WATER TEMPERATURE MODELLING BY MEANS OF GENETIC PROGRAMMING

M. Arganis[1*], R.Val[2], J. Prats[3], K. Rodríguez[4], R. Domínguez[1] and J. Dolz[3]

[1] *Instituto de Ingeniería, UNAM, Edificio 5, Cub. 414-F, Ciudad Universitaria, 04510 México, D.F., México. rdm@pumas.iingen.unam.mx*
[2] *Facultad de Ingeniería, UNAM, Ciudad Universitaria, 04510 México, D.F., México*
[3] *Universidad Politécnica de Catalunya, Barcelona, España., C/ Jordi Girona, 1-3, 08034 Barcelona, Spain. jordi.prats@upc.edu, j.dolz@upc.edu*
[4] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Ciudad Universitaria, 04510 México, D.F., México. katya@uxdea4.iimas.unam.mx*

**Corresponding author, e-mail MArganisJ@iingen.unam.mx*

ABSTRACT

In this work, an application of Genetic Programming (an evolutionary computational tool) is presented with the aim of modeling the behavior of the water temperature in a river. Recorded data corresponding to the water temperature behavior at the Ebro River, Spain, are used as analysis case, showing a performance improvement on the model developed when data are standardized. This improvement is reflected in a reduction of the mean square error.

KEYWORDS

Genetic Programming (PG), water temperature, Ebro River, measured data, fitting curves, standardization.

INTRODUCTION

Evolutionary computing has been widely used in hydraulics and hydrology; for example the studies of Savic *et al*. (1999), Madsen *et al*. (2000), and Dorado *et al*. (2002) related to rainfall-runoff processes; modeling of an urban aquifer was discussed by Hong and Rosen (2001); or the modifications of genetic programming algorithms attempting to get an agreement with the problem dimension in natural and compounded channels as applied by Keijzer and Babovic (2002), Harris *et al*. (2003), and Keijzer *et al*. (2005).

On the other hand, water temperature is an important parameter to take into account because of the changes it can experience due to human activities. In the last three decades diverse studies about weather changes have been made, related to the increase of extreme events such as floods and droughts (e.g. Lehner *et al*., 2006), the increasing air and water temperatures (e.g. Seguí, 2003; Webb & Nobilis, 1994), ice melting and greenhouse effect (e.g. Greve, 2000), with all their consequences in the surrounding ecosystems (e.g. Schindler, 1997; Álvarez Cobelas *et al*., 2005).

Water temperature in the lower Ebro River, Spain, was studied by Val (2003) and in the last two years an important effort has been made to obtain equations to predict water temperature

associated to meteorological variables, measured at the Ribarroja station located at this river (Arganis *et al*., 2005; Arganis *et al*., 2007). Downstream of Ribarroja, the Flix Dam is located. Part of the Ebro River's water is extracted about 5 km downstream from this reservoir for cooling purposes at the nuclear power plant of Ascó; the water is subsequently returned to the river and flows downstream towards Miravet (Figure 1).

Figure 1. Station locations at the Ebro River, Spain

In order to preserve the ecological balance it is very important to have a continuous inspection of water quality in that portion of the river. Freshwater organisms are mostly ectotherms and are therefore largely influenced by water temperature. Some of the expected consequences of a water temperature increase are life-cycle changes (Hellawell, 1986; Winfield & Nelson, 1991), and shifts in the distribution of species with the arrival of allochthonous species (Schindler, 1997; Walther *et al*., 2002) and the expansion of epidemic diseases (Harvel *et al*., 2002) as a possible result. Also, aquatic flora and fauna depend on dissolved oxygen to survive and this water quality parameter is a function of water temperature as well.

The reason to count with models that allow the representation of water temperature behavior in terms of time is because each time that a possible abnormal increase in this parameter happens, the consequences and implications for the physical and chemical properties of water with their corresponding effects in aquatic life are numerous; some models have been applied to maximum water temperatures by means of non linear relationships between air temperature and water temperature (Caissie *et al*., 2001) but there are other important variables involved in water temperature variation during a given period of time.

METHODOLOGY

Genetic programming algorithm

A typical genetic programming algorithm consists of a set of functions, which can involve arithmetic operators (+, -, *, /,…), transcendental functions (***sin, cos, tan…., ln, exp,…)***, even relational operators (>, <, =) or conditional operators (IF); and a terminal set with variables and constants (x_1, x_2 ,x_3,…x_n). An initial population is randomly created with a number of individuals formed by nodes (operators plus variables, and constants) previously defined according to the problem domain. An objective function must be defined to evaluate the fitness of each individual (in this case each individual will be a resultant model or program of the random combination of nodes). Selection, crossover and mutation operators are then applied to the best individuals and a new population is created. The whole process is repeated until the given generation number is reached (Cramer, 1985; Koza, 1989).

In this document, for simplicity, only four arithmetic operators were considered:

$$FS=\{+,-,*,/\} \quad (1)$$

Twelve independent variables, one dependent variable and a vector of real constants were selected. Thus, in the non-standardized case the terminal set is:

$$TS=\{h_{r98}, T_{a98}, v_{v98}, r_{s98}, h_{r99}, T_{a99}, v_{v99}, r_{s99,} h_{r2000}, T_{a2000}, v_{v2000}, r_{s2000,} T_{w2000,} \bar{b}\} \quad (2)$$

where: h_{r98}, h_{r99}, h_{r2000} are the hourly average relative humidity values recorded in the years 1998 to 2000, in decimals; T_{a98}, $T_{a99,}$ T_{a2000} are average air temperature values from years 1998, 1999 and 2000, in ^{o}C; v_{v98}, $v_{v99,}$ v_{v2000} are the average wind speeds from years 1998, 1999 and 2000, in m/s;r_{s98} , r_{s99}, r_{s2000} are average solar radiations from years 1998, 1999, 2000, in W/m^2; T_{w2000} is the hourly average water temperature measured from year 2000, in ^{o}C; $\bar{b}$ is a real constant vector.

Tests were made with daily and weekly averaged water temperatures. In the standardized case all the last variables are dimensionless.

Objective Function

The objective function considered in this problem was defined as the minimization of the mean square error between calculated and measured data:

$$FO = Min\left[\sum_{i=1}^{n}\frac{(T_{w_i} - T_{w1_i})^2}{n}\right] \quad (3)$$

where: T_w measured data, T_{w1} calculated data, i counter from *1* to data number n; the genetic programming algorithm was implemented in MATLAB (The MathWorks, 1992).

Standardization

The variables were standardized by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{T_w - \overline{T}_w}{\sigma_{T_w}} \quad (4)$$

where: Z standardized variable, dimensionless, T_w variable before standardization, with physical dimensions, $\overline{T}_w$ mean of T_w, with the same units than T_w (the arithmetic average was used), and σ_{T_w} standard deviation of T_w, with the same units than T_w

Input data

Meteorological and water temperature data were taken in gauging stations installed in the Ebro River. Data consists of 10-minute averages of the measurements taken every minute. Water temperatures were measured just downstream of the hydroelectric power plant of Flix. The meteorological variables were measured at the measuring station located on the Ribarroja Dam. The daily and weekly average were calculated for all the variables and taken as input data: relative humidity (h_r), air temperature (T_a), wind speed (v_v), and solar radiation (r_s) as independent variables and water temperature (T_w) as the dependent variable.

The first trial was done with the original data, and the second one with the standardized ones. In each case a text file of thirteen columns was built containing the independent variables from columns one to twelve and the dependent variable in column thirteen.

GP parameter settings for both experiments are shown in Table 1.

Table 1. GP Parameter Settings

Parameter	Value
Number of individuals	250
Maximum number of nodes	30
Maximum number of generations	3,000
Cross probability	0.9
Mutation probability	0.09
Real constant mutation probability	0.03

Model Linearity

In order to validate the applicability of the method, the correlation coefficient between measured and calculated data was obtained:

$$r_{T_w T_{w1}} = \frac{Cov(T_w, T_{w1})}{\sigma_{T_w} \sigma_{T_{w1}}} \tag{5}$$

where: $Cov(T_w, T_{w1}) = \frac{1}{n}\sum_{i=1}^{n}(T_{w_i} - \overline{T}_w)(T_{w1_i} - \overline{T}_{w1})$, is the covariance between T_w and T_{w1} variables; $\sigma_{T_w}, \sigma_{T_{w1}}$ are the standard deviation of T_w and T_{w1}, respectively

RESULTS AND DISCUSSION

Daily average data

The processing time took about 25 minutes (1500 s). The equations obtained without and with standardization were as follows:

$$T_{w12000} = T_{a98} + \frac{T_{a2000}}{T_{a98}} + \frac{r_{s98}}{\dfrac{(T_{a2000} - 2r_{s98})h_{r98}T_{a2000}}{v_{v2000}{}^2 T_{a98}}} \tag{6}$$

$$T_{w1_{2000_z}} = 0.5018T_{a98} + 0.4982T_{a99} - 0.2108r_{s98} - 0.1195v_{v99} + 0.1195v_{v2000} - 0.1195h_{r2000} \tag{7}$$

where: $T_{w1_{2000}}$ is the daily average water temperature value estimated in 2000, in ^{o}C, h_{r98}, h_{r99}, are the daily average relative humidity values recorded in 1998 and 1999, in decimals, T_{a98}, T_{a99} are the daily average air temperatures of 1998 and 1999, in ^{o}C, r_{s98} is the daily average solar radiation of 1998, in ^{o}C; the *z* prefix indicates a standardized variable.

By applying an inverse standardization process:

$$T_{w1_{2000}} = \sigma_{T_{w2000}} T_{w1_{2000_z}} + \mu_{T_{w2000}} \tag{8}$$

In equation 8, data from 2000 are estimated according to equations 10 and 11, but considering daily measurements. The mean square errors (MSE) obtained by using equations 6 and 8 are set on Table 2. The mean (μ_r) and the standard deviation of residuals (σ_r) of this experiment appear in Table 3. Water temperature variations against time and the obtained differences are plotted on Figures 2 and 3. Figures 4 and 5 show a comparison between measured and

calculated daily average water temperatures with respect to the identity function. (1 day=86,400 s).

Table 2. Mean square error values. Daily average data

Equation	MSE, ^{o}C
6	8.279
8	4.978

Table 3. Statistics of residuals. Daily average data

Equation	μ_r (^{o}C)	σ_r (^{o}C)
6	0.0762	2.8802
8	0.0213	2.2342

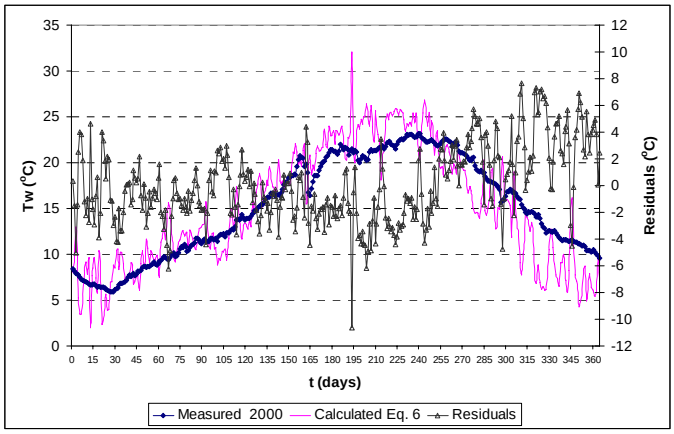

Figure 2. Water temperature values and residuals. Trial without standardization. One-day average values

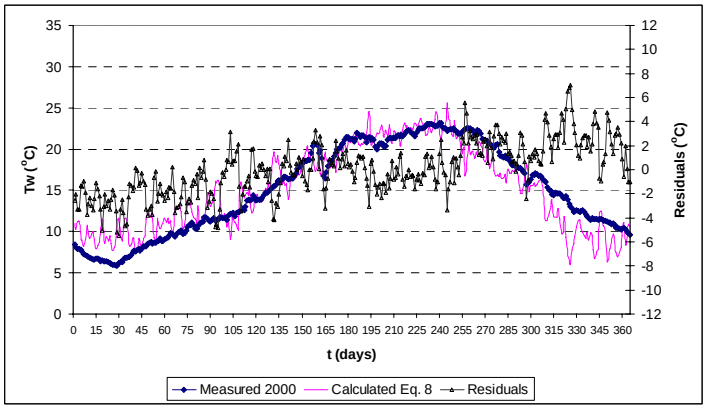

Figure 3. Water temperature values and residuals. Trial with standardization. One-day average values

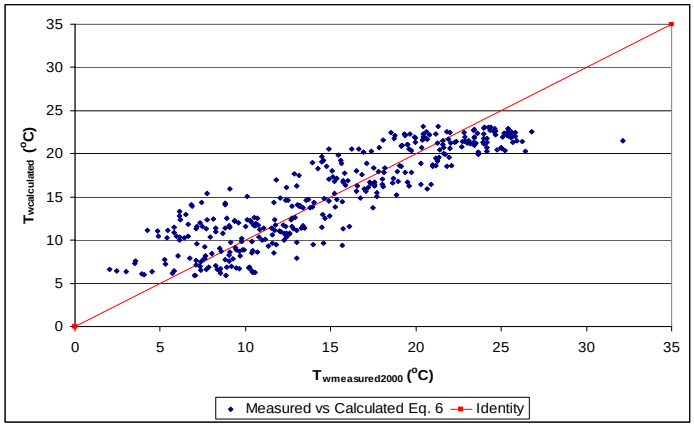

Figure 4. Comparison between measured and estimated data (Eq. 6). Correlation coefficient $r_{Tw,Tw1}$=0.8939

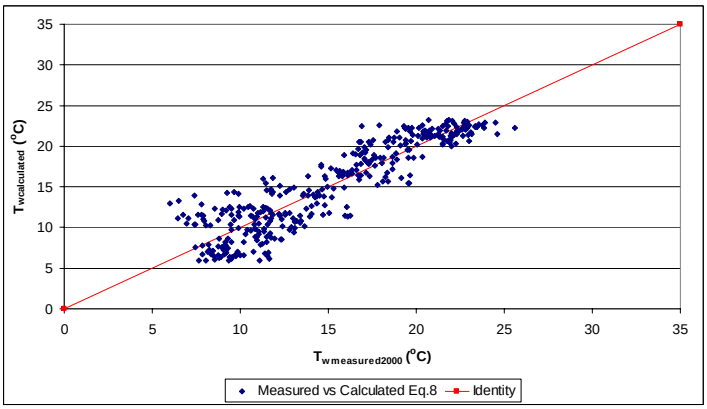

Figure 5. Comparison between measured and estimated data (Eq. 8). Correlation coefficient $r_{Tw,Tw1}$=0.9091

Results for daily analyses report a reduction of nearly 40% in mean square error with the equation obtained using standardized data. In this case the standard deviation of residuals is also smaller (12% lower than using non-standardization).

Weekly average data

In this last experiment, processing time was similar to daily results (about 1380 s). The equations obtained without and with standardization were:

$$T_{w1_{2000}} = 2h_{r98} - v_{v99} - 1.4558v_{v2000} + T_{a98} + \frac{(\frac{T_{a2000}}{T_{a99} - v_{v98} - v_{v99} - v_{v2000} + h_{r2000}}) + v_{v98}}{v_{v2000}} \quad (9)$$

$$T_{w12000z} = 0.2962T_{a98} + 0.6819T_{a99} + 0.6397T_{a2000} - 0.2668r_{s98} - 0.3215r_{s99} - 0.3852r_{s98}T_{a2000} + 0.3852r_{s98}r_{s99} - 0.0928 \quad (10)$$

where: $T_{w1_{2000}}$ is the weekly average water temperature value estimated in 2000, in ^{o}C, $h_{r98,}$ h_{r2000} are the weekly average relative humidity values recorded in 1998 and 2000, in decimals, T_{a98}, $T_{a99,}$ T_{a2000} are the weekly average air temperatures of 1998,1999 and 2000, in ^{o}C, v_{v98}, v_{v99}, v_{v2000} are the weekly average wind speeds from years 1998, 1999 and 2000, in m/s, $r_{s98,}$ $r_{s99,}$ are the weekly average solar radiation values of 1998 and 1999, in ^{o}C, and the *z* prefix indicates standardized variable. Equation 10 must be non-standardized to get the average weekly temperature approach:

$$T_{w1_{2000}} = \sigma_{T_{w2000}} T_{w1_{2000_z}} + \mu_{T_{w2000}} T_{w_{2000}} \quad (11)$$

Mean square errors and statistics of residuals appear in Tables 4 and 5. Figures 6 to 9 show the behavior of water temperature in this weekly analysis. (1 week=604,800 s)

Table 4. Mean square error values. Weekly average data

Equation	MSE, ^oC
9	4.538
11	2.176

Table 5. Statistics of the residuals. Weekly average data

Equation	μ_r (^oC)	σ_r (^oC)
9	0.0186	2.1509
11	0.0239	1.4892

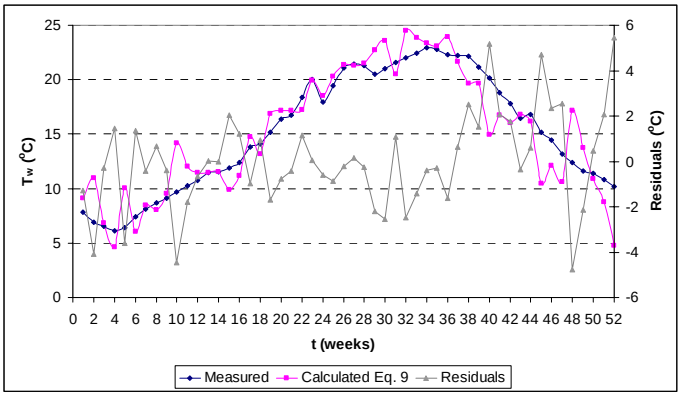

Figure 6. Water temperature values and residuals. Trial without standardization. Weekly average values

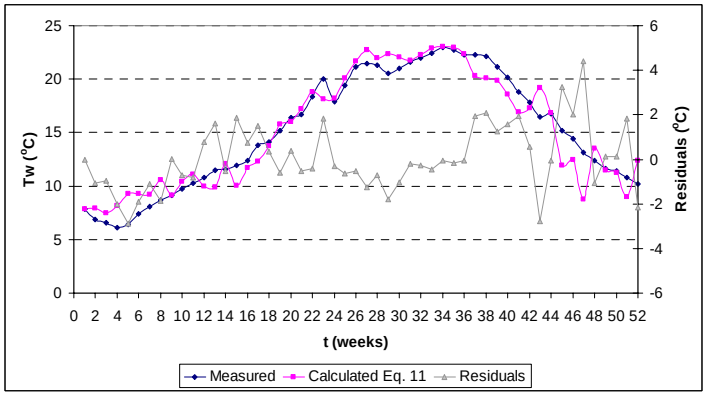

Figure 7. Water temperature values and residuals. Trial with standardization. Weekly average values

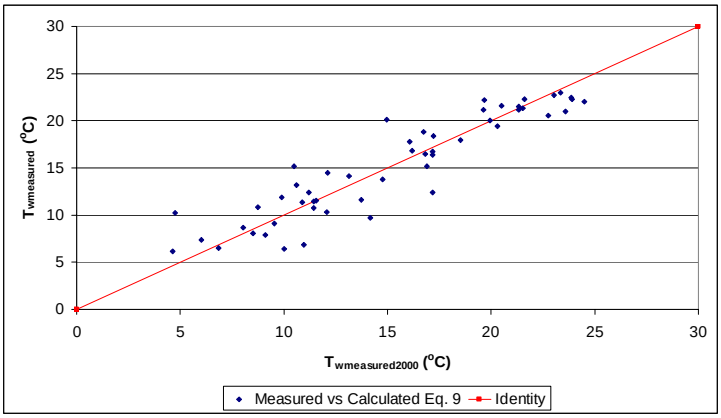

Figure 8. Comparison between measured and estimated data (Eq. 9). Correlation coefficient $r_{Tw,Tw1}$=0.9241

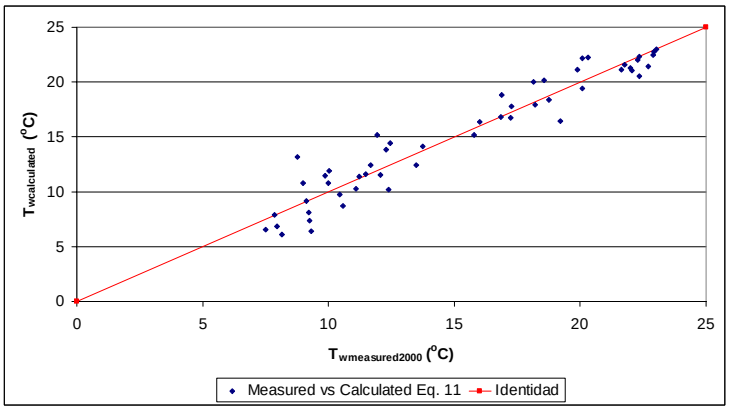

Figure 9. Comparison between measured and estimated data (Eq. 11). Correlation coefficient $r_{Tw,Tw1}$=0.9612

The results obtained for the weekly analysis show a reduction of 52% in the mean square error when data are previously standardized, and about 31% reduction in the standard deviation of residuals. The correlation coefficient is also close to one.

CONCLUSIONS

Different models which allow the estimation of water temperature in the Ebro River in a given year were obtained, taking into account climatic variables measured in the same year, but also considering their variability in two previous years. The GP algorithm was fed with hourly, daily and weekly average measured data without and with standardization, in order to analyze the resulting equations when the shape of input data varies from one form to another.

Intrinsically, measured data of water temperature have more oscillations in hourly average data than in daily or weekly average data. Particularly in the experiment using hourly data, the GP algorithm showed some difficulties to reproduce such oscillations with the considered arithmetic operators in both cases (without and with standardization). Nevertheless, by using standardized data, mean square errors were lower than those without standardization and a lower dispersion in data could be obtained. Similar situations occurred in the case of daily data.

When weekly data were considered, GP algorithms gave models able to follow the behavior of water temperature, particularly those obtained with standardized data based on the mean square errors, the standard deviation of residuals and the correlation coefficient.

According to these results, the convenience of the standardization process is evident to be able to get some improvements in generating water temperature models by means of genetic programming.

ACKNOWLEDGEMENT

Thanks to the Instituto de Ingeniería, UNAM and to FLUMEN, UPC, Barcelona, Spain for their support.

REFERENCES

Álvarez Cobelas, M., J. Catalán, & D. García De Jalón.(2005). Impactos sobre los ecosistemas acuáticos continentales. In: *Evaluación preliminar de los impactos en España por efecto del cambio climático*. J. M. Moreno (Coord.): pp. 113-146. Ministerio de Medio Ambiente, Madrid, Spain.

Arganis, M. L., S. R. Val, V. K Rodríguez, M. R. Domínguez, & R. J. Dolz (2005). Comparación de curvas de ajuste a la Temperatura del Agua de un río usando programación genética. *Congreso Mexicano de Computación Evolutiva COMEV*, mayo 2005, Universidad Nacional Autónoma de Aguascalientes. 8 pp.

Arganis, M. L., S. R. Val, M. R Domínguez, V. K. Rodríguez, R. J. Dolz & J. Eaton (2007). Comparison between equations obtained by means of multiple linear regression and genetic programming to approach measured climatic data in a river. *IWA Watermex*, 2007 May 7-9, Washington, D. C. 8 pp.

Caissie, D., N. El-Jebi & M.G Satish. (2001). Modelling maximum daily water temperature in a small stream using air temperatures. *Journal of Hydrology*, 251: 14-28.

Cramer, N. L. (1985). A representation for the adaptive generation of simple sequential programs. In: *Proceedings of International Conference on Genetic Algorithms and the Applications*, Grefenstette, J.J. Editor: 183-187.

Dorado, J., J. R. Rabuñal, J. Puertas, A. Santos, & D. Rivero (2002). Prediction and Modelling of the Flow of a Typical Urban Basin through Genetic Programming. *Lecture Notes in Computer Science*. EvoWorkshops 2002, LNCS 2279: 190–201.

Greve, R. (2000). On the response of the Greenland ice sheet to greenhouse climate change. *Climatic Change*, 46: 289-303.

Harris, E. L., V. Babovic, & R.A. Falconer (2003). Velocity predictions in compound channels with vegetated floodplains using genetic programming. *Intl. J. River Basin Management*, Vol. 1, No. 2: 117–123.

Harvell, C. D., C. E Mitchell, J. R Ward, S. Altizer, A. P Dobson, R. S. Ostfeld & M. D. Samuel (2002). Climate warming and disease risks for terrestrial and marine biota. *Science*, 296: 2158-2162.

Hellawell, J. M. (1986). *Biological indicators of freshwater pollution and environment management*. Elsevier, London. 546 pp.

Hong, Y. S & M. R. Rosen (2002). *Journal of Hydrology*, Volume 259 Issues 1-4, 1 March, pp 89-104.

Kaijzer, M. & V. Babovic (2002). Declarative and Preferential Bias in GP-based Scientific Discovery. *Genetic Programming and Evolvable Machines*, 3: 41–79.

Kaijzer, M., M Baptist, V. Babovic, & U.J. Rodriguez (2005). Determining equations for vegetation induced resistance using genetic programming. *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation. GECCO'05*, June 25–29, Washington, DC, USA. 1999-2006.

Koza, J. R. (1989). Hierarchical genetic algorithms operating on populations of computer programs. In: *Proceeding of the 11th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1: 768-774.

Lehner, B., P. Döll, J. Alcamo, T. Henrichs, & F. Kaspar (2006). Estimating the impact of global change on flood and drought risks in Europe: A continental integrated analysis. *Climatic Change*, 75: 273-299.

Madsen, H., M. B. Butts, S. T. Khu, & S. Y. Lyong (2000). Data assimilation in rainfall-runoff forecasting. *Hydroinformatics 2000, 4th Conference of Hydroinformatics*, Cedar Rapids, Iowa, USA, 23-27 July, pp 1-8

Savic, D., G. Walters, & J. W. Davidson (1999). A Genetic Programming Approach to Rainfall-Runoff Modelling. *Water Resources Management*, 13: 219–231.

Schindler, D. W (1997). Widespread effects of climatic warming on freshwater ecosystems in North America. *Hydrological Processes*, 11: 1043-1067.

Seguí, J (2003). *Análisis de la serie de temperatura del Observatorio del Ebro 1894-2002*. Observatori de l'Ebre, Roquetes, Spain. 83 pp.

The MathWorks (1992). *MATLAB Reference Guide*. The MathWorks, Inc.

Val, S. R. (2003). *Incidencia de los embalses en el comportamiento térmico del río. Caso del sistema de embalses Mequinenza-Ribarroja-Flix en el río Ebro*. PhD Thesis. Universitat Politècnica de Catalunya, Barcelona, Spain. 196 pp.

Walther, G-R., E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. C. Beebee, J-M. Fromentin, O. Hoegh-Guldberg, & F. Bairlein (2002). Ecological responses to recent climate change. *Nature*, 416: 389-395.

Webb, B. & W. F. Nobilis (1994). Water temperature behaviour in the River Danube during the twentieth century. *Hydrobiologia*, 291: 105-113.

Winfield, I. J. & J. S. Nelson (1991). *Cyprinid fishes. Systematics, biology and exploitation*. Chapman & Hall, London. 667 pp.