build. The GA-ANN model seems to predict mainly underestimated values, thus leading to increased risk. For the final selection, given that each version (108, 109 or 110) may indicate better performance in a specific subgroup of the overall error criteria when compared to each other, the final candidate may be chosen based on a specific criterion that is considered as more important than the others. Based on the aforementioned analysis, model 110 prevails (in terms of RMSE and E).

Table 6. Comparative results on the test (validation) set on different metrics

Model	MAPE	E	AIC	RMSE	MSRE
Model 108	28,448	0,763	11431	13,348	0.357
Model 109	52,093	0,739	11651	13,994	0.986
Model 110	39,499	0,772	11433	13,075	0.611
GA-ANN	38.058	0.600	12482	17.331	0.348
AR Model	18.028	0.622	6652	16.837	0.268

4. Conclusions

Recently, several methods have more or less efficiently dealt with the selection of input variables to artificial neural network (ANN) models in the hydrology and water resources domain. While the ultimate purpose is to approximate an effective method accounting for non linear input variables to ANN models, very few approaches could reach this target. Moreover, none of these has considered as inputs precipitation and temperature coupled with atmospheric circulation patterns describing the variability of wet/dry and warm/cold weather influencing the river flows. In addition, none methodology has considered the

inherent seasonality of precipitation and temperature as input variable to ANN models. The seasonality had only been considered as wavelet transform for space–time pre-processing of satellite or ground station precipitation, evaporation, and runoff data in neural network of rainfall–runoff modeling. To this regard, the proposed methodology for selection of input variables involved to ANN models for river flow forecasting taking into consideration variables linked to atmospheric circulation and seasonality is multi-stage taking into account:

- The stepwise limitation of the input variables under selection allowing the non linear correlation between each input variable and the flow under forecasting by using the ANN. In this way, 110 different cases of ANN models have been examined instead of 17,640 potential cases.
- Simulation of annual seasonality through an appropriate pair of sine and cosine functions.

A set of hydrology criteria including *RMSE*, *r*, *MAPE*, *MPE*, *VE*, *ME*, MF, *NMBE*, *NRMBE*, *E*, and E1, was used through the technique of voting analysis. Each criterion served a specific purpose for the model selection. For example, the *TS* and *MPE* statistics measured the effectiveness of the model to accurately predict the data. The statistics *r*, *E*, *NMBE,* and *NRMBE* quantified the efficiency of the model in capturing the complex, dynamic and nonlinear precipitation- river flow process, while the global error statistics such as *r*, *E*, and *NRMBE* accounted for high flows due to involvement of square of the difference between observed and forecasted flows. The proposed methodology revealed as robust input variables the river flow of past 4 days, the precipitation of past 3 days and the seasonality. Its superiority has been proven by the comparison with GA-based ANN and AR- models, two methodologies that were selected following tradeoffs in the simplicity and time needed to create the models. Comparison on a critical subset of the criteria has indicated that the final model created by the proposed methodology portrays significant advantages especially in the more risky cases of high data flows and the increased need for their prediction to timely identify flooding phenomena. The latter may be attributed to the prevailing wet and warm circulation patterns at 700 hPa geo-potential heights influencing the river flows of the study catchment (Panagoulia et al., 2006a). In future work different kinds of training ANNs techniques will be applied for the results improvement, while different catchments river flows can be tested for generalization purposes. One limitation of the work is the training time needed for the investigation of e.g. 110 models in the main methodology used. However the model is trained once and may be used on a daily basis for the forecasts. Retraining may also be performed in the future in case of new data arrivals that indicate a difference in patterns, which is expected to be a rare occasion. Furthermore, parallel computational implementations of ANNs (e.g. like Apache Mahout or Spark ML) may be used in order to speed up the ANN investigation and training process.

References

Bowden G.J., Maier H.R. and Dandy, G.C. (2005), Input determination for neural network models in water resources applications, Part 2, Case study: forecasting salinity in a river. *Journal of Hydrology,* **301**, 93–107.

Chau, K-W., Wu, C.L. (2010), A hybrid model coupled with singular spectrum analysis for daily rainfall prediction, *Journal of Hydroinformatics* , **12.4**, 458-473.

Chen, X.Y., Chau, K.W., Busari, A.O. (2015), A comparative study of population-based optimization algorithms for downstream river flow forecasting by a hybrid neural network model, *Engineering Applications of Artificial Intelligence*, **46**, 258-268.

Davies, L., Gather, U. (1993),The identification of multiple outliers, *Journal of the American Statistical Association,* **88** (423), 782–792.

Dawson, C. W., Harpham, C., Wilby, R. L., and Chen, Y. (2002), Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China, *Hydrological Earth System Science*, **6**, 619-626, doi:10.5194/hess-6-619-2002,

Fernando T.M.K.G., Maier H.R., Dandy G.C. (2009), Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach, *Journal of Hydrology,* **367**,, 165-76.

Gholami, V., Chau, K-W., Fadaee, F., Torkaman, J., Ghaffari, A. (2015), Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers, *Journal of Hydrology*, **529**, 1060-1069.

Hippert, H. S., Pedreira, C. E., Souza R.C. (2001), Neural networks for short-term load forecasting: A review and evaluation, *IEEE Trans. on Power Systems*, **16 (1)**, 44-55.

Jain A., Prased Indurthy S.V. (2003), Comparative analysis of event-based rainfall-runoff modeling techniques – deterministic, statistical and artificial neural networks, *Journal of Hydrologic Engineering*, 93-98.

Kousiouris, G.A. Menychtas, D. Kyriazis, K. Konstanteli, S. Gogouvitis, G. Katsaros, T. Varvarigou, (2012),"Parametric Design and Performance Analysis of a Decoupled Service-Oriented Prediction Framework based on Embedded Numerical Software", *IEEE Transactions on Services Computing*, IEEE computer Society Digital Library, IEEE Computer Society.

Lintner, B. R., Chiang, J. C. H. (2007), Adjustment of the remote tropical climate to El Niño conditions, *Journal of Climate*, **20**, 2544–2557.

Maier H. R., Jain A., Dandy G.C., Sudheer K.P. (2010), Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Review paper, *Environmental Modelling & Software,* **25**, 891-909.

Maier, H.R., Dandy, G.C. (2001), Neural network based modelling of environmental variables: a systematic approach, *Mathematical and Computer Modelling* 33.

May, R.J., Maier, H.R., Dandy, G.C., Fernando, T.M.K.G, (2008), Non-linear variable selection for artificial neural networks using partial mutual information, Environmental Modelling & Software, **23,** (10–11), 1312–1326.

Nayak P.C., Venkatesh B., Krishna B., Jain S.K. (2013), Rainfall-runoff modeling using conce-ptual, data driven, and wavelet based computing approach, *Journal of Hydrology,* **493,** 57–67.

Panagoulia, D., Bárdossy, A., Lourmas, G. (2006b), Diagnostic statistics of daily rainfall variability in an evolving climate, *Advances in Geosciences*, **7**, 349–354.

Panagoulia, D., Bárdossy, A., Lourmas, G. (2008), Multivariate stochastic downscaling model for generating precipitation and temperature of climate change based on atmospheric circulation atmospheric circulation, *Global NEST Journal*, **10(2),** 263–272.

Panagoulia, D., Grammatikogiannis, A., Bárdossy, A. (2006a), An automated classification method of daily circulation patterns for surface climate data downscaling based on optimised fuzzy rules, *Global NEST Journal*, **8(3)**, 218–223.

Philander, S.G.H. (1990), El Niño, La Niña, and the Southern Oscillation, Academic Press, 293.

Sharma, A. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 – a strategy for system predictor identification, *Journal of Hydrology* **239**, 232–239.

Taormina, R., Chau, K-W. (2015), Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and extreme learning machines, *Journal of Hydrology* , **529**, 1617-1632.

Wang, W. C., Chau, K.-W., Xu, D.-M., Chen, X.-Y. (2015), Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition, *Water resource Management*, **29**, 2655-2675.

Wu, C.L., Chau, K.-W., Li, Y.S. (2009), Methods to improve neural network performance in daily flows prediction, *Journal of Hydrology* , **372**, 80-93.