Environ Sci Pollut Res (2015) 22:4230–4241
DOI 10.1007/s11356-014-3669-y

RESEARCH ARTICLE

Artificial neural network modelling of biological oxygen demand in rivers at the national level with input selection based on Monte Carlo simulations

Aleksandra Šiljić · Davor Antanasijević · Aleksandra Perić-Grujić · Mirjana Ristić · Viktor Pocajt

Received: 5 June 2014 / Accepted: 29 September 2014 / Published online: 5 October 2014

Abstract Biological oxygen demand (BOD) is the most significant water quality parameter and indicates water pollution with respect to the present biodegradable organic matter content. European countries are therefore obliged to report annual BOD values to Eurostat; however, BOD data at the national level is only available for 28 of 35 listed European countries for the period prior to 2008, among which 46 % of data is missing. This paper describes the development of an artificial neural network model for the forecasting of annual BOD values at the national level, using widely available sustainability and economical/industrial parameters as inputs. The initial general regression neural network (GRNN) model was trained, validated and tested utilizing 20 inputs. The number of inputs was reduced to 15 using the Monte Carlo simulation technique as the input selection method. The best results were achieved with the GRNN model utilizing 25 % less inputs than the initial model and a comparison with a multiple linear regression model trained and tested using the same input variables using multiple statistical performance indicators confirmed the advantage of the GRNN model. Sensitivity analysis has shown that inputs with the greatest effect on the GRNN model were (in descending order) precipitation, rural population with access to improved water sources, treatment capacity of wastewater treatment plants (urban) and treatment of municipal waste, with the last two having an equal effect. Finally, it was concluded that the developed GRNN model can be useful as a tool to support the decision-making process on sustainable development at a regional, national and international level.

Keywords GRNN · BOD · River water · MCS · MLR · Sustainability

Responsible editor: Michael Matthies

A. Šiljić · A. Perić-Grujić · M. Ristić · V. Pocajt
Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia

D. Antanasijević (✉)
Innovation Center of the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia
e-mail: dantanasijevic@tmf.bg.ac.rs

Introduction

The Technical Advisory Committee of the Global Water Partnership in 2002, at the Johannesburg World Summit on Sustainable Development (WSSD), defined integrated water resources management (IWRM) as a process which promotes the coordinated development and management of water, land and related resources in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems (Rahaman and Varis 2005). Surface water quality is an important segment of IWRM that has now been accepted internationally as the way forward for efficient, equitable and sustainable development and management of the world's limited water resources and for coping with conflicting demands (UNDESA 2014). On the other hand, following the 1992 Sustainable Development Strategy, the member states of the UN were obliged to integrate the principles of sustainable development into national policies and programs, i.e. to formulate, elaborate and adopt national strategies for sustainable development. The mechanism for the implementation of the strategy involves a set of indicators to be monitored, namely sustainable development indicators (SDI) determined to be internationally comparable. The only water quality indicator monitored at the national level is biochemical oxygen demand (BOD) in rivers. Eurostat, as the official European Statistical Office, collects BOD data from all European countries and publishes the indicator regularly among SDI, Theme 8—Natural resources. At this moment, attainable through

Eurostat, BOD data at a national level is available for 28 of 35 listed European countries for the period prior to 2008, among which 46 % of data is missing (Eurostat 2013a).

BOD is a measure of organic pollution in an aquatic system and it is inversely related to the dissolved oxygen (DO) level in water, i.e. high BOD values indicate a low level of DO or even anoxic conditions in water (Singh et al. 2009). Also, it represents the approximate amount of biodegradable organic matter in the water and provides a strong indication of the extent of water pollution (Marcotullio 2007). Therefore, it is established as the most significant water quality parameter needed to identify and implement strategies for the protection of water resources (Basant et al. 2010). In general, the BOD level depends on the wastewater discharge (industrial effluents, treatment plants, municipal untreated wastewater) and agricultural run-off (European Environment Agency EEA 2012a).

Regulations such as the Urban Wastewater Treatment Directive (European Economic Community EEC 1991a, EC 1998) and the Nitrates Directive (EEC 1991b) and the associated measures for the implementation of these directives have significantly reduced point source pollution in recent decades. Nevertheless, overflow of wastewater from sewage systems as well as by discharges from wastewater treatment plants and industries still keep polluting European water resources (EEA 2012b). River water pollution can be linked to the type of wastewater produced by urban, industrial and agricultural activities that flows into surface and subsurface waters (Vittori Antisari et al. 2010). Additionally, the increase in human population and economic activities has grown in scale in recent decades (Mustapha et al. 2013).

In recent years, data-driven computational intelligence techniques such as artificial neural networks (ANNs) took the lead ahead of physical-based methods, especially in hydrology (Awchi 2014). ANNs are able to capture nonlinear relationships between variables in a complex system by "learning" from available data presented to them. Thus, ANNs have become a powerful tool in predicting different parameters in aquatic ecosystems, e.g. for wastewater treatment process control and optimization (Mjalli et al. 2007; Kulkarni and Chellam 2010; Pendashteh et al. 2011). Usually, ANNs were applied separately for the prediction of BOD and DO, but simultaneous prediction of both parameters has also been investigated (Basant et al. 2010).

Key challenges in the development of an ANN-based model are the considered and effective selection of model inputs and architecture and the ongoing improvement of prediction performance during the selection process. With regard to the first challenge, if a large number of input variables are presented to an ANN model, it will usually increase the network size, slow the processing time and reduce the overall efficiency of the network (Arhami et al. 2013). Sensitivity analysis has been widely used to distinguish key inputs, to determine their relative importance on the output and to help to exclude inputs that do not have a significant effect on the performance of the ANN (Dogan et al. 2009, Ranković et al. 2010). Among others, Monte Carlo simulation (MCS) can be used for sensitivity analysis. The MCS technique is simple to implement, not time-consuming and was successfully applied to a wide range of problems, for example as a method for the water quality risk assessment with regard to exceeding regulatory limits (Jiang et al. 2013) and for uncertainty analysis (Shrestha et al. 2009, Arhami et al. 2013, Dehghani et al. 2014).

In the present study, widely available sustainability, economical and industrial indicators were used as inputs for the modelling of national BOD level. In order to obtain an optimized ANN model, with a reduced number of inputs and enhanced performance, the MCS technique was implemented as an input selection method. Finally, the best performing ANN model was analyzed using a set of statistical indicators and compared against a corresponding multiple linear regression (MLR) model.

Materials and methods

Input data

The dataset used in this study was generated based on the national level data reported by European countries, from two statistical databases: European Statistical Office (Eurostat 2013b) and World Bank (World Bank 2013). Initial inputs were selected among available economic, industrial, environmental and sustainability indicators that cover the following sectors: industry, agriculture, waste and water management, tourism and transport. The list of chosen indicators and their descriptive statistics for the period 2000–2008 are presented in Table 1. Where necessary, normalization was performed per capita, per hectare of arable land and per value of EU27, to allow comparison between countries. Descriptive statistics of BOD data at the national level for each country and for the entire BOD dataset are shown in Table 2.

The complete dataset was partitioned into subsets according to the availability of BOD data from Eurostat: the data from the years 2000–2007 was used for training and validation of the ANN model (140 data patterns), while the data from 2008 was used to test the model (19 data patterns, equaling 12 % of the available data).

Artificial neural network model development

Artificial neural network is a numerical algorithm inspired by the functioning of biological neurons. Neuron m receives an input signal vector x_i, where $i=1, 2, \ldots, L$, from total L input channels and then computes the weighted sum of components

Table 1 Descriptive statistics of inputs selected for BOD modelling

Input	Unit	Mean	St. dev.	Minimum	Maximum
Gross domestic product (GDP)[a]	–	0.97	0.68	0.11	2.96
Rural population (RPOP)	%	29.29	11.98	2.61	49.84
Rural population with access to improved water source (RPOPAW)	%	98.25	5.58	71.20	100.00
Total environmental taxes (ENVT)	%	7.28	1.49	4.13	10.69
Tourism intensity (TI)	Nights/1000 inhabitants	4680.21	3464.42	683.00	20,100.00
Municipal waste generation (MWG)	kg pc	503.94	138.45	239.00	794.00
Treatment of municipal waste (TMW)	%	92.72	7.81	75.00	109.00
Population connected to urban wastewater treatment with at least secondary treatment (PWT2)	%	69.37	25.19	12.30	113.87
Population without wastewater treatment (PWT0)	%	7.15	9.42	0.00	38.00
Treatment capacity of wastewater treatment plants (urban) (TCAP)	kg O_2/day/cap	0.08	0.17	0.00	0.76
Precipitation (PREC)	m^3 pc	7922.18	4680.44	2006.91	21,992.99
Industry value added (GDPI)	% of GDP	28.70	5.84	15.11	41.54
Environmental investment by industry (ENVI)	% of GDP	0.19	0.15	0.02	0.84
Agriculture value added (GDPA)	% of GDP	3.38	2.84	0.36	14.98
Use of inorganic fertilizers—phosphorus (UFP)	kg/ha	14.35	9.85	1.01	46.24
Use of inorganic fertilizers—nitrogen (UFN)	kg/ha	123.14	81.32	25.49	378.37
Arable land (ALAND)	%	29.59	12.08	8.34	56.61
Area under organic farming (ORG)	%	3.93	3.77	0.20	17.40
Livestock density index (LDI)	unit/ha	1.05	0.75	0.27	3.62
Freight inland waterway transport (FIWT)[a]	–	0.03	0.08	0.00	0.47

[a] Unitless because of normalization per value of EU27

x, multiplying each component x_i by the coefficient w_{mi} that reflects the importance of the input channel i (Fig. 1) (Cardoso et al. 2008).

The neuron m activation a_m, is given by the following expression:

$$a_m = \sum_{i=1}^{L} w_{mi} x_i + b_m \quad (1)$$

where b_m is the bias (correction) that allows positive activation a_m when x_i equals zero. The numerical output signal of neuron m (or neuron response), s_m, is based on the weighted sum of all inputs and results from the computation of an activation function $f(a_m)$. Different types of sigmoid functions are commonly used as activation functions, e.g. log-sigmoid (logistic) and symmetric log-sigmoid functions, along with linear or trigonometric functions, such as the sine and hyperbolic tangent (tanh) functions.

The neural network's ability to process the input data and produce output results is created by connecting the independent neurons into layers. The process of obtaining unknown coefficients w_{mi} and b_m required to approximate the prescribed function is called training, and the input dataset is called the training set. Supervised training comprises proposing initial values to the coefficients and then adjusting them in order to minimize the error between the predicted and actual output (Cardoso et al. 2008). The training set is used to fit ANN model weights, the validation set to select the model that provides the best level of generalization and the test set to evaluate the chosen model against the remaining data (Dehghani et al. 2014).

ANN is a powerful tool for modelling, especially in non-linear multivariable systems, where variables may have complex interrelationships that cannot be easily defined, like in water resource systems (Hadzima-Nyarko et al. 2014). The structure of the ANN implies usually three or four different layers: the input layer, the hidden layer(s) and the output layer, with no connections within the layers. In general, the ANN works by introducing the values of available inputs to the neurons of the first layer, where each input value is multiplied by a coefficient (weight) and forwarded to the neurons in the hidden layer in which the weighted sum is computed. Finally, the results are then forwarded to the output layer, where they are compared to the actual values and eventually presented as the ANN predictions.

The general regression neural network (GRNN) learning algorithm was reported to demonstrate good results in environmental modelling (Antanasijević et al. 2013a, b, 2014; Heddam 2014a, b). GRNN is a one-pass supervised learning network consisting of four layers with a variable number of neurons depending on the number of input/output variables

Table 2 Descriptive statistics of national level BOD data for the study period

Country/region	BOD (mg O_2/l)[a]			
	Mean	St. dev.	Minimum	Maximum
Belgium	3.06	0.76	2.19	4.41
Bulgaria	3.92	0.54	3.27	4.73
Czech Republic	3.32	0.29	2.87	3.84
Denmark	1.54	0.21	1.16	1.79
Germany	2.29	0.19	2.17	2.50
Ireland	1.15	0.23	0.95	1.68
Spain	3.09	0.69	2.36	4.49
France	2.23	0.61	1.38	3.22
Italy	2.94	0.58	2.48	4.05
Cyprus	2.87	1.58	1.50	4.30
Latvia	1.85	0.32	1.48	2.44
Luxembourg	2.79	1.19	1.47	4.66
Hungary	3.55	0.35	2.87	3.91
Netherlands	1.82	0.45	1.24	2.33
Austria	1.19	0.09	1.04	1.33
Poland	4.00	0.53	3.14	4.57
Romania	4.69	0.56	4.04	5.95
Slovenia	3.07	1.73	0.86	5.25
Slovakia	2.84	0.44	2.20	3.64
United Kingdom	1.81	0.08	1.66	1.91
BOD (mg O_2/l)[a]	2.71	1.16	0.86	5.95

[a] Data not available for Bulgaria 2006; Germany 2000–2004 and 2007; Cyprus 2000–2004; Latvia 2000; Luxembourg 2003; Hungary 2008; Netherlands 2003–2006; Romania 2000

and data patterns in the dataset used for training. In the case of the GRNN used in this study, the input layer consisted of 20 neurons (one for each input variable), the pattern layer had one neuron for each pattern in the training dataset (140 neurons), the summation layer was composed of two neurons (one for the only output variable plus one), and the output layer had one neuron for the output variable (Fig. 2).

The training process between the input layer and the hidden layer is composed of unsupervised learning, while the training between the hidden layer and the summation layer is composed of supervised learning, with the aim to minimize the difference (the mean square error) between the model output and target value (Kim and Kim 2008).

The input and hidden neurons were linked using a linear scaling function, while the hidden and summation neurons were connected using an exponential activation function:

$$f(D) = \exp\left(\frac{-D}{2\sigma^2}\right) \quad (2)$$

where D is the distance between training patterns calculated using Euclidean distance method and σ is the smoothing factor.

The accuracy of the GRNN is determined by a smoothing factor, which represents the width of the Gaussian curve calculated for each probability density function during the network training (Antanasijević et al. 2013b). The smoothing factor should be greater than 0 and may usually range from 0.01 to 1, to demonstrate what can be considered good results. An iterative algorithm is used to determine the proper smoothing factor (Ward Systems Group 2008).

Input selection approach

In general, there are three approaches for the selection of input variables used in ANN modelling (Shahin et al. 2008):

- Techniques based on a priori knowledge of the characteristics of potential input variables
- Model-based techniques that rely on training many ANNs with different combinations of input variables and selection of the network that has the best performance, and
- Model-free techniques that use linear-dependent measures, such as correlation, or nonlinear measures of dependence, such as mutual information, to obtain the significant model inputs prior to developing the ANN models

In this study, a model-based approach was used, because it has been shown to optimize the model by eliminating the inputs based on their real influence on the output results, rather than on the influence assumed by a statistical technique. To gauge the influence of input variables on the variations of model predictions, the MCS was applied. MCS input selection involves repeated generation of random input values from their probability distributions, as well as the creation and evaluation of a new GRNN model with a reduced number of inputs. The input which induces the smallest change on the output value has been eliminated from the dataset in each step of the selection procedure, as presented in more detail in "GRNN model development and MCS input selection." MCS was performed using Statistica 10 (StatSoft. Inc. 2010).

Multiple linear regression

Regression analysis is one of the most frequently used statistical techniques for modelling relationships between variables. The linear regression model consists of the regression variables, dependent variables and regression coefficients, and data for complex systems are described using MLR models (Awchi 2014). The general MLR equation is as follows:

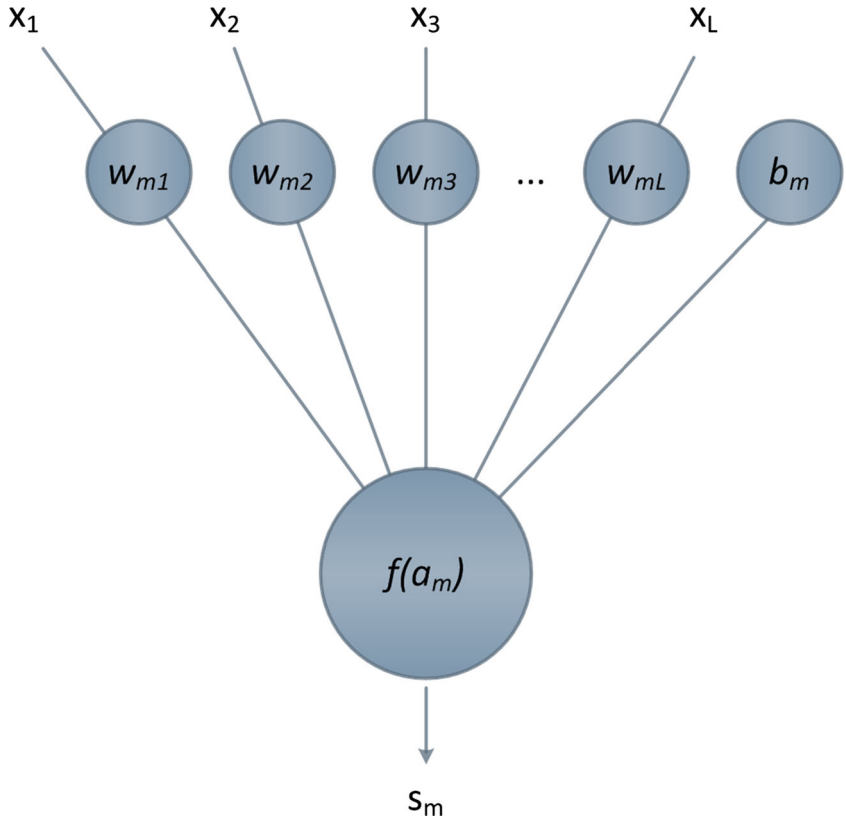

Fig. 1 Artificial neuron

$$y = c_0 + c_1x_1 + c_2x_2 + \ldots + c_nx_n \quad (3)$$

where y is the dependent variable (output), c_i (i=0,…., n) with the coefficients generally estimated by least squares and x_i (i=1,…, n) are the explanatory variables (inputs).

Performance indicators

The analysis of the GRNN model performance was based on the statistical comparison of outputs predicted by the model and observed/actual data. The performance of each model and its ability to produce accurate results were determined using the following criteria:

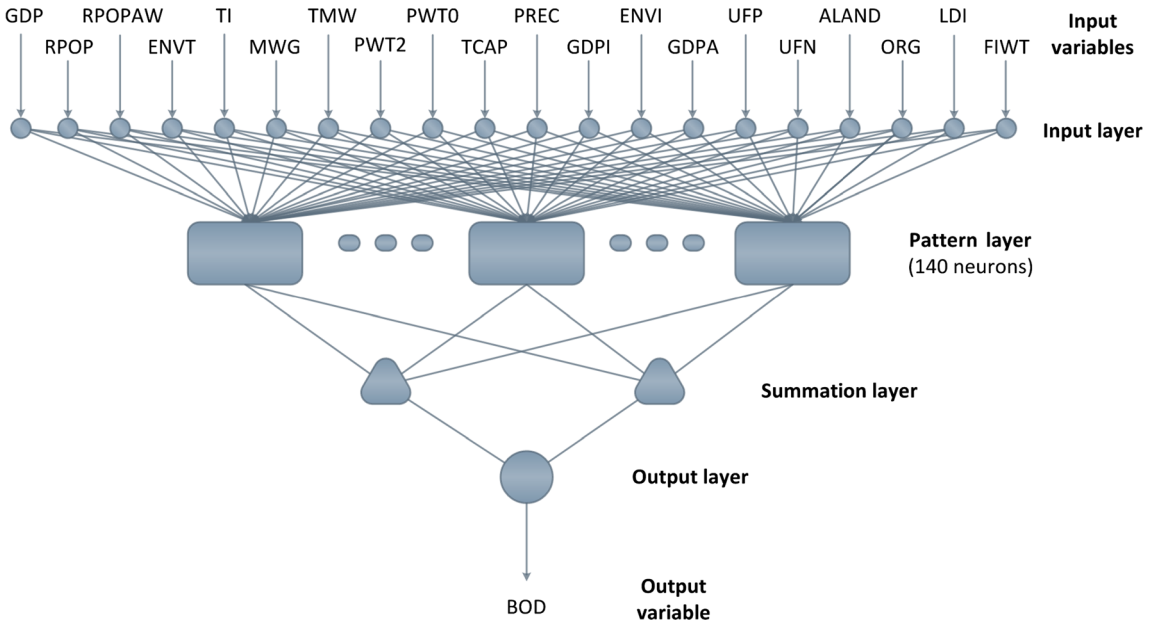

Fig. 2 GRNN architecture

- Index of agreement (IA)

$$IA = 1 - \frac{\overline{(C_p - C_o)^2}}{\left[\left|C_p - \overline{C_o}\right| + \left|C_o - \overline{C_o}\right|\right]^2} \quad (4)$$

- Percentage of predictions within a factor 1.25 (FA1.25) of the observed values

$$0.8 < \frac{C_p}{C_o} < 1.25 \quad (5)$$

- Mean absolute percentage error (MAPE)

$$\mathrm{MAPE} = 100\frac{1}{k}\sum\frac{|C_o - C_p|}{C_o} \quad (6)$$

- Root mean squared error (RMSE)

$$\mathrm{RMSE} = \left[\overline{(C_p - C_o)^2}\right]^{1/2} \quad (7)$$

- Mean absolute error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum|C_p - C_o| \quad (8)$$

where C_p and C_o are the predicted and observed BOD values, respectively.

IA, FA1.25 and MAPE were used as the key statistical parameters for the evaluation and comparison of models in this study. IA is a relative and bound measure that allows for cross comparisons between models and is limited to the range 0–1. FA1.25 shows the percentage of cases in which the ratio between predicted and observed value falls within the range 0.8–1.25, i.e. it indicates the proportion of data for which the model results are somewhat in proximity of the actual values. MAPE indicates the mean relative error of model predictions. The RMSE and MAE describe the average difference between model predictions and observations/actual values in the units of the observed value (Hadzima-Nyarko 2014). Because of the power term in the RMSE calculation, it is more sensitive to extreme values than that of MAE (Antanasijević et al. 2013b).

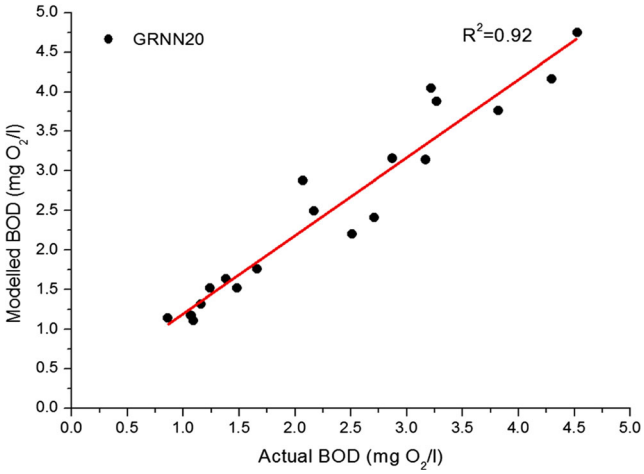

Fig. 3 The comparison of actual BOD values and BOD predictions obtained using the GRNN20 model (test data)

Results and discussion

GRNN model development and MCS input selection

In order to determine the optimal set of training parameters for the GRNN, several models were created, trained and tested with all available inputs. The GRNN model which was trained using linearly scaled input values, Euclidian distance metrics

Table 3 Simulated PDFs with the Kolmogorov-Smirnov test significance values

Input	PDF	Kolmogorov-Smirnov test Significance (*p*)
GDP	Folded normal	0.08012
RPOP	Gaussian mixture	0.44590
RPOPAW	Gaussian mixture[a]	0.00000
ENVT	Normal	0.78384
TI	Weibull	0.29407
MWG	Gaussian mixture	0.99010
TMW	General extreme value[a]	0.00307
PWT2	Gaussian mixture	0.14572
PWT0	Gaussian mixture[a]	0.00115
TCAP	General extreme value	0.28079
PREC	Log normal	0.67845
GDPI	Gaussian mixture	0.75093
ENVI	Gaussian mixture	0.79490
GDPA	Gaussian mixture	0.46568
UFP	Gaussian mixture	0.39239
UFN	Gaussian mixture	0.93511
ALAND	General extreme value	0.29402
ORG	General Pareto	0.24213
LDI	General extreme value	0.33630
FIWT	Gaussian mixture[a]	0.00000

[a] The best-ranked PDF

and iterative algorithm for the determination of the smoothing factor demonstrated the best performance (IA=0.97, FA1.25=84 % and MAPE=13 %). Since the GRNN model with 20 inputs (GRNN20) provided good agreement with the actual BOD values (Fig. 3), it can be concluded that the initial input selection was satisfactorily performed. The next step comprised the elimination of less significant inputs, with the aim to preserve or if possible improve the GRNN model performance. The GRNN model with the reduced number of inputs can be applied in a wider range of causes with an associated reduced time needed for data preparation and also with a reduced level of uncertainties.

The MCS input selection is based on the estimation of probability density functions (PDFs) and on re-sampling of the input values for each input used in the model, according to the obtained PDFs. Since many PDFs are tested in the MCS analysis, the Kolmogorov-Smirnov non-parametric test is often used to test the null hypothesis that two independent samples are not different according to their distribution characteristics (Dehghani et al. 2014). Table 3 shows the PDFs obtained for the 20 inputs used in this study and the values of Kolmogorov-Smirnov test which was used for the selection of the best fitting PDF, based on its significance (p).

The MCS dataset, used for the determination of the influence of inputs on the performance of the GRNN model to accurately predict the national BOD level, was generated by re-sampling 50 values for each of the 20 inputs in the defined range (Table 1). Therefore, the MCS dataset was constructed with 20 blocks of 50 patterns per input (1000 data patterns), where each block had one input with MCS values within the defined range, while other inputs were set to the mean of measured values (Table 1). The quantification of influence relating to a specific input on the model output was performed by calculating the BOD range, as the difference between

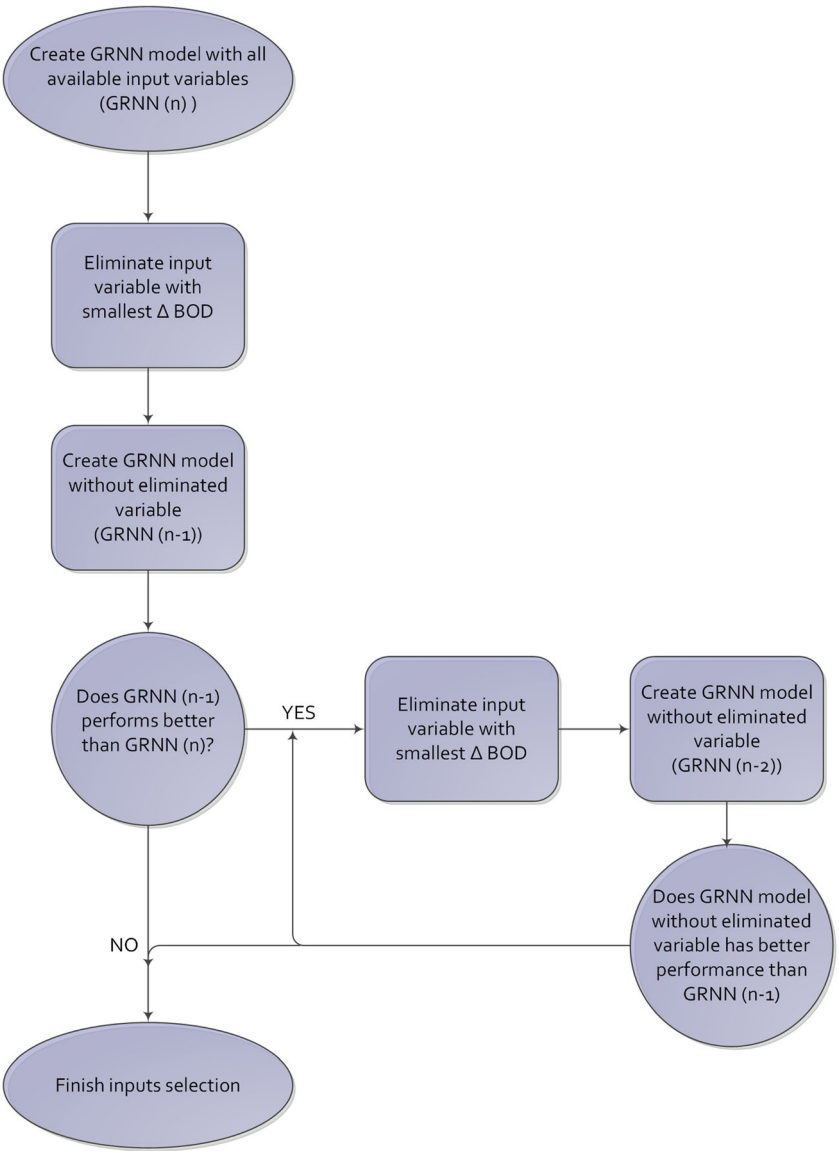

Fig. 4 Schematic representation of the elimination procedure in input variable selection

Table 4 ΔBOD values (mg/l) obtained during the MCS input selection process

Inputs	Model						
	GRNN20	GRNN19	GRNN18	GRNN17	GRNN16	GRNN15	GRNN14
GDP	0.946	0.958	0.489	0.760	0.762	1.029	0.721
RPOP	0.933	0.926	1.055	1.067	1.067	1.294	2.622
RPOPAW	1.764	1.905	2.112	2.085	2.085	2.100	1.923
ENVT	0.939	0.805	0.787	0.886	0.883	0.856	0.682
TI	0.980	1.004	1.098	1.098	1.103	1.108	1.349
MWG	1.004	1.006	1.121	1.120	1.120	1.104	0.872
TMW	2.045	2.008	1.831	1.919	1.919	1.917	1.832
PWT2	1.165	1.184	1.239	1.231	1.234	1.223	1.223
PWT0	0.950	1.248	1.261	1.261	1.261	1.268	1.252
TCAP	1.932	1.919	1.820	1.916	1.916	1.921	1.531
PREC	2.180	2.180	2.180	2.180	2.180	2.179	2.278
GDPI	1.369	1.256	0.752	0.621	*0.561*	–	–
ENVI	*0.213*	–	–	–	–	–	–
GDPA	1.033	1.060	1.079	1.071	1.084	1.090	0.595
UFP	0.660	*0.584*	–	–	–	–	–
UFN	1.371	1.215	0.844	1.041	1.040	1.009	1.014
ALAND	0.709	0.620	0.482	0.749	0.754	*0.633*	–
ORG	1.669	1.435	1.058	1.036	1.036	1.041	2.106
LDI	0.972	0.890	*0.207*	–	–	–	–
FIWT	0.577	0.586	0.311	*0.339*	–	–	–

Italic numbers indicate the minimum ΔBOD value

maximum and minimum predicted BOD value ($\Delta BOD = BOD_{max} - BOD_{min}$) for each block in the MCS dataset. Therefore, the significance of each input to the model is proportional to the ΔBOD value. The MCS input selection procedure is schematically summarized in Fig. 4.

During the input selection process, the number of neurons in the hidden layer and output layer were the same as in the case of the GRNN20 model, while the number of neurons in the input layer varied between 20 and 14, with each subsequent model having one input neuron less than the previous one, since the number of inputs was also reduced in the same manner. ΔBOD values obtained during the MCS input selection are shown in Table 4. The values of model performance indicators and inputs excluded for each particular GRNN model are presented in Table 5. As it can be observed, the best results were obtained with the GRNN15 model, which was trained with the number of inputs reduced by 25 % comparing to the initial GRNN20 model, i.e. without environmental investment by industry (ENVI), use of inorganic fertilizers—phosphorus (UFP), livestock density index (LDI), freight inland waterway transport (FIWT) and industry value added (GDPI).

Table 5 Performance indicators of models with different number of inputs

Model label	Inputs	Performance indicators				
		IA	FA1.25 (%)	MAPE (%)	RMSE (mg/l)	MAE (mg/l)
GRNN20	All	0.97	84	13	0.36	0.27
GRNN19	ENVI excluded	0.97	84	13	0.36	0.27
GRNN18	UFP excluded	0.98	95	12	0.35	0.26
GRNN17	LDI excluded	0.98	95	12	0.35	0.26
GRNN16	FIWT excluded	0.98	95	12	0.35	0.26
GRNN15	GDPI excluded	*0.98*	*95*	*12*	*0.35*	*0.26*
GRNN14	ALAND excluded	0.98	89	12	0.35	0.26

Italic numbers indicate the best performance

The optimized GRNN15 model predictions of BOD levels were in good agreement with the actual (measured) BOD values (Fig. 5a) and have a reduced divergence (Table 5 and Fig. 5b) in comparison with the initial GRNN20 model.

Except for Luxembourg, all predictions of BOD level obtained using the GRNN15 model for the test dataset (2008) fall within the FA1.25 boundaries. Luxembourg is a very specific country, with the size, structure of economy and hydrological infrastructure significantly different from those of other EU countries. Therefore, Luxembourg was represented within the training data set with a very limited number of learning patterns, which in turn significantly impaired the ability of GRNN model to produce accurate predictions.

Comparison with MLR model

A multiple linear regression (MLR) model was developed for comparison purposes, using the same data as for the GRNN15 model (Table 6).

The performance of the MLR model (IA=0.87, MAPE=27 %, RMSE=0.64 mg/l, MAE=0.51 mg/l) (Fig. 6) is inferior in comparison with the GRNN15 model (Table 5 and Fig. 5). The R^2 coefficient for the GRNN15 model prediction compared with actual values was 0.92 (Fig. 5a), while for the corresponding MLR model it was only 0.69 (Fig. 6a), showing that the GRNN model performed considerably better compared to the linear MLR model.

Also, in the case of the MLR model, only 47 % of predictions fall within the FA1.25 range, while the value of relative errors were between −25 %, in the case of Cyprus, and 94 % for the Netherlands (Fig. 7). The maximum value of relative error for GRNN15 predictions is 40 % and it is obtained for Luxembourg. The BOD levels for France and Slovenia predicted by the MLR model also strongly deviate from the actual BOD values (relative error ≥70 %), while in case of the GRNN15 model prediction for those countries, relative errors were collectively less than 20 % (Fig. 7).

Table 6 The coefficients of MLR model with the standard error

Input	Coefficient value	Standard error
GDP	7.71×10^{-1}	2.74×10^{-1}
GDPA	1.41×10^{-1}	5.76×10^{-2}
ENVT	-1.09×10^{-1}	7.05×10^{-2}
TCAP	-1.07×10^{-1}	7.50×10^{-1}
ORG	-6.53×10^{-2}	2.54×10^{-2}
MWG	-3.35×10^{-2}	1.20×10^{-2}
RPOPAW	2.51×10^{-2}	1.84×10^{-2}
RPOP	2.09×10^{-2}	9.99×10^{-3}
PWT2	-1.17×10^{-2}	4.37×10^{-3}
ALAND	-6.74×10^{-3}	1.04×10^{-2}
TMW	-3.09×10^{-3}	1.10×10^{-3}
UFN	-2.90×10^{-3}	1.84×10^{-3}
PWT0	7.50×10^{-4}	1.21×10^{-2}
PREC	-1.07×10^{-4}	2.11×10^{-5}
TI	5.66×10^{-6}	3.68×10^{-5}
Intercept	6.32	2.17

Sensitivity analysis

Sensitivity analysis was performed to determine the effect of the inputs on the final GRNN15 model. The most significant input in the GRNN15 model, regarding the obtained ΔBOD value (Table 4), is precipitation (PREC) and it is followed by rural population with access to improved water source (RPOPAW), treatment capacity of wastewater treatment plants (urban) (TCAP) and treatment of municipal waste (TMW). In the periods of high precipitation, pollutants are being washed away from agricultural fields, while low or no precipitation in

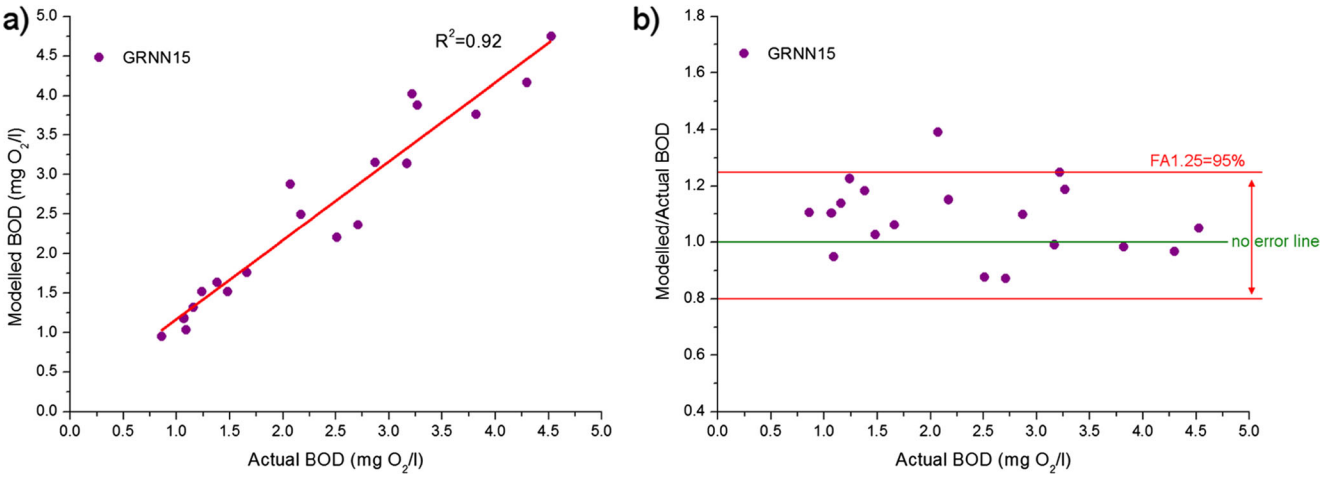

Fig. 5 Performance of the GRNN15 model for test data: **a** agreement of actual and modelled BOD and **b** FA1.25 plot

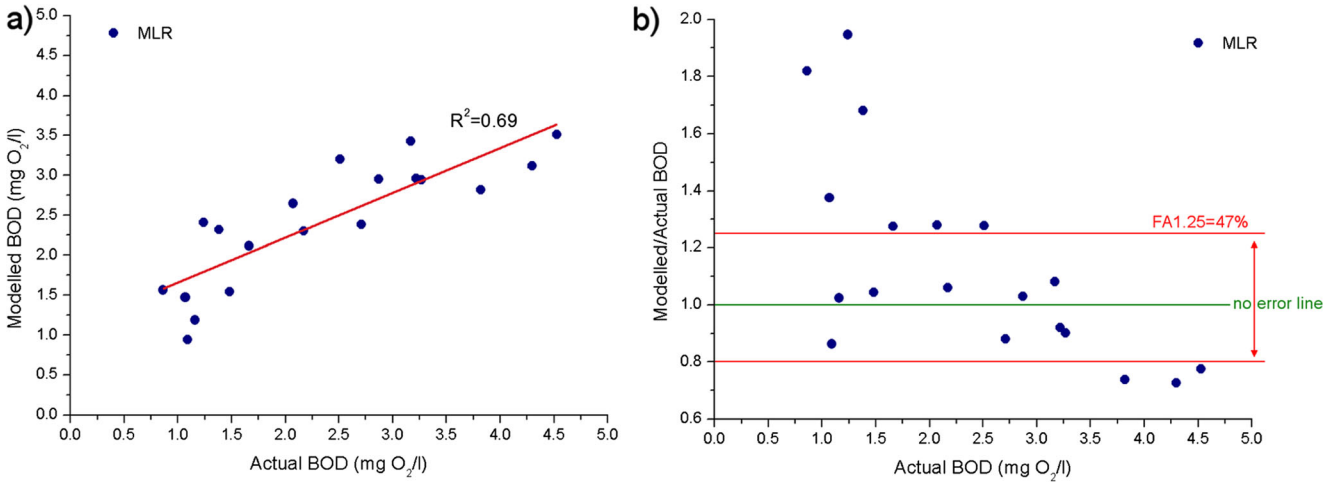

Fig. 6 Performance of the MLR model for test data: **a** agreement of actual and modelled BOD and **b** FA1.25 plot

dry periods influences limited migration of polluting matter, but negatively impacts water level, so less water is available to dilute organics brought by wastewater discharge. The rural population having access to an improved water source was close to or equal to 100 % in most cases, except in the case of Romania, where BOD stays among the highest values over the examined period. This contribution of TCAP can be attributed to the fact that the higher the capacity to treat wastewater, the better the quality of effluent will be when discharged to recipients. Likewise, the larger the share of municipal waste treated, the lower the quantity of landfill waste and hence a direct effect on the leachate draining to local streams.

Conclusions

This paper reports the development of an artificial neural network (ANN) model based on sustainability, economical and industrial indicators for the prediction of annual BOD values at a national level in rivers, and its optimization using the Monte Carlo simulation (MCS) technique for the selection of inputs. The BOD model was created using a general regression neural network (GRNN) architecture, trained with the data from 20 EU countries for the years 2000–2007 and tested with the data from 2008. The performance of the GRNN model was evaluated using the index of agreement (IA), mean absolute percentage error (MAPE), the mean absolute error (MAE), the root mean squared error (RMSE) and the percentage of predictions within a factor of 1.25 of the observed values (FA1.25).

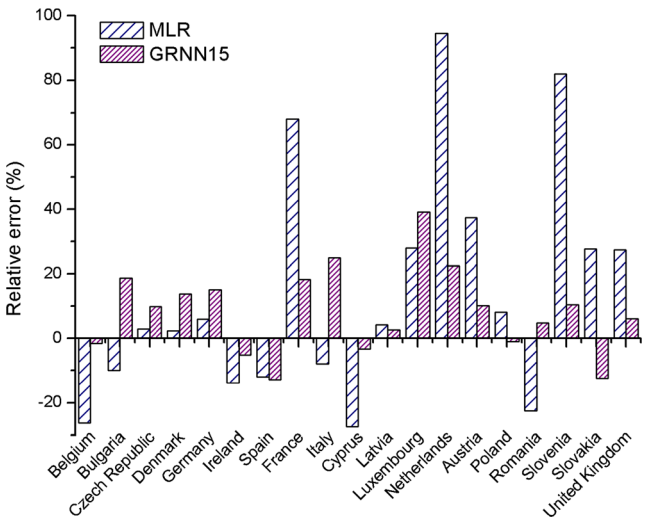

Fig. 7 Comparison between the GRNN15 and MLR model based on relative error for test data

The performance of the optimized GRNN model (IA= 0.98, FA1.25=95 %), obtained after the elimination of inputs using the MCS procedure, was improved in comparison with the non-optimized ANN model (IA=0.97, FA1.25=84 %), which included all available input variables. It can be concluded that MCS input selection is an efficient technique for the elimination of less important inputs, whose application in this case resulted in a more accurate model with 25 % less inputs needed. A comparison of the optimized ANN model with a conventional MLR shows that the ANN model has much better forecast performance (R^2=0.92) than the MLR model (R^2=0.69) when both models were trained and tested using the same datasets and input variables. Sensitivity analysis has shown that inputs with the greatest effect on the GRNN model were (in descending order) precipitation, rural population with access to improved water source, treatment capacity of wastewater treatment plants (urban) and treatment of municipal waste, with the last two having equal effect.

The developed GRNN model can be applied for the prediction of BOD values in cases where BOD data is missing, for the prediction of future BOD values, and even for the simulation of possible consequences of regulatory, infrastructural and industrial changes on the BOD level in rivers, thus

supporting the decision-making process on sustainable development at a regional, national and international level.

Further research is planned in the application of different ANN architectures and optimization techniques to explore opportunities for further improvement of results and prediction capabilities, in extending the application of the model to non-European countries with a diversified level of development (e.g. USA, China, India) and in the development of models for BOD forecasting at both a regional and local level.

Acknowledgements The authors are grateful to the Ministry of Education, Science and Technological Development of the Republic of Serbia, Project No. 172007, for the financial support.

References

Antanasijević D, Pocajt V, Povrenović D, Ristić M, Perić-Grujić A (2013a) PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci Total Environ 443:511–519

Antanasijević D, Ristić M, Perić-Grujić A, Pocajt V (2013b) Forecasting human exposure to PM10 at the national level using an artificial neural network approach. J Chemometr 27:170–177

Antanasijević D, Ristić M, Perić-Grujić A, Pocajt V (2014) Forecasting GHG emissions using an optimized artificial neural network model based on correlation and principal component analysis. Int J Greenh Gas Con 20:244–253

Arhami M, Kamali N, Mahdi Rajabi M (2013) Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. Environ Sci Pollut R 20: 4777–4789

Awchi TA (2014) River discharges forecasting in northern Iraq using different ANN techniques. Water Resour Manag 28:801–814

Basant N, Gupta S, Malik A, Kunwar P, Singh K (2010) Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water—a case study. Chemom Intell Lab 104:172–180

Cardoso J, Almeida J, Dias J, Coelho P (2008) Structural reliability analysis using Monte Carlo simulation and neural networks. Adv Eng Softw 39:505–513

Dehghani M, Saghafian B, Nasiri Saleh F, Farokhnia A, Noori R (2014) Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. Int J Climatol 34: 1169–1180

Dogan E, Sengorur B, Koklu R (2009) Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. J Environ Manage 90:1229–1235

European Commission (EC) (1998) Directive 1998/15/EC of 27 February 1998 amending Council Directive 91/271/EEC with respect to certain requirements established in Annex I. Off J Eur Communities 41: 29–30

European Economic Community (EEC) (1991a) Council Directive 91/271/EEC of 21 May 1991 concerning urban waste-water treatment. Off J Eur Communities 34:40–52

European Economic Community (EEC) (1991b) Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources. Off J Eur Communities 34:1–8

European Environment Agency (EEA) (2012a) Oxygen consuming substance in rivers (CSI 019), http://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/oxygen-consuming-substances-in-rivers-5. Accessed 7 Feb 2014

European Environment Agency (EEA) (2012b) European waters—assessment of status and pressures (EEA Report No 8/2012), Office for Official Publications of the European Union, Luxembourg. doi: 10.2800/63266

Eurostat (2013a), Sustainable development indicators, theme 8: natural resources, http://epp.eurostat.ec.europa.eu/portal/page/portal/sdi/indicators/theme8. Accessed 17 Dec 2013

Eurostat (2013b), Statistics, http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes. Accessed 17 Dec 2013

Hadzima-Nyarko M, Rabi A, Sperac M (2014) Implementation of artificial neural networks in modeling the water-air temperature relationship of the River Drava. Water Resour Manag 28:1379–1394

Heddam S (2014a) Generalized regression neural network-based approach for modelling hourly dissolved oxygen concentration in the Upper Klamath River, Oregon, USA. Environ Technol 35:1650–1657

Heddam S (2014b) Modelling hourly dissolved oxygen concentration (DO) using dynamic evolving neural-fuzzy inference system (DENFIS)-based approach: case study of Klamath River at Miller Island Boat Ramp OR USA. Environ Sci Pollut R 21:9212–9227

Jiang Y, Nan Z, Yang S (2013) Risk assessment of water quality using Monte Carlo simulation and artificial neural network method. J Environ Manage 122:130–136

Kim S, Kim HS (2008) Neural networks and genetic algorithm approach for nonlinear evaporation and evapotranspiration modeling. J Hydrol 351:299–317

Kulkarni P, Chellam S (2010) Disinfection by-product formation following chlorination of drinking water: artificial neural network models and changes in speciation with treatment. Sci Total Environ 408(19): 4202–4210

Marcotullio PJ (2007) Urban water-related environmental transitions in Southeast Asia. Sustain Sci 2:27–54

Mjalli FS, Al-Asheh S, Alfadala HE (2007) Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. J Environ Manage 83(3):329–338

Mustapha A, Zaharin Aris A, Juahir H, Firuz Ramli M, Umar Kura N (2013) River water quality assessment using environmentric techniques: case study of Jakara River Basin. Environ Sci Pollut R 20: 5630–5644

Pendashteh AR, Fakhru'l-Razi A, Chaibakhsh N, Abdullah LC, Madaeni SS, Abidin ZZ (2011) Modeling of membrane bioreactor treating hypersaline oily wastewater by artificial neural network. J Hazard Mater 192:568–575

Rahaman MM, Varis O (2005) Integrated water resources management: evolution, prospects and future challenges. Sustain: Sci Pract Policy 1:15–21

Ranković V, Radulović J, Radojević I, Ostojić A, Lj Č (2010) Neural network modeling of dissolved oxygen in the Gruza reservoir, Serbia. Ecol Model 221:1239–1244

Shahin MA, Jaksa MB, Maier HR (2008) State of the art of artificial neural networks in geotechnical engineering. Electron J Geotech Eng 8:1–26

Shrestha D, Kayastha N, Solomatine D (2009) A novel approach to parameter uncertainty analysis of hydrological models using neural networks. Hydrol Earth Syst Sc 13:1235–1248

Singh K, Basant A, Malik A, Jain G (2009) Artificial neural network modeling of the river water quality—A case study. Ecol Model 220: 888–895

StatSoft. Inc., 2010. Statistica (data analysis software system), version 10. Tulsa, USA

United Nations Department of Economic and Social Affairs (UNDESA) (2014) Integrated Water Resources Management (IWRM), http://www.un.org/waterforlifedecade/iwrm.shtml. Accessed 30 Mar 2014

Vittori Antisari L, Trivisano C, Gessa C, Gherardi M, Simoni A, Vianello G, Zamboni N (2010) Quality of municipal wastewater compared to surface waters of the river and artificial canal network in different areas of the eastern Po Valley (Italy). Water Qual Expo Health 2:1–13

Ward Systems Group (2008) Neuroshell 2 Help. http://www.wardsystems.com/manuals/neuroshell2/index.html?idxhowuse.htm. Accessed 10 Feb 2014

World Bank (2013) World DataBank. http://databank.worldbank.org/data/databases.aspx. Accessed 17 Dec 2013