Applying Monte Carlo and classification tree sensitivity analysis to the Zayandehrood River

Nader Nakhaei and Amir Etemad-Shahidi

ABSTRACT

Water quality modeling is an important issue for both engineers and scientists. The QUAL2K model is a simulation tool that has been used widely for this purpose. Uncertainty and sensitivity analysis is a major step in water quality modeling. This article reports application of Monte Carlo analysis and classification tree sensitivity analysis in the modeling of the Zayandehrood River. First the model was calibrated and validated using two sets of data. Then, three input values (stream flow, roughness and decay rate) were considered for both analyses. The Monte Carlo analysis was conducted using triangular distribution of probability of occurrence for the input parameters. The classification tree analysis classifies outcome values into non-numeric categories. Considering the relationships between the input parameters in the classification tree analysis is the most important advantage of it. The analyses demonstrated the key input variables for three points of the river. The dissolved oxygen levels were mainly sensitive to the decay rate coefficient along the river.

Key words | classification tree analysis, Monte Carlo analysis, QUAL2K, Zayandehrood River

Nader Nakhaei
Amir Etemad-Shahidi (corresponding author)
Hydro-Environment COE,
School of Civil Engineering,
Iran University of Science and Technology,
P.O. Box 16765-163,
Tehran,
Iran
E-mail: *etemad@iust.ac.ir*

LIST OF SYMBOLS

A_c	the cross-sectional area
c_i	constituent concentration
CDF	cumulative distribution function
CV	coefficient of variation
D	deviance at nodes
E	dispersion coefficient
F_{ox}	attenuation due to low oxygen
k	decay rate
k_a	reaeration rate
n	roughness coefficient
n_{ij}	number of cases which are assigned to class j at node i
o	oxygen concentration
P	the wetted perimeter
p_{ij}	probability distribution at i over the classes j
pdf	probability density function
Q	stream flow
S	sources and sinks of the pollutant including reactions and mass transfer
S_0	bottom slope
SD	standard deviation
t	time
V	volume
W	loading of pollutant

doi: 10.2166/hydro.2011.094

INTRODUCTION

Nowadays water quality management has become a serious issue due to the large amount of agricultural, municipal and industrial wastewaters which are discharged into water bodies. Rivers are one of the main water sources that suffer from these anthropogenic effects mainly because of ease of access to them and also because of their potential for self-purification. One of the widely used tools to predict river water quality is using computer models to simulate the stream and pollutants fate and transport. Using these models, not only the status quo can be simulated, but also, the future management options can be simulated before applying them.

Water quality models such as QUAL series have been used successfully in different modeling investigations (e.g. Palmieri & Carvalho 2006, Mannina & Viviani 2010). A water quality modeling study on the Nakdong River in Korea, performed by both QUAL2K and QUAL2E, demonstrated that the latest version is more accurate than the former (Park & Lee 2002). It has been also reported that this model can be used as a good tool for determining reaction kinetics of a river (Ghosh 1996). Application of QUAL2K for water quality modeling in the Bagmati River showed that the model represented the field data accurately (Kannel *et al.* 2007). In their study various water quality management options to control DO, such as pollution loads modification and local oxygenation were investigated. In some cases with insufficient data for using complex 2D or 3D models QUAL2K can be perfectly suitable (Paliwal *et al.* 2007). It has also been reported that a simplified model and a more complicated model such as Mike 11 yield similar results (Radwan *et al.* 2003).

There is an argument stating that as a model becomes more complex in terms of the increased number of parameters and variables, the error between simulations and measurements decreases and the overall model sensitivity increases (Snowling & Kramer 2001). It has been also argued that the most complex model is not necessarily the most useful one (Lindenschmidt 2006).

In parallel with modeling, uncertainty and sensitivity of input and output data must be investigated in order to have a more accurate estimation of water quality. Uncertainty analysis is a way to describe the model and parameter uncertainties. Along with uncertainty analysis, sensitivity analysis will enable modelers to estimate the range of likely outputs according to the selected range of input variables. The ranges of input variables which are selected for sensitivity analysis can be given to the model to calculate the outcomes and determine the most effective variables. The contribution to output uncertainty by an input is related to: (a) uncertainty of the input variable and (b) sensitivity of the output to that particular input (Mishra *et al.* 2003). It is reported that uncertainty should be considered throughout the modeling study starting from the very beginning (Refsgaard *et al.* 2007). The inevitability of significant uncertainty and the need to account for it in water quality management have been recognized in the development of some modeling tools. For example, QUAL2EUNCAS (Brown & Barnwell 1987) and DESERT (Ivanov *et al.* 1996) have this option. Considering these analyses lead to more accurate prediction of scenarios which needs to be simulated. There are several methods for sensitivity and uncertainty analysis, such as the Bayesian forecasting system (e.g. Georgakakos & Krzysztofowicz 2001; Krzysztofowicz & Herr 2001) and GLUE (Beven & Binley 1992). Mannina (2010) showed that the GLUE method can be used successfully for uncertainty assessment. The Monte Carlo simulation (MCS) is a common way for uncertainty and sensitivity analysis which is a suitable method for hydrological modeling (McIntyre *et al.* 2003). In hydrological applications, uncertainty analysis is often taken to be synonymous to MCS (Mishra 2009). In many investigations MCS has been applied for hydrological uncertainty and sensitivity analysis. In the Yamuna River, located in India, this method indicates satisfactory results on simulated data (Paliwal *et al.* 2007), In the Kennet River, the advantages and limitations of this method has been illustrated (McIntyre *et al.* 2005). MCS analysis has also been used to indicate model structure errors (Wagener *et al.* 2003; Smith & Wheater 2004) and assessing significance of input and calibration data errors (McIntyre & Wheater 2004). MCS has also been used for uncertainty analysis in small river basins (Marsili-Libelli & Giusti 2008). There are many traditional one parameter (at a time) sensitivity analysis which distributes a parameter in an acceptable range and calculates the corresponding changes in the output (Anderson & Woessner 1992). Applying this analysis for several parameters may lead to an approximate comparison between output correspondent distributions for each input variable. Such analysis can only be valid locally because local sensitivity analysis can not consider two or more input parameter's sensitivity at the same time (Mishra *et al.* 2009). Therefore, global sensitivity analysis which has the ability to perform sensitivity analysis with more than one input–output association needs to be used. In this way, the comparison between key parameters is much easier and more accurate. An example of this method is the Stepwise rank regression analysis which can be used for building input–output models to identify key contributors to output variance (Mishra *et al.* 2009). Stepwise rank regression analysis delineated by Helton (1993) has been used for Waste Isolation Pilot Plan project

(Helton & Anderson 1999) and has also been reported to be useful for ground water modeling (Mishra *et al.* 2009). Another global sensitivity analysis method is Mutual Information (Entropy) Analysis which can determine the strength of nonmonotonic relations of input–output (Mishra *et al.* 2009). This analysis has been used in combination with contingency table analysis (Mishra *et al.* 2003). Application of this method showed capability of it in selecting key parameters in neural network based input–output modeling (Bonnlander & Weigend 1994). The Regional sensitivity analysis (RSA) is another widely used global sensitivity analysis method in hydrological modeling (e.g. McIntyre *et al.* 2003; Mertens *et al.* 2005; Pappenberger *et al.* 2008; Tang *et al.* 2007; Jacquin & Shamseldin 2009). The RSA method is mainly concerned with the effect of unilateral variations of the parameters (Saltelli *et al.* 2004). Another global sensitivity analysis introduced by Sobol (1993), known as Sobol's variance decomposition (SVD), has received increasing attention mostly from hydrologists (e.g. Francos *et al.* 2003; Wang *et al.* 2006; Ratto *et al.* 2007; Tang *et al.* 2007; Jacquin & Shamseldin 2009). There are also other global sensitivity analysis such as Fourier amplitude sensitivity test (FAST) (Cacuci & Ionescu-Bujor 2004) and partial rank correlation coefficient (Frey & Patil 2002).

Classification tree analysis is also a global sensitivity analysis that demonstrates effect of parameters and/or combination of them on output values. This method can report the sensitivity by categorizing the outcomes in non-numeric factors. It has been used widely for medical decision making, agriculture (McQueen *et al.* 1995) and well log analysis (Perez *et al.* 2005). The capability of this method in global sensitivity analysis in order to justify the key parameters has been described in detail by Mishra *et al.* (2003).

This paper reports uncertainty and sensitivity analysis for the Zayandehrood River, in Iran, with Monte Carlo and Classification tree analysis, respectively. First-order reliability analysis on the decay rate and the reaeration rate and other parameters has been applied for this river using the QUAL2E model (Abrishamchi *et al.* 2005). The computer model applied for modeling is the U.S. Environmental Protection Agency's QUAL2K. The aims of this study are (a) to use both Monte Carlo uncertainty analysis and classification tree analysis as a new global sensitivity analysis tool, (b) to use classification tree analysis on Monte Carlo simulated model probabilistic realizations, and (c) to investigate the key sensitive parameters at different points on the Zayandehrood River.

STUDY AREA AND METHODOLOGY

Study area

The Zayandehrood River is one of the biggest rivers in the Middle East which is located in the center of Iran around Isfahan city. It emanates from Dimeh spring which is located within 140 km distance from Isfahan. The river's length is about 360 km with a drainage area around 4200 km^2 (Figure 1). The Zyandehrood reservoir, located 110 km from Isfahan city, is vital for the region because of agricultural and drinking usages and also due to the entertainment attraction for people who live nearby. From the Zayandehrood Reservoir an adjusted flow is released. Water flows in a valley until Kaleh Bridge (roughly 100 km) where there are no agricultural farms. The river water has a good quality until Kaleh Bridge because the area is mountainous and there is no waste discharge. Afterwards, the Zayandehrood River flows toward Varzaneh village and then divides into several branches and runs through Ghavkhuni Swamp (Figure 1).

There are many industrial plants that pollute the river directly or indirectly. There are also several cities and villages along the river which discharge their raw or partially treated wastewater into the river (Abrishamchi *et al.* 2005). The Abshar treatment plant, located in east of the Isfahan city, discharges municipal wastewater of Isfahan and some other smaller cities after treatment into the river. However, its performance is poor because of its low capacity and sometimes raw waste overflows into the river. From this point onwards, the quality of water begins to deteriorate rapidly as DO (dissolved oxygen) concentrations become low and BOD (biochemical oxygen demand) levels rise. These problems and Zayandehrood River's vital role indicate that water quality management is required in order to simulate its quality and also to prevent the water quality of getting deteriorated.

The modeling tool (QUAL2K)

Qual2k is the newest version of Qual series which can simulate streams and river quality in a one-dimensional steady

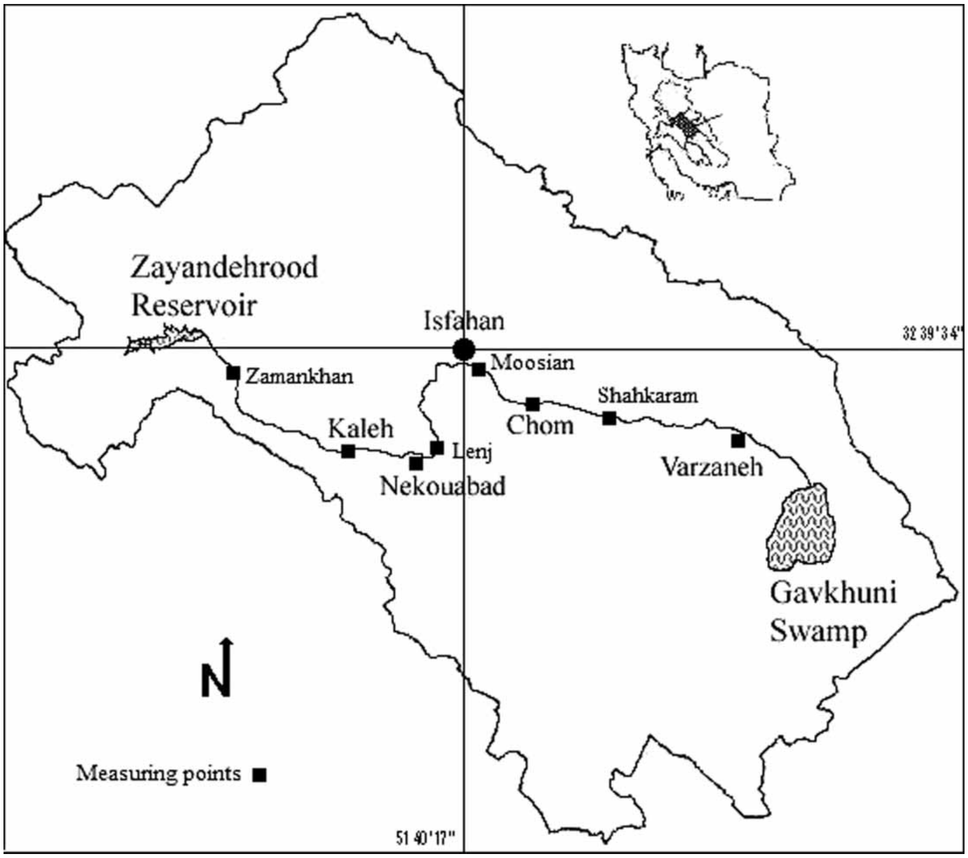

Figure 1 | The Zayandehrood River, its drainage area and the measurement stations.

state condition. In the Qual2k model the river is divided into several reaches and each reach is divided into equal segments. These segments are the model's shortest elements. This model can simulate fate and transport of many parameters and contaminants such as temperature, BOD, DO, phytoplankton, various kinds of nutrients, pH and etc (Chapra *et al*. 2007).

Qual2k has a general mass balance (Equation (1)) for a constituent concentration, c_i in a water column of reach i (Figure 2):

$$\frac{dc_i}{dt} = \frac{Q_{i-1}}{V_i}c_{i-1} - \frac{Q_i}{V_i}c_i - \frac{Q_{ab,i}}{V_I}c_i + \frac{E_{i-1}}{V_i}(c_{i-1} - c_i) + \frac{E_i}{V_i}(c_{i+1} - c_i) + \frac{W_i}{V_i} + S_i \tag{1}$$

where Q_i = flow at reach i ($m^3\,s^{-1}$), $Q_{ab,i}$ = abstraction flow at reach i ($m^3\,s^{-1}$), V_i = Volume of reach i (m^3), W_i = loading of pollutant to reach i (mg s^{-1}), S_i = sources and sinks of the pollutant including reactions and mass transfer (mg s^{-1} m^{-3}), E_i = dispersion coefficient between reaches i and $i+1$ ($m^3\,s^{-1}$), E_{i-1} = dispersion coefficient between reaches i –1 and i ($m^3\,s^{-1}$) and t = time (s).

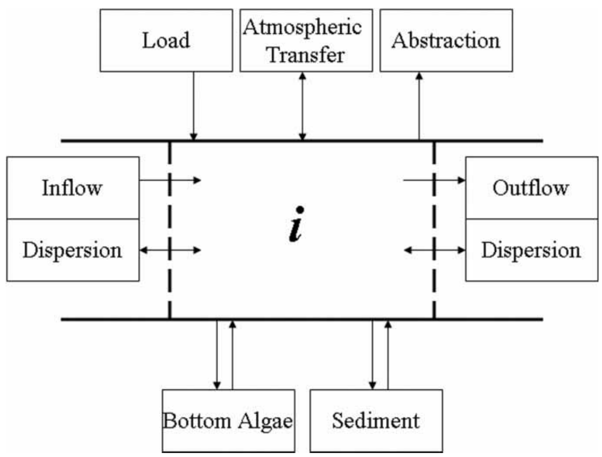

Figure 2 | Mass balance in reach *i* in QUAL2K.

The Manning equation which is used to express the relationship between flow and depth in the model is

$$Q = \frac{S_0^{1/2}}{n}\frac{A_c^{5/3}}{P^{2/3}} \tag{2}$$

where $S_0 =$ bottom slope, $n =$ the roughness factor (s m$^{-1/3}$), $A_c =$ the cross-sectional area (m^2) and $P =$ the wetted perimeter (m).

The sources and sinks (S_i) of DO in the Qual2k model in this investigation are formulated as

$$S_i = k_a(o_s - o) - F_{ox}k\,c_{BOD} \tag{3}$$

where $k_a =$ the reaeration rate (s^{-1}), $o_s =$ saturation concentration of oxygen (mg m^{-3}), $o =$ concentration of oxygen (mg m^{-3}), $F_{ox} =$ attenuation due to low oxygen (dimensionless), $k =$ decay rate of BOD (s^{-1}), $c_{BOD} =$ concentration of BOD (mg m^{-3}).

There are several reaeration formulas which can be used in the Qual2k model. However, Chapra *et al.* (2007) has expressed that they might be inappropriate for some cases and the reaeration rate can be manually calibrated.

Monte Carlo simulation analysis

Monte Carlo Simulation (MCS) analysis is based on the characterizing input distribution. The probability density function (*pdf*) shows the distribution of each input values. This distribution can be uniform which means that there is an equal likelihood of occurrence for parameter values. On the other hand, the normal distribution assumes that the probability of occurrence of the input is like a bell-shaped curve. It has been suggested to use prior knowledge based on previously monitored events for probability distribution of parameters if possible (Freni & Mannina 2010). The triangular distribution can be used to approximate the normal distribution (Chapra 1997).

The first step for uncertainty analysis based on MCS is to calculate the *pdf* due to boundaries of the selected input. Afterwards, the cumulative distribution function (*CDF*) is

$$CDF(x) = \int_{-x}^{x} pdf(x) \tag{4}$$

Since the area under the *pdf* is equal to 1, the *CDF* varies from 0 to 1. *CDF*(x) demonstrates the probability that the input parameter will be less than x. In fact, this function is the key to perform the MCS analysis. A set of random numbers between 0 and 1 can be chosen and propagated through the *CDF* in order to calculate values of x. In this way, calculated values are distributed properly. This method is also called Inverse Transform method and the details are given by Rubinstein (1981) and Fishman (1995). Subsequently, the random input values have to be used in the model to simulate distributions of output variables. MCS results are generally presented in the form of *CDF*s and/or histograms. For time-dependent models, the *CDF* can be extracted at different time slices (Mishra 2009).

Classification tree analysis

Classification tree analysis is a global sensitivity analysis which enables modelers to categorize the outcomes based on the set of input variables. This partitioning occurs with binary splits that are applied by the suitable classifier (Mishrat *et al.* 2003). Classification tree requires three basic elements (Breiman *et al.* 1984): selection of splits, determining when to continue the split or to terminate the node (where the split occurs) and assigning a class to a terminal node (leaf). The selection of splits must be in a way to provide maximum increase in the purity of the leaves. A leaf is called completely pure is no misclassified data in it. Several methods can define the purity of the nodes. In this investigation the classifiers are selected based on an overall reduction in deviance (a measure for purity of the leaves) for all possible splits over all the input values (Breiman *et al.* 1984). The deviance at node i is defined as

$$D_i = -2\sum_j n_{ij}.\ln(p_{ij}) \tag{5}$$

where n_{ij} is the number of cases which are assigned to class j at node i, p_{ij} is the probability distribution at i over the classes j that can be estimated for each node:

$$p_{ij} = \frac{n_{ij}}{n_i} \tag{6}$$

The total reduction in deviance from splitting node s into nodes t and u is

$$D_s - D_t - D_u = 2\sum_j \left[n_{tj} \ln\left(\frac{n_{tj}n_s}{n_{sj}n_t}\right) + n_{uj} \ln\left(\frac{n_{uj}n_s}{n_{sj}n_u}\right)\right] \quad (7)$$

The tree is built based on maximum reduction in deviance. The termination of splitting occurs in two conditions: (1) number of cases at a node becomes less than a set minimum, or (2) maximum reduction for a node drops under a set minimum (Breiman *et al.* 1984).

As the tree is built, each leaf refers to a particular factor describing output values. In this way, it is possible to determine the important input parameters. The earlier splits demonstrate the more important parameters because they refer to the most reduction in deviance and are more important in classification process. Later splits may also be important or may fall in the range of statistical noise. It is better to reduce the number of splits (prune) to the point that only invaluable parameters are left (Mishra *et al.* 2003). Pruning can be done by increasing the minimum reduction in necessary deviance for splitting. The result of classification tree can be shown in 2D and/or 3D scatter plots other than the tree itself. Usually the scatter plots demonstrate most important inputs (Mishra *et al.* 2003). This helps to identify and compare the key values visually.

RESULTS AND DISCUSSION

The river was divided into 18 reaches and (from the reservoir till the swamp) with different characteristics (Table 1). The measured hydraulic characteristics of the river were stream flow, velocity, depth of the water and the top width of the river. There were nine measurement stations in January 2003 and eight measurement stations in July 2005 for hydraulic characteristics (Table 1). The concentration of DO and BOD

Table 1 | Details of reaches and measured parameters

Reach\Measuring point upstream	Length (km)	Slope	Measured top width in Jan 2003 (m)	Measured top width in July 2005 (m)	Measured depth in Jan 2003 (m)	Measured depth in July 2005 (m)	Measured velocity in Jan 2003 (m/s)	Measured velocity in July 2005 (m/s)	Stream flow in Jan 2003 (cms)	Stream flow in July 2005 (cms)
1\Tanzimi (Reservoir)	25	0.0022	62.50	63.70	0.865	1.124	0.341	0.992	14	60.2
2	25	0.0022	57.00	58.50	–	–	–	–	–	–
3\Zamankhan	31.5	0.0021	57.00	58.50	0.283	0.970	0.830	1.400	–	–
4	31.5	0.0021	35.00	44.00	–	–	–	–	–	–
5\Kaleh	13.65	0.0026	35.00	44.00	0.351	1.164	0.377	1.280	–	–
6	13.65	0.0026	25.00	41.50	–	–	–	–	–	–
7\Nekouabad	2.31	0.0017	25.00	41.50	0.550	0.56	0.391	1.20	–	–
8	2.31	0.0017	24.00	44.00	–	–	–	–	–	–
9\Lenj	14	0.0018	24.00	44.00	0.296	0.580	0.925	1.20	–	–
10	14	0.0018	31.00	36.00	–	–	–	–	–	–
11\Moosian	17.5	0.0013	31.00	36.00	0.200	0.840	0.864	1.792	–	–
12	17.5	0.0013	22.00	15.00	–	–	–	–	–	–
13\Chom	17.5	0.00077	22.00	15.00	0.541	0.580	0.588	0.520	–	–
14	17.5	0.00077	22.00	15.00	–	–	–	–	–	–
15\Shahkaram	35	0.00077	22.00	15.00	0.37	–	0.5	–	–	–
16	35	0.00077	19.00	7.00	–	–	–	–	–	–
17\Varzaneh	15.75	0.0008	19.00	7.00	0.180	0.600	0.684	0.600	–	–
18\downstream: swamp	15.75	0.0008	19.00	7.00	–	–	–	–	–	–

Table 2 | Measured concentrations of DO and BOD

Distance from downstream (km)	Measured DO in Jan 2003 (mg/l)	Measured BOD in Jan 2003 (mg/l)	Measured DO in July 2005 (mg/l)	Measured BOD in July 2005 (mg/l)
340.80	8.47	6.00	7.50	5.00
292.40	8.38	5.00	7.00	4.00
232.40	8.35	8.00	8.10	6.00
211.00	8.81	26.00	5.00	18.00
180.10	8.91	12.00	6.50	20.00
170.00	8.90	18.00	6.00	16.00
156.30	8.26	7.00	5.20	14.00
136.50	7.40	27.00	3.00	18.00
109.12	8.25	30.00	3.00	15.00
70.88	8.66	25.00	2.40	17.00
61.50	8.45	24.00	1.60	17.00
49.01	9.25	27.00	1.50	20.00
28.17	8.20	23.00	1.00	21.00
23.00	–	–	1.00	22.00

were also measured in two separate seasons (Table 2). The water quality parameters were measured at 13 stations in January 2003 and at 14 stations in July 2005. Since the Qual2k is a steady model, only fixed values of discharge, DO and BOD were specified as the upstream boundary condition.

Two sets of data (dry and wet seasons) were used to calibrate and validate the model. The correlation coefficient (r) was used to evaluate the performance of modeling. The calibration and validation of the model were conducted based on the comparison of the measured and modeled velocities, water depths, DO and BOD levels. The first step was to calibrate the roughness factor using the measured velocities and water depths (Figures 3(a) and (b)). As seen, the measured and modeled values are similar. The calibrated roughness factors were then validated using the wet season data (Figures 3(c) and (d)). Then the decay rate of BOD and the reaeration rate were calibrated and validated based on the measured BOD and DO

Table 3 | Calibrated values of roughness factor, decay rate and the reaeration rate in each reach of the Zayandehrood River

Reach number	Roughness factor (n)	Decay rate (k)	Reaeration rate (k_a)
1	0.0560	0.05	0.700
2	0.0350	0.05	0.700
3	0.0350	0.05	0.200
4	0.0450	0.05	0.200
5	0.0450	0.07	0.650
6	0.0500	0.07	0.550
7	0.0350	0.10	0.850
8	0.0250	0.10	0.850
9	0.0250	0.45	0.550
10	0.0250	0.45	0.550
11	0.0250	0.45	0.350
12	0.0250	0.30	0.350
13	0.0250	0.10	0.200
14	0.0250	0.10	0.200
15	0.0280	0.10	0.200
16	0.0260	0.10	0.240
17	0.0260	0.40	0.240
18	0.0250	0.40	0.250

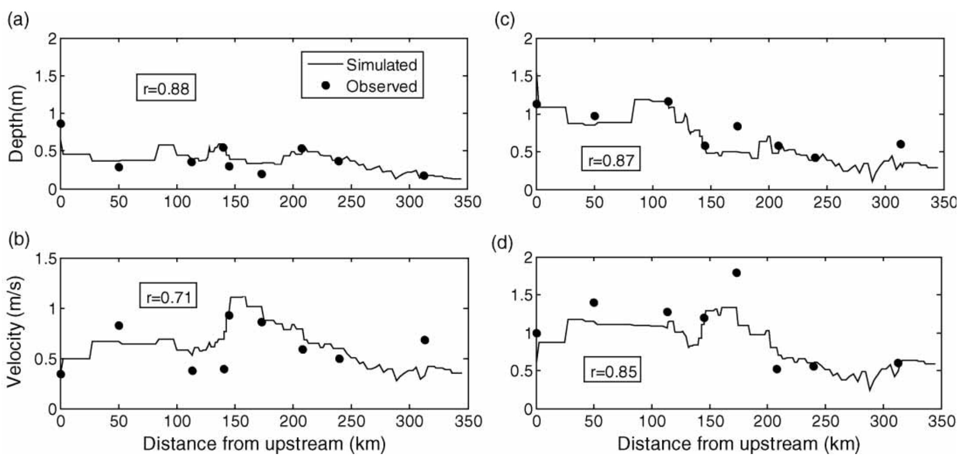

Figure 3 | (a) Simulated and measured depth of the river in the calibration period (January 2003), (b) Simulated and measured velocity in the calibration period (January 2003), (c) Simulated and measured depth of the river in the validation period (July 2005) and (d) Simulated and measured velocity in the validation period (July 2005).

concentrations. Table 3 shows the calibrated values of the mentioned parameters in each reach. Figures 4(a) and (b) show the result of calibration for DO and BOD concentration in January 2003, respectively. As seen, the DO concentration is high along the river except the last 50 km. This is because of large amounts of waste discharges which increase the BOD level. Figures 4(c) and (d) show the validation results of DO and BOD concentration in July 2005. In this case the DO levels are lower mainly at last 200 km of the river and BOD concentration increases gradually. This was caused by the large quantity of abstractions in this part of the year and more waste discharges. As seen in Figure 4, the accuracy of measured and simulated data based on the *r* value is acceptable as discussed by Kannel *et al.* (2007). In brief, the Q2K model was successful in simulating the DO and BOD levels. From this point onwards the uncertainty and sensitivity methods were conducted for July 2005.

Application of Monte Carlo analysis

In application of MCS analysis for the Zayandehrood River, three stations were selected to determine the DO level. These three points were about 258 km, 172 km and 85 km from downstream. The first step for MCS analysis is to choose the distribution for each key input variable in order to calculate the *pdf*. The distributions of input parameters were chosen to be triangular as an approximation of normal distribution. Three input values have been taken into consideration for uncertainty and sensitivity analysis: flow rate (Q), decay rate (k) and roughness factor (n).

The effect of stream flow and BOD decay rate on the uncertainty of BOD and DO founded to be dominant in the Zayandehrood River (Abrishamchi *et al.* 2005). In addition, due to frequent dredging of river bed and other projects which affect the river bed, the roughness factor may vary especially in the last 200 km. Therefore, the uncertainty and the sensitivity of the Zayandehrood River to these parameters is a matter of concern. Typical ranges (Chapra 1997; Chapra *et al.* 2007) were selected for decay rate and roughness factor in order to be used in the triangular distribution. The boundaries for the stream flow were calculated through available daily measured flows. The maximum measured flow and the minimum measured flow have been selected as the extremes. According to selected values and the probability distribution the following *pdf*s were considered:

$$pdf(Q) = 0.00167347Q - 0.101245 \quad 60.5 \leq Q(m^3s^{-1}) < 85 \tag{8}$$

$$pdf(Q) = -0.00167347Q + 0.183245 \quad 85 < Q(m^3s^{-1}) \leq 109.5 \tag{9}$$

$$pdf(n) = 130.6126n - 3.2653 \quad 0.025 \leq n < 0.1125 \tag{10}$$

$$pdf(n) = -1930.6126n + 26.12251 \quad 0.1125 \leq n \leq 0.2 \tag{11}$$

$$pdf(k) = -19.755k + 9.878 \quad 0.05 \leq k(d^{-1}) < 0.275 \tag{12}$$

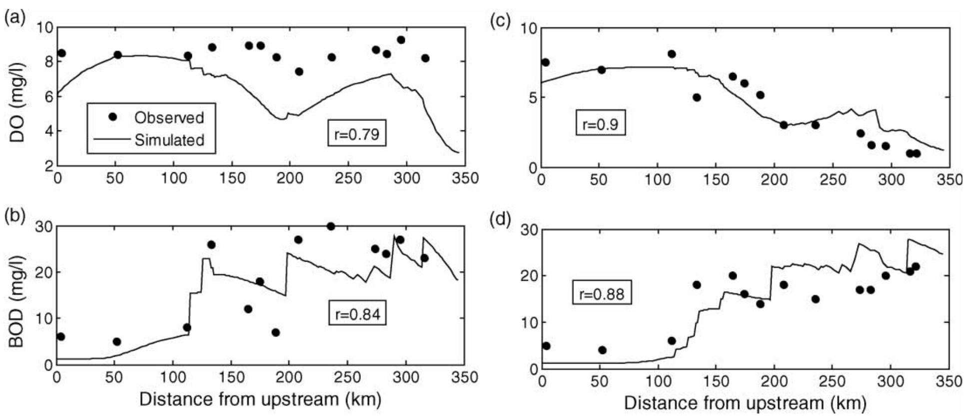

Figure 4 | (a) Simulated and measured DO concentration in the calibration period (January 2003), (b) Simulated and measured BOD concentration in the calibration period (January 2003), (c) Simulated and measured DO concentration in the validation period (July 2005), (d) Simulated and measured BOD concentration in the validation period (July 2005).

$$pdf(k) = 19.755k - 0.98777 \quad 0.275 \leq k(d^{-1}) \leq 0.5 \tag{13}$$

The *CDF*s were calculated using Equation (4). Random numbers between 0 and 1 were spread through the *CDF* to obtain a set of random input values. One hundred values were generated in this way and then these values were used to simulate concentration of DO. Using more than 100 values has been tested and yielded similar results. In this way, 900 DO values were produced. The distribution of outcome values demonstrated the sensitivity of the model to each input parameter. Nine (three for each input) histograms according to probability of occurrence and ranges of DO were then generated from these outputs (Figure 5). The focus of this part was to determine the most sensitive parameter by analyzing the statistical properties (such as Coefficient of Variation and skewness) of DO distributions. As seen, histograms of the outcome values are skewed generally. Apparently, the decay rate is the most significant input variable because the Coefficient of Variation (CV) of the DO is higher when varying the decay rate than those of others in different stations (Figures 5(b), (e) and (h)). However, the calculated CVs in the upstream imply that the model is more sensitive to the roughness factor than the stream flow. On the other hand, it shows that the stream flow is more important than the roughness factor in the other points. The concentration of DO in the upstream is not widely distributed (Figures 5(a), (b) and (c)) mainly because a fixed DO value is considered as the upstream boundary condition and it takes some kilometers for DO histograms to be widened. Therefore, the DO concentration does not vary significantly in the upstream taking into account the uncertainty in the parameters. As a result of that, the skewness of histograms may not be a matter of concern in the upstream. The CV due to the variation of the decay rate is higher than those of the others. When the decay rate is changed, a marginal reduction of DO amount can be seen in the distribution of it in the upstream, showing the low sensitivity of DO to this parameter. The *pdf* of DO (when stream flow varies) at the second point (172 km from downstream) is left-skewed (Figure 5(d)). The variation of DO versus stream flow at this point shows that, as the stream flow increases the DO levels increases as well. However, this increasing trend levels off as the stream flow becomes very high (Figure 6(a)). This causes more occurrences of higher DO values (the right side of the *pdf*). The *pdf* of DO (when decay rate varies) in this point can be approximated by a

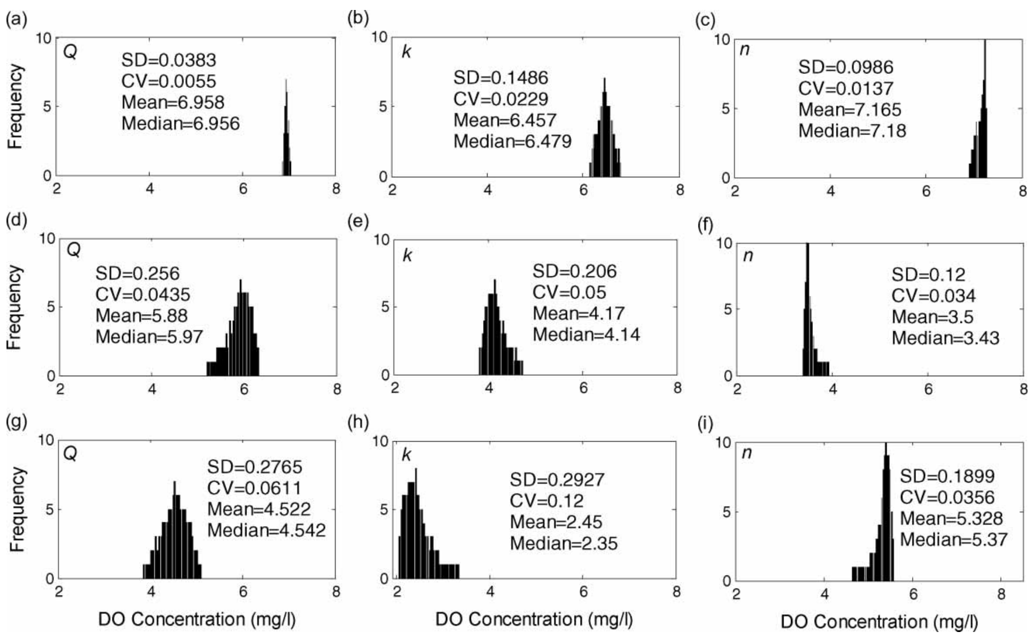

Figure 5 | *pdf* of DO (a) 258 km from the downstream when stream flow varies, (b) 258 km from the downstream when decay rate varies. (c) 258 km from the downstream when roughness factor varies. (d) 172 km from the downstream when stream flow varies. (e) 172 km from the downstream when decay rate varies. (f) 172 km from the downstream when roughness factor varies. (g) 85 km from the downstream when stream flow varies. (h) 85 km from the downstream when decay rate varies. (i) 85 km from the downstream when roughness factor varies.

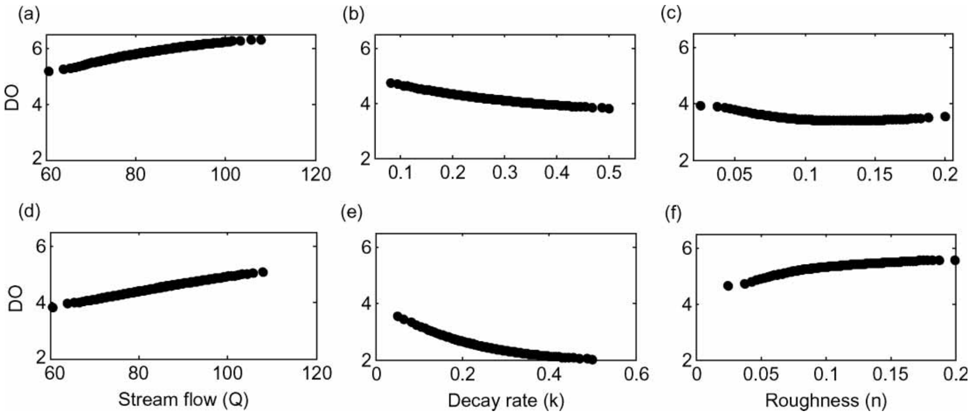

Figure 6 | (a) Relationship between stream flow and DO, 172 km from downstream. (b) Relationship between decay rate and DO, 172 km from downstream (c) Relationship between roughness coefficient and DO, 172 km from downstream (d) Relationship between stream flow and DO, 85 km from downstream. (e) Relationship between decay rate and DO, 85 km from downstream. (f) Relationship between roughness coefficient and DO, 85 km from downstream.

triangular (Figure 5(e)). The variation of DO versus decay rate in this point demonstrates that the DO level is inversely related to the decay rate (Figure 6(b)). Therefore, the mode happened in the middle of the DO range. The *pdf* of DO, when roughness varies, is right-skewed (Figure 5(f)). The variation of DO versus roughness coefficient shows that as the roughness increases (until roughly 0.125) the DO levels decreases due to increase of depth and as a result of that decrease of reaeration rate. From this point onwards, there is an increasing trend of DO (Figure 6(c)). This increase of DO is because of increase of traveling time (due to increase of roughness) and therefore, reduction of BOD in the downstream. The reduction of BOD causes the DO levels to increase as the traveling time increases. At this point of the river (172 km from downstream) according to calculated CVs, the DO is more sensitive to the decay rate and after that the key parameters are stream flow and the roughness, in that order.

The *pdf* of DO (when decay rate varies) is right-skewed in the point 85 km from the downstream (Figure 5(h)). This is because DO decreases with a convex curvature (Figure 6(e)). In the case that roughness coefficient varies, this trend is different (Figure 5(i)). Low flows in the downstream are due to the abstractions along the river. Therefore, in this part the depth of the water and as a result of that the reaeration is more important than the roughness. As the roughness increases, the DO levels increasing trend levels off (Figure 6(f)). This explains the low sensitivity of DO to the roughness coefficient in this point. The reason for increasing trend of DO is same as that of the last point. As the roughness increases, the velocity decreases and therefore the traveling time increases. This causes a decrease in the amount of BOD and DO in the upstream. However, in the downstream of the river the reaeration and lower BOD cause an increase in DO concentration in comparison with cases with lower values of roughness. In the downstream, as the stream flow increases, the DO concentration in this part increases too (Figure 6(d)). This increase is obviously because of the increase in the velocity and increase of reaeration rate. The *pdf* in this situation implies that the increasing rate of dissolved oxygen is steady so the maximum frequency occurs approximately in the middle of the histogram (Figure 6(d)). At this point, based on the CV values, the DO is more sensitive to the decay rate rather than stream flow and the roughness coefficient. In brief, from the results of the MCS analysis for the Zayandehrood River, it can be concluded that the decay rate is the most significant parameter for DO values.

Application of classification tree analysis

To evaluate the sensitivity of the model using classification tree analysis, the first step is to determine the distribution ranges of the input factors. Similar to MCS analysis, three input variables were selected: flow rate (Q), decay rate (k) and the roughness factor (n). In order to categorize the output parameters (DO) into non-numeric factors, the USEPA's water quality standards (www.epa.gov) was used.

According to this classification of water quality, if the DO level becomes less than 5 mg/l, the water can only be used for agricultural purposes. DO is categorized as high if it is above 5 mg/l and low if it is less than 5(mg/l). The same three points of the river have been chosen for the sensitivity analysis. The DO levels predicted by the mentioned simulations were categorized into high and low classes. The classification tree can be used for any number of DO categories. However, the result of classification tree analysis in this investigation is the sensitivity to DO binary classification. Figure 7 shows the classification tree of DO levels at 85 km from the downstream and Figure 8 is the partition plot for this point of the river which demonstrates the location of high and low concentrations of DO according to key input values. It is observed that with only two input values (decay rate and stream flow) the water quality can be classified. The first split with respect to the decay rate ($k = 0.198$) separates 40 lows out of 44 total lows and the second split (again with respect to the decay rate) separates 30 highs among total 34 lows ($k = 0.14$). The last split based on stream flow ($Q = 83.413$) only separates four highs among five highs in a leaf and four lows and a misclassified high in another leaf. As seen, the DO level is more sensitive to the decay rate at this point and the stream flow is the next important parameter. The distribution of lows and highs is quite homogeneous along the stream flow values but not along the decay rate values. Therefore, the decay rate

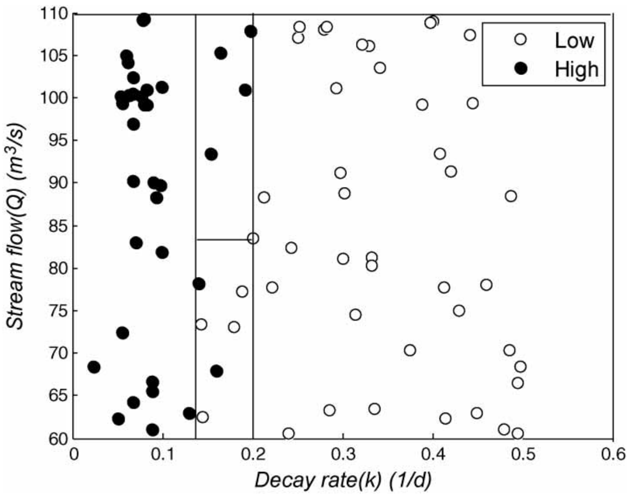

Figure 8 | Partition plot showing high and low concentrations of DO based on key input parameters, 85 km from the downstream.

values are discriminating highs and lows. Practically, in the first split 35 highs and 40 lows with only four misclassified (lows) have been classified. This implies the sensitivity of DO to the decay rate in the downstream of the river. Figure 8 shows where key input values yield a high or low concentration of dissolved oxygen based on their distribution range. The classifier lines in the partition plot are the classifiers generated by the classification tree.

Figure 9 shows the classification tree for the point 172 km from downstream. As seen, the classification tree was produced using only two input variables. The DO concentration is more sensitive to the decay rate and after that

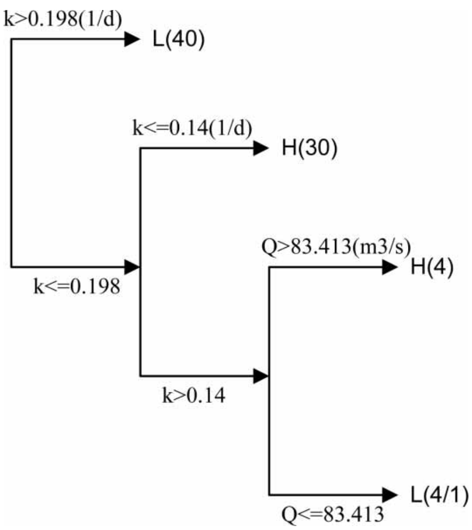

Figure 7 | The generated classification tree, 85 km from the downstream.

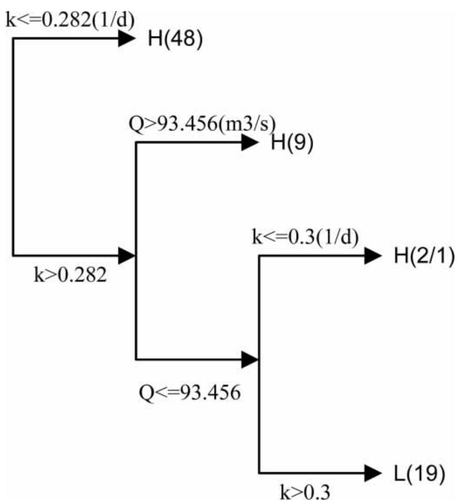

Figure 9 | The generated classification tree, 172 km from the downstream.

depends on the stream flow in this point. The first split, based on the decay rate ($k = 0.282$) separates 48 highs from total of 59 highs. The second split based on the stream flow ($Q = 93.456$) value separates nine highs among total of remaining 11. The last split based on the decay rate ($k = 0.3$), separates 19 lows of total 20 from only two highs and one misclassified low. In fact, after the second split and without the last one, 57 highs and 19 lows have been classified and only two highs have been misclassified. This shows that in this point, the decay rate is more important than the stream flow. Figure 10 is the partition plot for this point of the river which demonstrates the location of high and low concentrations of DO according to key input values.

The last point of the river that was used in the classification tree sensitivity analysis is 258 km far from the downstream. Figure 11 shows the classification tree generated for this point. As seen, the tree was built with two key input values (decay rate and roughness factor). Same as the previous points, the DO concentration is more sensitive to the decay rate. However, in this point, the stream flow is not important. According to the splits, the second key value in this point is the roughness factor. The first split separates 74 highs from total 76 highs based on the decay rate ($k = 0.48$). The second split is based on the roughness ($n = 0.105$) and separates all of three lows from two remaining highs. The partition plot for this point can be seen in Figure 12. The classifiers shown by the lines separate

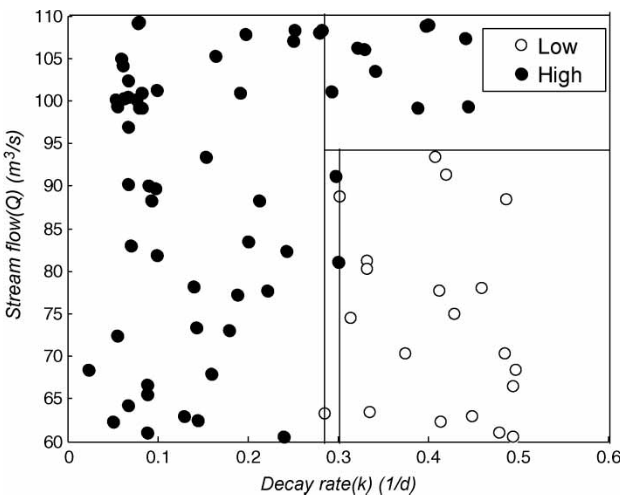

Figure 10 | Partition plot showing high and low concentrations of DO based on key input parameters, 172 km from the downstream.

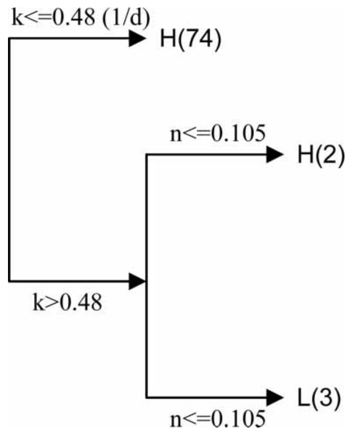

Figure 11 | The generated classification tree, 258 km from the downstream.

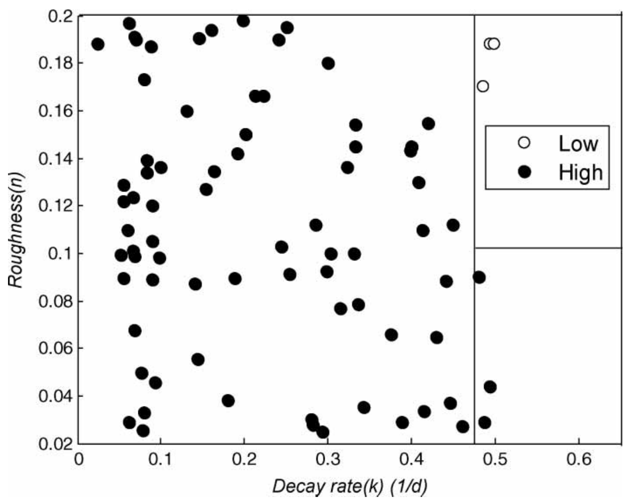

Figure 12 | Partition plot showing high and low concentrations of DO based on key input parameters, 258 km from the downstream.

high and low values of DO concentration. Using only the first split separates 74 highs and three lows with two misclassified highs.

An important point to note here is that in all three selected points, the DO concentration is more sensitive to decay rate than to the other factors and as the river runs thorough the downstream this sensitivity becomes more. The reason might be that more wastes are discharged in the downstream. This forces the DO values to decrease according to the decay rate. However, the sensitivity of the output to the stream flow is higher in the middle parts of the river and then it decreases through the upstream mainly because of large amounts of abstractions in the last

150 km of the river. The sensitivity of the output to the roughness factor was mainly seen at the upstream. The variations of the water depth of the water which affects the reaeration are partially due to the changes of roughness. As the roughness increases the depth of the water will increase and vice versa.

From MCS analysis and Classification Tree analysis it can be concluded that the DO was mainly sensitive to the decay rate in the downstream, upstream and the middle of the Zayandehrood River. In addition, based on the calculated CVs (Figure 5) and splitting order in the trees, it can be said that in the upstream and of the river the second important parameters were the roughness factor and stream flow, respectively. The decay rate and the stream flow were reported to have a dominant effect on prediction of uncertainties of DO and BOD (Abrishamchi *et al*. 2005). However, they have not considered the roughness coefficient in their analysis.

The most important advantage of the classification tree analysis is that in this method more than one input can be taken into account at the same time. In this way, the relationships between the input parameters for producing the output can be considered.

SUMMARY AND CONCLUSIONS

The sensitivity analysis helps modelers in decision making process. Application of MCS analysis and classification tree analysis for the Zayandehrood River, one of the biggest rivers in the Middle East, was investigated in this study. The DO concentration as an important indicator of water quality was considered as the main output parameter. In the MCS analysis the *pdf*s for three input parameters were considered to be triangular. After the decay rate the most important input parameters were found to be the roughness factor in the upstream (258 km from downstream) and the stream flow in the downstream (172 km and 85 km from downstream).

In the classification tree analysis all three parameters have been considered simultaneously. In this method, the results were divided into non-numeric parameters. The amount of DO was classified as high if it was more than 5 mg/l and low if it was less than 5 mg/l. In the upstream of the river the DO concentrations were mostly sensitive to the decay rate and then to the roughness factor. In the middle of the river, the concentration of DO was more sensitive to the decay rate and after that to the stream flow. The classification tree analysis demonstrates that in the downstream of the river, the DO was almost entirely dependent to the decay rate and it was less sensitive to the stream flow.

The main findings of this investigation are:

- The classification tree sensitivity analysis, as a global sensitivity analysis, is capable of determining the most effective input parameters in the model. Classification tree is more accurate in sensitivity analysis since it can consider the effects of all parameters simultaneously.
- The shape of *pdf*s could be explained by the variation of DO concentration against input variables. Results of sensitivity analysis based on the CV (coefficient of variation) showed that the decay rate coefficient is the most significant input variable.
- The comparison between the simulated and measured DO and BOD in the calibration and validation steps showed the capability of the QUAL2K.
- One of the advantages of the classification tree analysis is that it can show the different contribution of two or three input parameters to the outcomes by means of the partition plots that can be used easily for water quality classification.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Isfahan Environmental Protection Agency and the Isfahan Regional Water Board for providing the field data. We are grateful to Professor Srikanta Mishra from INTERA Inc. for his help. This study was partly supported by the Deputy of Research, Iran University of Science and Technology.

REFERENCES

Abrishamchi, A., Tajrishy, M. & Shafieian, P. 2005 Uncertainty analysis in qual2e model of Zayandehrood River. *Water Env. Res.* **77**, 279–286.

Anderson, M. P. & Woessner, W. W. 1992 *Applied Ground Water Modeling*. Academic, San Diego, California.

Beven, K. J. & Binley, A. 1992 The future of distributed models: model calibration and uncertainty prediction. *Hydrol. Process*. **6**, 279–298.

Bonnlander, B. V. & Weigend, A. S. 1994 Selecting input variables using mutual information and nonparametric density estimation. In *Proceedings of the International Symposium on Artificial Neural Networks (ISANN 94), Tainan, Taiwan*, pp. 42–50.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984 *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey, California.

Brown, L. C. & Barnwell, T. O. 1987 *The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS (EPA/600/3-87-007)*. U.S. Environmental Protection Agency, Athens, Georgia.

Cacuci, D. G. & Ionescu-Bujor, M. 2004 A comparative review of sensitivity and uncertainty analysis of large-scale systems – II: statistical methods. *Nucl. Sci. Eng*. **147** (3), 204–217.

Chapra, S. C. 1997 *Surface Water Quality Modeling*. McGraw-Hill, New York.

Chapra, S. C., Pelletier, G. J. & Tao, H. 2007 *QUAL2K: A Modeling Framework for Simulating River and Stream Water Quality, Version 2.07: Documentation and Users Manual*. Civil and Environmental Engineering Department, Tufts University, Medford, Massachusetts.

Fishman, G. S. 1995 *Monte Carlo Concepts, Algorithms, and Applications*. Springer, New York.

Francos, A., Elorzab, F. J., Bouraouia, F., Bidoglioa, G. & Galbiatia, L. 2003 Sensitivity analysis of distributed environmental simulation models: understanding the model behaviour in hydrological studies at the catchment scale. *Reliab. Eng. Syst. Safe*. **79** (2), 205–218.

Frey, H. C. & Patil, S. R. 2002 Identification and review of sensitivity analysis methods. *Risk Anal*. **22** (3), 553–578.

Freni, G. & Mannina, G. 2010 Bayesian approach for uncertainty quantification in water quality modelling: the influence of prior distribution. *J. Hydrol*. **392** (1–2), 31–39.

Georgakakos, K. P. & Krzysztofowicz, R. 2001 Probabilistic and ensemble forecasting. *J. Hydrol*. **249** (1–4), 1 (editorial).

Ghosh, N. C. 1996 Application of QUAL2E for water quality modeling of Kali River (UP). *Ind. J. Env. Health* **38** (3), 160–170.

Helton, J. C. 1993 Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. *Reliab. Engng Syst. Safety* **42** (2/3), 327–367.

Helton, J. C. & Anderson, D. R. (eds) 1999 Performance assessment for radioactive waste disposal. *Risk Anal*. **19** (5), 759–1016.

Ivanov, P., Masliev, I., Kularathna, M., De Marchi, C. & Somlyódy, L. 1996 *DESERT User Manual*. International Institute for Applied Systems Analysis, Laxenburg, Austria and Institute for Water and Environmental Problems, Barnaul, Russia.

Jacquin, A. P. & Shamseldin, A. Y. 2009 Sensitivity analysis of Takagi–Sugeno–Kang rainfall–runoff fuzzy models. *Hydrol. Earth Syst. Sci*. **13**, 41–55.

Kannel, P. R., Leea, S., Leeb, Y. S., Kanelc, S. R. & Pelletier, G. J. 2007 Application of automated QUAL2Kw for water quality modeling and management in the Bagmati River. Nepal. *J. Ecol. Model*. **202**, 503–517.

Krzysztofowicz, R. & Herr, H. D. 2001 Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation dependent model. *J. Hydrol*. **249** (1–4), 46–68.

Lindenschmidt, K.-E. 2006 The effect of complexity on parameter sensitivity and model uncertainty in river water quality modelling. *Ecol. Model*. **190** (1–2), 72–86.

Mannina, G. 2010 Uncertainty assessment of a water quality model for ephemeral rivers using GLUE analysis. *J. Environ. Engng – ASCE* **137** (3), 177–186.

Mannina, G. & Viviani, G. 2010 Water quality modelling for ephemeral rivers: model development and parameter assessment. *J. Hydrol*. **393** (3–4), 186–196.

Marsili-Libelli, S. & Giusti, E. 2008 Water quality modelling for small river basins. *Environ. Model. Softw*. **23** (4), 451–463.

McIntyre, N., Jacksona, B., Wade, A. J., Butterfield, D. & Wheater, H. S. 2005 Sensitivity analysis of a catchment-scale nitrogen model. *J. Hydrol*. **315**, 71–92.

McIntyre, N. R., Wagener, T., Wheater, H. & Chapra, S. C. 2003 Risk-based modelling of surface water quality: a case study of the Charles River, Massachussetts. *J. Hydrol*. **274**, 225–247.

McIntyre, N. & Wheater, H. S. 2004 A tool risk-based management of surface water quality. *Env. Model. Softw*. **19**, 1131–1140.

McQueen, R. J., Garner, S. R., Nevill-Manning, C. G. & Witten, I. H. 1995 Applying machine learning to agricultural data. *Comput. Electron. Agric*. **12** (4), 275.

Mertens, J., Madsen, H., Kristensen, M., Jacques, D. & Feyen, J. 2005 Sensitivity of soil parameters in saturated zone modelling and the relation between effective, laboratory and in situ estimates. *Hydrol. Process*. **19** (8), 1611–1633.

Mishra, S. 2009. Uncertainty and sensitivity analysis techniques for hydrologic modeling. *J. Hydroinf*. **11**, 282–296.

Mishra, S., Deeds, N. E. & RamaRao, B. S. 2003 Application of classification trees in the sensitivity analysis of probabilistic model results. *Rel. Engng Syst. Safety* **79** (2), 123–129.

Mishra, S., Deeds, N. E. & Ruskauff, G. 2009 Global sensitivity analysis techniques for probabilistic ground water modeling. *Ground Water* **47** (5), 727–744.

Paliwal, R., Sharma, P. & Kansal, A. 2007 Water quality modelling of the River Yamuna (India) using QUAL2E-UNCAS. *J. Env. Manag*. **83**, 131–144.

Palmieri, V. & Carvalho, R. J. 2006 Qual2e model for the Corumbataí River. *J. Ecol. Model*. **198**, 269–275.

Pappenberger, F., Beven, K. J., Ratto, M. & Matgen, P. 2008 Multimethod global sensitivity analysis of flood inundation models. *Adv. Water Resour*. **31** (1), 1–14.

Park, S. S. & Lee, Y. S. 2002 A water quality modeling study of the Nakdong River, Korea. *J. Ecol. Model*. **152**, 65–75.

Perez, H. H., Datta-Gupta, A. & Mishra, S. 2005 The role of electrofacies, lithofacies, and hydraulic flow units in permeability prediction from well logs: a comparative analysis using classification trees. *SPE Reser. Engng Eval*. **8** (2), 143–155.

Radwan, M., Willems, P., El-sadek, A. & Berlamont, J. 2003 Modelling of dissolved oxygen and biochemical oxygen demand in river water using a detailed and a simplified model. *Int. J. River Basin Manag.* **1** (2), 97–103.

Ratto, M., Young, P. C., Romanowicz, R., Pappenberger, F., Saltelli, A. & Pagano, A. 2007 Uncertainty, sensitivity analysis and the role of data based mechanistic modeling in hydrology. *Hydrol. Earth Syst. Sci.* **11**, 1249–1266.

Refsgaard, J. C., Van der Sluijs, J. P., Højberg, A. L. & Vanrolleghem, P. A. 2007 Uncertainty in the environmental modelling processa framework and guidance. *Environ. Model. Softw.* **22**, 1543–1556.

Rubinstein, Y. R. 1981 *Simulation and the Monte Carlo Method*. Wiley, New York.

Saltelli, A., Tarantola, S., Campolongo, F. & Ratto, M. 2004 *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Wiley, Chichester.

Smith, R. & Wheater, H. S. 2004 Multiple objective evaluation of a simple phosphorus transfer model. *Hydrol. Process.* **18** (9), 1703–1720.

Snowling, S. D. & Kramer, J. R. 2001 Evaluating modelling uncertainty for model selection. *Ecol. Model.* **138** (1), 17–30.

Sobol, I. M. 1993 Sensitivity analysis for non-linear mathematical models. *Math. Model. Comput. Expt.* **1** (4), 407–414.

Tang, Y., Reed, P., Wagener, T. & van Werkhoven, K. 2007 Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. *Hydrol. Earth Syst. Sci.* **11**, 793–817.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S. & Gupta, H. V. 2003 Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. *Hydrol. Process.* **17** (2), 455–476.

Wang, X., Frankenberger, J. R. & Kladivko, E. J. 2006 Uncertainties in DRAINMOD predictions of subsurface drain flow for and Indiana silt loam using the GLUE methodology. *Hydrol. Process.* **20** (14), 3069–3084.

First received 17 July 2010; accepted in revised form 8 November 2010. Available online 12 April 2011