Hindawi Publishing Corporation
Advances in Civil Engineering
Volume 2009, Article ID 353960, 10 pages
doi:10.1155/2009/353960

Research Article

Genetic Programming and Standardization in Water Temperature Modelling

Maritza Arganis,[1] Rafael Val,[2] Jordi Prats,[3] Katya Rodríguez,[4] Ramón Domínguez,[1] and Josep Dolz[3]

[1] *Instituto de Ingeniería, UNAM, Ciudad Universitaria, Edificio 5, Cub. 403, 04510 México, DF, Mexico*
[2] *Facultad de Ingeniería, UNAM, Ciudad Universitaria, 04510 México, DF, Mexico*
[3] *Department of Hydraulic Engineering, Maritime and Environmental, Universidad Politécnica de Catalunya, C/ Jordi Girona1-3, 08034 Barcelona, Spain*
[4] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Ciudad Universitaria, 04510 México, DF, Mexico*

Correspondence should be addressed to Maritza Arganis, marganisj@iingen.unam.mx

Received 21 May 2008; Revised 20 April 2009; Accepted 21 June 2009

Recommended by Bryan W. Karney

An application of Genetic Programming (an evolutionary computational tool) without and with standardization data is presented with the aim of modeling the behavior of the water temperature in a river in terms of meteorological variables that are easily measured, to explore their explanatory power and to emphasize the utility of the standardization of variables in order to reduce the effect of those with large variance. Recorded data corresponding to the water temperature behavior at the Ebro River, Spain, are used as analysis case, showing a performance improvement on the developed model when data are standardized. This improvement is reflected in a reduction of the mean square error. Finally, the models obtained in this document were applied to estimate the water temperature in 2004, in order to provide evidence about their applicability to forecasting purposes.

1. Introduction

Evolutionary computing has been widely used in hydraulics and hydrology, for example, the studies ofSavic et al. [1], Madsen et al. [2], and Dorado et al. [3], related to rainfall-runoff processes, modeling of an urban aquifer as was discussed by Hong and Rosen [4], 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 [5], Harris et al. [6], and Keijzer et al. [7]. On the other hand, water temperature is an important parameter to consider 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. [8]) the increasing air and water temperatures (e.g., Seguí [9]; Webb and Nobilis [10]), ice melting, and greenhouse effect (e.g., Greve [11]), with all their consequences in the surrounding ecosystems (e.g., Schindler[12], Álvarez Cobelas et al. [13]).

The motivation to work with models that allow the representation of water temperature behavior year after year is because each time a possible abnormal increase in this parameter occurs, 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 nonlinear relationships between air temperature and water temperature (Caissie et al. [14]), but there are other important variables involved in water temperature variation during a given period of time. 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 [15]; Winfield and Nelson [16]), shifts in the distribution of species with the arrival of allochthonous species (Walther et al. [17]), and the expansion of epidemic diseases (Harvell et al. [18]) as

Figure 1: Station locations at the Ebro River, Spain.

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.

2. Study Site

The field data used in this study were taken from the lower Ebro River, Spain. This river has a basin of 85 000 km^2 and an average year inflow of 17 000 hm^3 in natural regime. Three dams are located along the river (Figure 1) which change the water temperature regime (Val [19]): Mequinenza (1534 hm^3), Ribarroja (210 hm^3), and Flix (11 hm^3). Five kilometers downstream the Flix dam, the water is taken from the river with cooling purposes in the Nuclear Central Ascó. Water is returned to the river with a higher temperature and flows downstream to Miravet. In this zone several meteorological gauge stations were installed, including some measuring water temperature (Figure 1). These data were applied in studies made by Val [19] and Prats et al. [20]. Besides, an important effort has been recently made to obtain equations to predict water temperature associated to meteorological variables that are easily measured, centered at the Ribarroja station (Arganis et al. [21, 22]).

3. Methodology

3.1. Evolutionary Algorithms. Evolutionary algorithms, also known as Evolutionary Computation (EC), the optimization tool used in this work, use computational models of evolutionary processes in the design and implementation of computer-based problem solving. A general definition and classification of these evolutionary techniques is given in Bäck [23]. He defines an EA as a search and optimization algorithm, inspired by the process of natural evolution, which maintains a population of structures that evolve according to the rules of selection and other operators such as recombination and mutation. Here, the structure of all evolution-based algorithms is shown in Algorithm 1.

In a similar way to that of natural evolution and heredity, these algorithms work on a population of N individuals

```
Program
    t = 0
    Create Initial Population P(t)
    Evaluate Initial Population P(t)
    While (not termination_criterion) do
        t = t + 1
        Select Individuals for Reproduction P(t) from P(t − 1)
        Alter P(t)
        Evaluate New Population P(t)
    end
```

Algorithm 1: Evolution-based algorithm.

$P(t) = \{x_1^t, \ldots, x_N^t\}$, representing search points in the space of potential solutions of a given problem. How well each individual x_i^t adapts each generation t to the problem under investigation is provided by a quality measure called the "*fitness*". The population evolves, generation by generation, towards better regions of the search space by means of genetic processes, such as *selection*, *recombination*, and *mutation*. The selection process uses the fitness measure to choose individuals of the previous generation ($P(t-1)$) to be reproduced, favoring those of higher quality. The recombination operation promotes the exchange of genetic information between parent individuals, thereby producing descendants. The mutation operation alters the genetic information by introducing some changes into the population. The evaluation process is repeated until a predefined termination criterion is met, or alternatively, until a maximum number of generation (iterations) is reached. This artificial evolution process is the foundation of the evolution-based algorithms used in this work, genetic programming.

3.2. Genetic Programming Algorithm. A typical genetic programming algorithm consists of a set of functions, which can involve arithmetic operators ($+, -, *, /, \ldots$), transcendental functions (sin, cos, tan,..., ln, exp,...), even relational operators ($>, <, =$) or conditional operators (IF), and a terminal

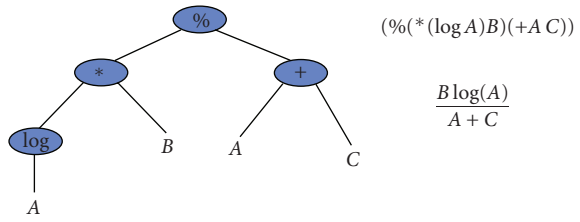

Figure 2: A mathematical expression represented hierarchically by its parse tree.

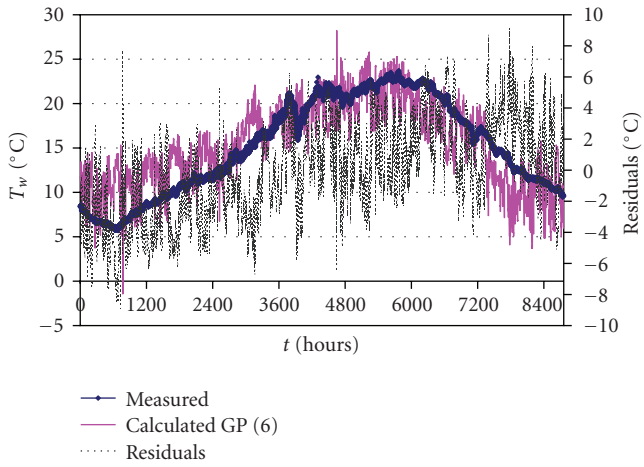

Figure 3: Water temperature values and residuals, experiment without standardization (hourly average values).

set with variables and constants $(x_1, x_2, x_3, \ldots x_n)$. An initial population is randomly created with a number of parse-tree individuals composed of nodes (operators plus variables, and constants) previously defined according to the problem domain (an example of GP individual is given in Figure 2). 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 [24], Koza [25]).

3.3. Brief Description of the Physical Phenomena and Their Related Variables. The water from a river is in a constant heat exchange with its surroundings: the atmosphere and the river bed. This process may reach equilibrium so that the heat lost by the water equals that which is absorbed. Normally, the water temperature increases throughout the river in a natural state as the altitude decreases. To this spatial variation a double temporal variation is superimposed. In a river reach temperature varies following both a daily and an annual cycle.

In the study performed by Val [19], an analysis of five kilometers of the Ebro River was performed, in a section downstream of the Flix hydroelectric center; in this reach, hourly temperature measurement data are available for different sections. It was observed that during the summer a 9°C difference may exist between the Flix Central site and the temperature before the dams. Additionally, downstream of Flix Central, the water temperature recovers, trying to reach thermal equilibrium with its surroundings. To estimate the heat that is absorbed by the river water as it progresses naturally through a certain reach, and its corresponding temperature variation, an energy balance is established between the caloric energy received and the caloric energy emitted by the water along that reach. This can be done based on the thermic balance presented by Edinger et al. [26]. This balance can be expressed as

$$A = H_{sn} + H_{an} - H_b - H_e - H_c + S, \quad (1)$$

where A is the total caloric power absorbed by water as it moves along a river reach, by square meter of free surface, measured in W/m^2. This is the result of the balance of the different heat inputs and outputs for water as it moves along the reach. H_{sn} is the net (incident minus reflected) total shortwave solar radiation (direct plus diffused) that is absorbed by the water by square meter of free surface, measured in W/m^2. This is a function of the incident solar radiation r_s and the reflected r_r, which is proportional to r_s, and this proportionality is given by the constant α which is also known as the albedo. H_{an} is the net longwave atmospheric radiation (incident minus reflected) absorbed by water by free surface square meter, measured in W/m^2. H_b is the longwave radiation emitted by water by free surface square meter, measured in W/m^2, determined as a function of average surface water temperature, T_w. H_e is the heat lost through evaporation by free surface square meter, measured in W/m^2, determined as a function of the wind velocity, v_v, the vapor pressure of saturation, and the air's vapor pressure. Relative humidity h_r is also a variable that affects the water-atmosphere heat exchange. H_c is the sensitive heat interchanged by conduction between the atmosphere and water by free surface square meter, measured in W/m^2, dependent upon the air temperature T_a and that of the water T_w. S is the heat exchanged with the substrate (river bed) by square meter of river reach.

The heat stored by a water mass as it moves along a river stretch of longitude L is estimated by $A = 4187(\Delta T Q C_e \rho / LB)$, where A is the caloric power absorbed by water, in (W/m^2), ΔT is the water temperature increment, in (°C), Q is the circulating flow, in (m^3/s), C_e is the specific heat of water in (Kcal/°CKg), ρ is the density of water, in (Kg/m^3), L is the longitude of the studied reach, in (m), and B_w is the effective width of the river, in (m).

On the other hand, through an analysis of the historical behavior of the time variation of water temperature during consecutive years, similar results were observed, both in the cyclical variation and in the tendency to increase or decrease. This leads to an expectation of a correlation between the temperature variation in year i and the temperature of previous years.

This background described led to the choice of the measured variables which were used in the prediction model.

Additionally, when physical variables are used to be fitted by means of genetic programming, several questions about

Table 1: GP parameter settings.

Parameter	Value
Number of individuals	250
Maximum number of nodes	30
Maximum number of generations	3000
Cross probability	0.9
Mutation probability	0.09
Node mutation probability	0.03

Table 2: Mean square error values.

Equation	MSE, °C
(6)	9.4336
(8)	6.4763

Table 3: Statistics of residuals.

Equation	μ_r (°C)	σ_r (°C)
(6)	−0.0004	3.0716
(8)	0.0230	2.5449

the dimensionality of the problem could be made. But this problem can be solved considering the possible existence of dimension in the obtained constants of the calculated model. New physical interpretations of the related variables can be done by analyzing the model terms.

In this document, for simplicity, only four arithmetic operators were considered: $FS = \{+, -, *, /\}$.

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

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

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

Tests were made with one hour, daily and weekly averaged water temperatures.

In the standardized case all the last variables are dimensionless.

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

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

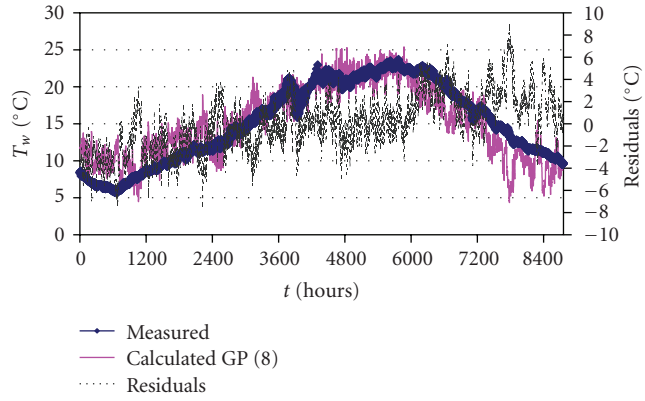

Figure 4: Water temperature values and residuals, experiment with standardization (hourly average values).

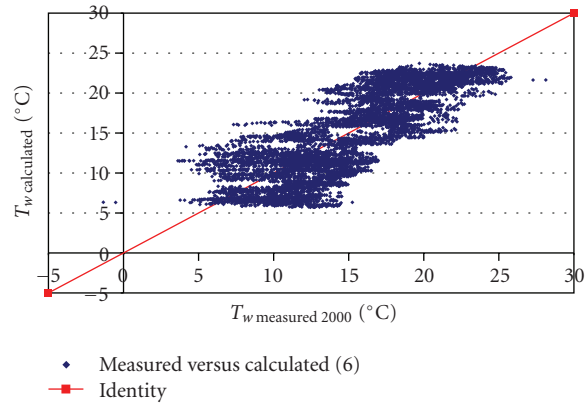

Figure 5: Comparison between measured and estimated data (6), correlation coefficient $r_{Tw,Tw1} = 0.82$.

where T_w measured data, T_{w1_i} calculated data, and i counter from 1 to data number n.

The genetic programming algorithm was implemented in MATLAB (The MathWorks [27]).

3.5. 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 is the standardized variable, dimensionless; T_w is the variable before standardization; $\overline{T}_w$ is the mean value of T_w, with the same units as T_w (the arithmetic average was used); σ_{T_w} is the standard deviation of T_w, with the same units as T_w.

Variables with large variances tend to have a larger effect on the resulting model than those with small variances that can be also relevant. Standardized variables can then be advantageous in that their means are zero and their second moments (variances) are one.

3.6. Input Data. Meteorological and water temperature data were taken in gauging stations installed in the Ebro

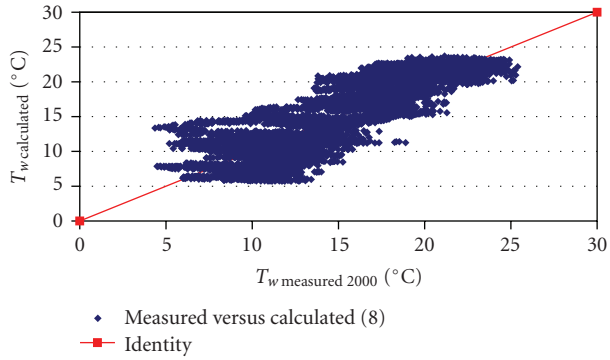

Figure 6: Comparison between measured and estimated data (8), correlation coefficient $r_{Tw,Tw1} = 0.88$.

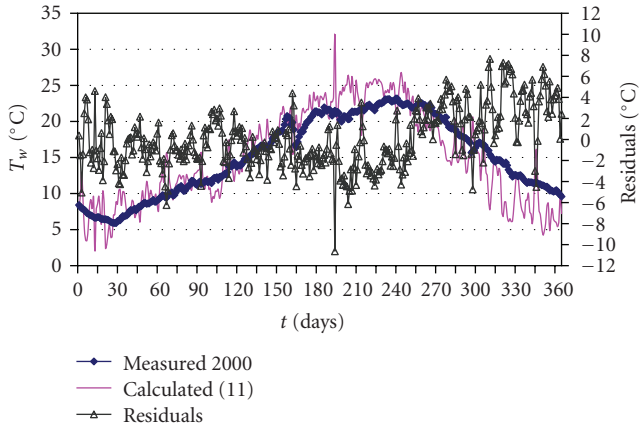

Figure 7: Water temperature values and residuals, experiment without standardization (daily average values).

River. Data consist of 10-minute averages of 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 hourly average was 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 experiment was carried out with the original data, and the second one with the standardized ones. GP parameter settings for both experiments are shown in Table 1

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

$$r_{T_w T_{w1}} = \frac{\text{Cov}(T_w, T_{w1})}{\sigma_{T_w}\sigma_{T_{w1}}},$$
$$\text{Cov}(T_w, T_{w1}) = \frac{1}{n}\sum_{i=1}^{n}\left(T_{w_i} - \overline{T}_w\right)\left(T_{w1_i} - \overline{T}_{w1}\right), \quad (5)$$

Table 4: Mean square error values. Daily average data.

Equation	MSE, °C
(11)	8.279
(13)	4.978

Table 5: Statistics of residuals. Daily average data.

Equation	μ_r (°C)	σ_r (°C)
(11)	0.0762	2.8802
(13)	0.0213	2.2342

where $\text{Cov}(T_w, T_{w1})$ is the covariance between the variables T_w and T_{w1}; $\sigma_{T_w}, \sigma_{T_{w1}}$ are the standard deviation of T_w and T_{w1}, respectively.

4. Results and Discussion

4.1. One-Hour Average Data. The genetic programming algorithm tendency is to produce relatively simple models. The equations produced in both experiments were

$$T_{w1_{2000}} = T_{a99}h_{r99} + \frac{2}{h_{r99}} + \frac{v_{v98} + v_{v2000}}{T_{a99} + h_{r2000} + 0.6776} + 1.3445, \quad (6)$$

$$T_{w1_{2000_z}} = 0.4732T_{a98} + 0.6409T_{a99} + 0.0321h_{r98} + 0.2316h_{r99} - 0.2366r_{s98}, \quad (7)$$

respectively, where $T_{w1_{2000}}$ is the hourly average water temperature value estimated in 2000, in °C, and the prefix z indicates a standardized variable.

In order to get T_{w2000} values, an inverse standardization process should be performed:

$$T_{w1_{2000_z}} = \widetilde{\sigma}_{w2000}(0.4732T_{a98} + 0.6409T_{a99} + 0.0321h_{r98} + 0.2316h_{r99} - 0.2366r_{s98}) + \widetilde{T}_{w2000}. \quad (8)$$

For forecasting purposes, mean and standard deviations were estimated as follows:

$$\widetilde{T}_{w2000} = \left(\frac{\overline{T}_{w98} + \overline{T}_{w99}}{2}\right), \quad (9)$$

$$\widetilde{\overline{\sigma}}_{w2000} = \left(\frac{\sigma_{w98} + \sigma_{w99}}{2}\right), \quad (10)$$

where $\widetilde{T}_{w2000}$ is the estimator of mean water temperature in 2000, in °C; $\overline{T}_{w98}$ is the water temperature in 1998, in °C; $\overline{T}_{w99}$ is the water temperature in 1999, in °C; $\widetilde{\overline{\sigma}}_{w2000}$ is the estimator of standard deviation of water temperature in 2000, in °C; σ_{w98} is the standard deviation of water temperature in 1998, in °C; σ_{w99} represents the standard

deviation of water temperature in 1999, in °C. Table 2 shows the mean square error (MSE) obtained with both models.

The mean (μ_r) and the standard deviation (σ_r) of the residuals, calculated as the difference between measured and calculated water temperatures, appear in Table 3. The measured and calculated data, including their differences for both experiments, are shown in Figures 3 and 4.

In Figures 5 and 6, the measured and calculated data with (6) and (8) are plotted against the identity function to obtain the correlation coefficient ($r_{Tw,Tw1}$), checking the linearity in the fitting. Results given by Figures 3–6 show an improvement in calculated data when standardization is applied; residuals are slightly reduced, fluctuations become softer, and this is verified by the correlation coefficient. The mean square error is reduced in about 30%, and there is less data dispersion (standard deviation of residuals decreases 17%).

4.2. Daily Average Data. In this case, the equations obtained without and with standardization were as follows:

$$T_{w12000} = T_{a98} + \frac{T_{a2000}}{T_{a98}} + \frac{r_{s98}}{(T_{a2000} - 2r_{s98})h_{r98}T_{a2000}/{v_{v2000}}^2 T_{a98}}, \quad (11)$$

$$T_{w1_{2000_z}} = 0.5018T_{a98z} + 0.4982T_{a99z} - 0.2108r_{s98z} - 0.1195v_{v99z} + 0.1195v_{v2000_z} - 0.1195h_{r2000_z}, \quad (12)$$

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

By applying an inverse standardization process,

$$T_{w1_{2000}} = \sigma_{T_{w2000}} T_{w1_{2000z}} + \mu_{T_{w2000}}. \quad (13)$$

In (13), data from 2000 are estimated according to (9) and (10), but considering daily measurements. The mean square errors (MSEs) obtained by using (11) and (13) are detailed on Table 4. The mean (μ_r) and the standard deviation of residuals (σ_r) of this experiment appear in Table 5.

Water temperature variations against time and the obtained differences are plotted on Figures 7 and 8. Figures 9 and 10 show a comparison between measured and calculated daily average water temperatures with respect to the identity function. 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 nonstandardization). Figure 11 shows an example of the performance of the best individual in each generation when the genetic programming algorithm was applied.

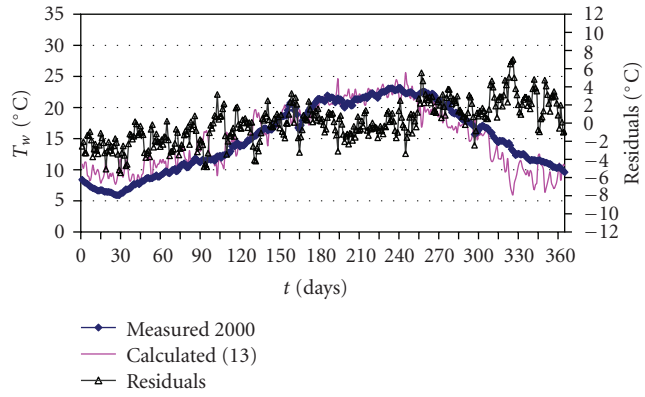

Figure 8: Water temperature values and residuals, experiment with standardization (daily average values).

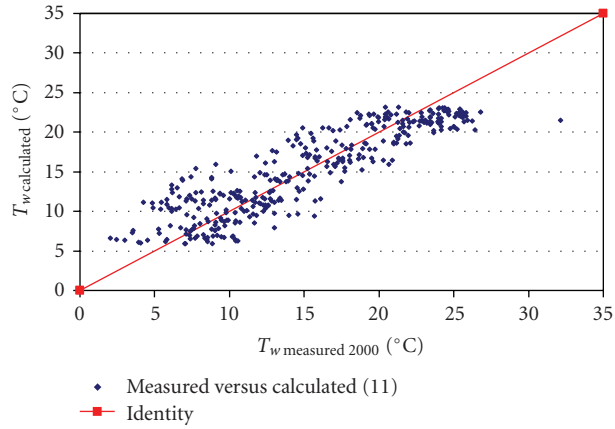

Figure 9: Comparison between measured and estimated data (11), correlation coefficient $r_{Tw,Tw1} = 0.8939$.

4.3. Weekly Average Data. In this last experiment, the equations obtained without and with standardization were

$$T_{w1_{2000}} = 2h_{r98} - v_{v99} - 1.4558v_{v2000} + T_{a98} + \frac{(T_{a2000}/(T_{a99} - v_{v98} - v_{v99} - v_{v2000} + h_{r2000})) + v_{v98}}{v_{v2000}}, \quad (14)$$

$$T_{w12000z} = 0.2962T_{a98} + 0.6819T_{a99} + 0.6397T_{a2000} - 0.2668r_{s98} - 0.3215r_{s99} - 0.3852r_{s98}T_{a2000} + 0.3852r_{s98}r_{s99} - 0.0928, \quad (15)$$

respectively, where $T_{w1_{2000}}$ is the weekly average water temperature value estimated in 2000, in °C; h_{r98} and h_{r2000} are the weekly average relative humidity values recorded in 1998 and 2000, in decimals; T_{a98}, T_{a99}, and T_{a2000} are the weekly average air temperatures of 1998, 1999, and 2000, in °C; v_{v98}, v_{v99}, and v_{v2000} are the weekly average wind speeds from years 1998, 1999, and 2000, in m/s; r_{s98} and r_{s99}, are the

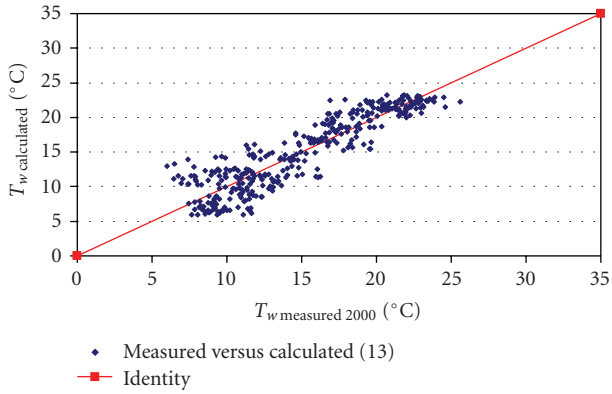

FIGURE 10: Comparison between measured and estimated data (13), correlation coefficient $r_{Tw,Tw1} = 0.9091$.

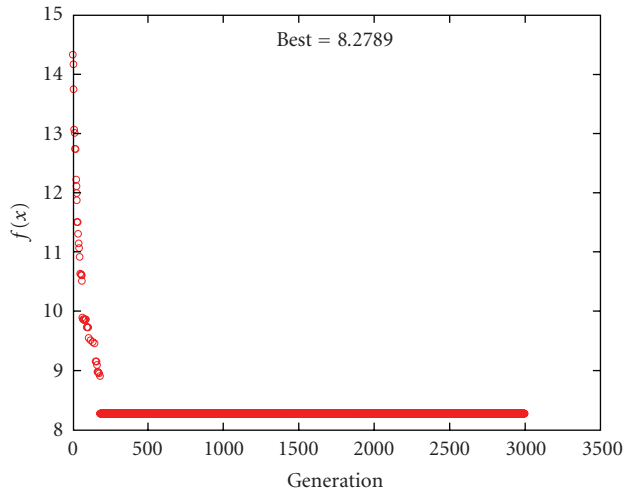

FIGURE 11: Convergence of a genetic programming run.

weekly average solar radiation values of 1998 and 1999, in W/m^2; the prefix z indicates standardized variable.

Equation (15) must be nonstandardized to get the average weekly temperature approach:

$$T_{w1_{2000}} = \sigma_{T_{w2000}} T_{w1_{2000_z}} + \mu_{T_{w_{2000}}}. \quad (16)$$

Mean square errors and statistics of residuals appear in Tables 6 and 7. Figures 12, 13, 14 and 15 show the behavior of water temperature in this weekly analysis.

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

5. Getting the Daily Water Temperature for the Year 2004

The climatic daily data measured from 2002 to 2003 in Flix and Miravet stations were taken to estimate water temperature in the year 2004, in order to check the accuracy

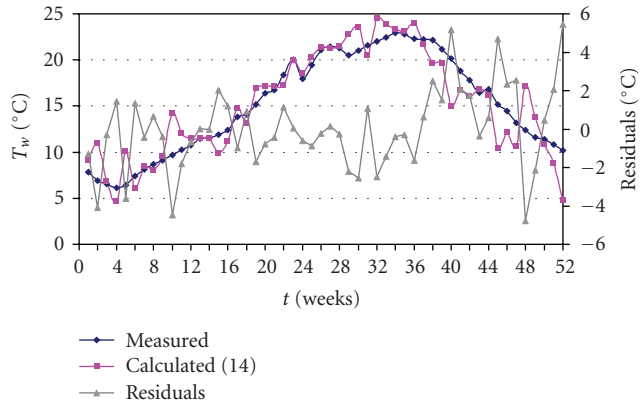

FIGURE 12: Water temperature values and residuals, experiment without standardization (weekly average values).

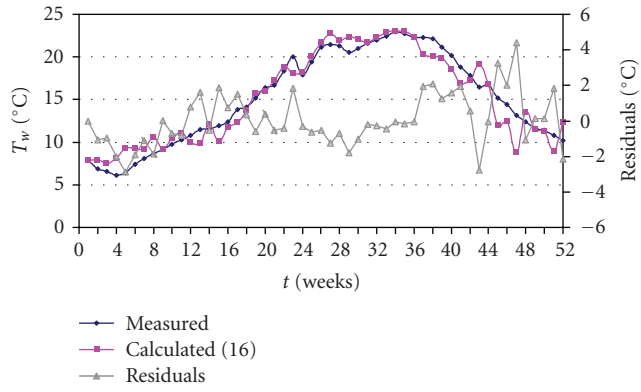

FIGURE 13: Water temperature values and residuals, experiment with standardization (weekly average values).

TABLE 6: Mean square error values. Weekly average data.

Equation	MSE, °C
(14)	4.538
(16)	2.176

TABLE 7: Statistics of the residuals. Weekly average data.

Equation	μ_r (°C)	σ_r (°C)
(14)	0.0186	2.1509
(16)	0.0239	1.4892

of models given by (11), (12), and (13). The climatic data for 2004 needed by the models were assumed as the average of the years 2002 and 2003. The average water temperature for 2004 was assumed 15°C with a standard deviation of 5°C.

A mean square error of 49.549 and a correlation coefficient of 0.6744 were obtained by applying (11) as it is shown in Figures 16 and 17, with an important variation in the residuals; by contrast, with (13) that demands standardized data, a mean square error of 13.027 and a correlation

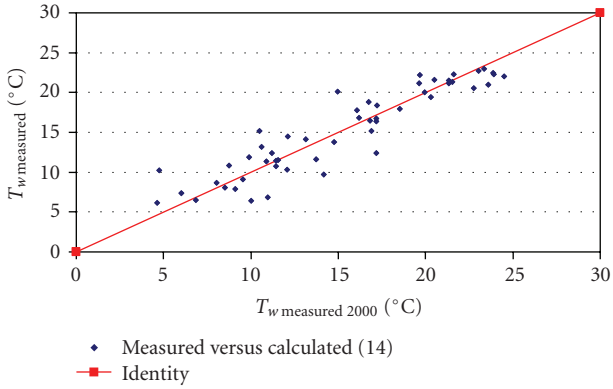

Figure 14: Comparison between measured and estimated data (14), correlation coefficient $r_{Tw,Tw1} = 0.9241$.

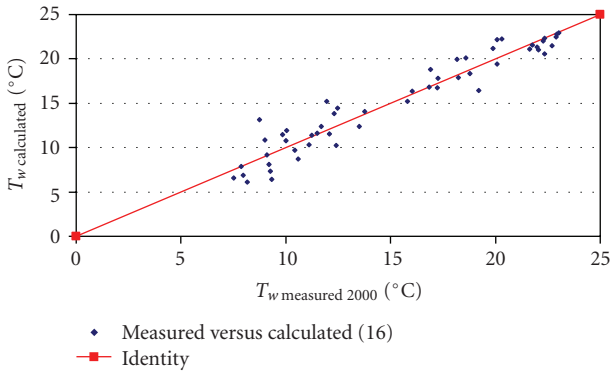

Figure 15: Comparison between measured and estimated data (16), correlation coefficient $r_{Tw,Tw1} = 0.9612$.

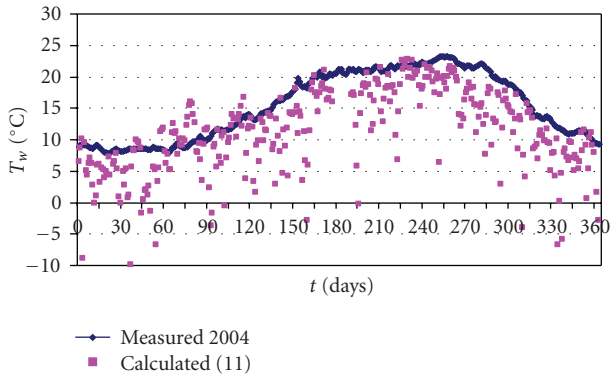

Figure 16: Measured and predicted water temperature for 2004, model without standardization (daily average values).

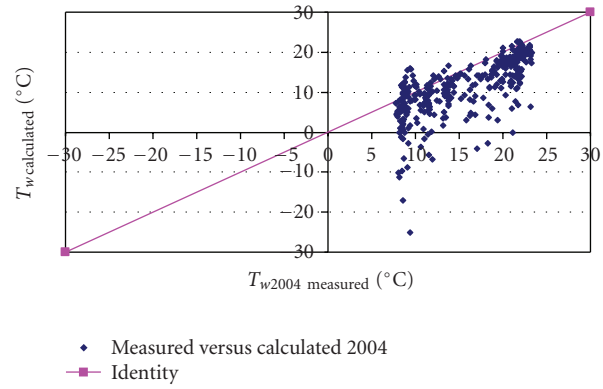

Figure 17: Comparison between measured and estimated data (11), correlation coefficient $r_{Tw,Tw1} = 0.6744$.

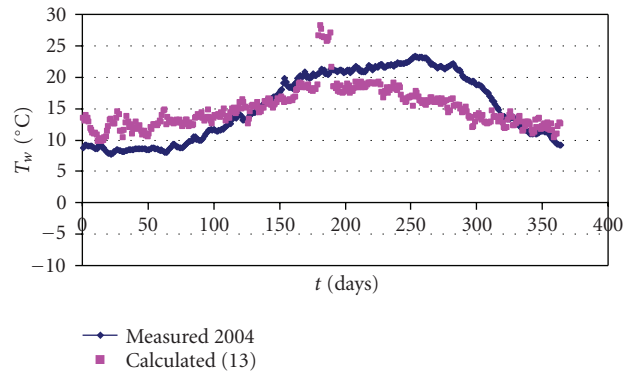

Figure 18: Measured and predicted water temperature for 2004, model with standardization (daily average values).

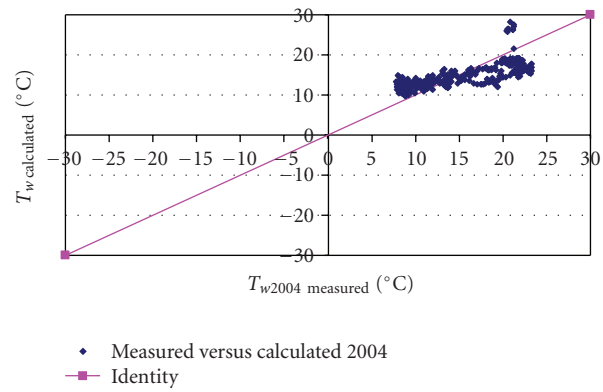

Figure 19: Comparison between measured and estimated data (13), correlation coefficient $r_{Tw,Tw1} = 0.7445$.

coefficient of 0.7445 were obtained; the residuals took values between −7°C and almost 8°C (Figures 18 and 19). Therefore, this last model had a better performance in daily data, in this year.

With both equations very big residuals for water temperature were obtained for some days of the estimated year.

6. 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, in an attempt to explain the possible evolution of the water temperature behavior.

The GP algorithm considered as input hourly, daily, and weekly average measured data without and with standardization, in order to analyze the resulting equations when the shape of the input data varies from one form to another.

Intrinsically, measured data of water temperature and climatic variables have more oscillations in hourly average data than in daily or weekly average data. Particularly, in the experiment using hourly data, the GP algorithm amplifies the water temperature oscillations, probably because in the actual physical process, the oscillations of the climatic variables are filtered. 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.

According to the mean square errors, the standard deviation of residuals, and the correlation coefficient, when weekly data were considered, GP algorithms produced models more capable to follow the behavior of water temperature. This was particularly true for those models obtained with standardized data.

Therefore equations such as those obtained herein can be used as a first approximation to predict changes in water temperature when changes occur in climatic variables such as air temperature, wind speed, relative humidity, and solar radiation, all of which affect the water temperature as well as the physical and chemical water conditions, including the flora and fauna of a river.

When the models for daily data were applied in another year, lower correlations between measured and predicted data were obtained, particularly with the model that does not take into account standardized variables.

According to these results, it is feasible to obtain some improvements in generating water temperature models by means of genetic programming, when the standardization process is incorporated.

Results also show limits on the models developed herein; the models produced oscillations in the water temperatures that do not correspond to the measured data; the results of forecasting from 2004 are only fair. That is probably due to the fact that some variables included in the physical phenomena are eliminated, and the filtering that occurs in nature is not reproduced; nevertheless, these results are considered useful as a first-order explanation of a complex process. However future work is suggested to compare the proposed method with physically based ones.

References

[1] D. A. Savic, G. A. Walters, and J. W. Davidson, "A genetic programming approach to rainfall-runoff modelling," *Water Resources Management*, vol. 13, no. 3, pp. 219–231, 1999.

[2] H. Madsen, M. B. Butts, S. T. Khu, and S. Y. Lyong, "Data assimilation in rainfall-runoff forecasting," in *Proceedings of the 4th Conference of Hydroinformatics*, pp. 1–8, Cedar Rapids, Iowa, USA, July 2000.

[3] J. Dorado, J. R. Rabuñal, J. Puertas, A. Santos, and D. Rivero, "Prediction and modelling of the flow of a typical urban basin through genetic programming," in *Proceedings of the European WorkShops on Applications of Evolutionary Computation (EvoWorkshops '02)*, vol. 2279 of *Lecture Notes in Computer Science*, pp. 190–201, 2002.

[4] Y.-S. Hong and M. R. Rosen, "Identification of an urban fractured-rock aquifer dynamics using an evolutionary self-organizing modelling," *Journal of Hydrology*, vol. 259, no. 1–4, pp. 89–104, 2002.

[5] M. Keijzer and V. Babovic, "Declarative and preferential bias in GP-based scientific discovery," *Genetic Programming and Evolvable Machines*, vol. 3, pp. 41–79, 2002.

[6] E. L. Harris, V. Babovic, and R. A. Falconer, "Velocity predictions in compound channels with vegetated floodplains using genetic programming," *International Journal of River Basin Management*, vol. 1, no. 2, pp. 117–123, 2003.

[7] M. Keijzer, M. Baptist, V. Babovic, and J. R. Uthurburu, "Determining equations for vegetation induced using genetic programming," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '05)*, pp. 1999–2006, Washington, DC, USA, June 2005.

[8] B. Lehner, P. Döll, J. Alcamo, T. Henrichs, and F. Kaspar, "Estimating the impact of global change on flood and drought risks in Europe: a continental, integrated analysis," *Climatic Change*, vol. 75, no. 3, pp. 273–299, 2006.

[9] J. Seguí, *Análisis de la Serie de Temperatura del Observatorio del Ebro 1894–2002*, Observatori de l'Ebre, Roquetes, Spain, 2003.

[10] B. W. Webb and F. Nobilis, "Water temperature behaviour in the River Danube during the Twentieth Century," *Hydrobiologia*, vol. 291, no. 2, pp. 105–113, 1994.

[11] R. Greve, "On the response of the Greenland ice sheet to greenhouse climate change," *Climatic Change*, vol. 46, no. 3, pp. 289–303, 2000.

[12] D. W. Schindler, "Widespread effects of climatic warming on freshwater ecosystems in North America," *Hydrological Processes*, vol. 11, no. 8, pp. 1043–1067, 1997.

[13] M. Álvarez Cobelas, J. Catalán, and D. García De Jalón, "Impactos sobre los ecosistemas acuáticos continentales," in *Evaluación Preliminar de Los Impactos en España por Efecto del Cambio Climático*, J. M. Moreno, Ed., pp. 113–146, Ministerio de Medio Ambiente, Madrid, Spain, 2005.

[14] D. Caissie, N. El-Jabi, and M. G. Satish, "Modelling of maximum daily water temperatures in a small stream using air temperatures," *Journal of Hydrology*, vol. 251, no. 1-2, pp. 14–28, 2001.

[15] J. M. Hellawell, *Biological Indicators of Freshwater Pollution and Environment Management*, Elsevier, London, UK, 1986.

[16] I. J. Winfield and J. S. Nelson, *Cyprinid Fishes. Systematics, Biology and Exploitation*, Chapman & Hall, London, UK, 1991.

[17] G.-R. Walther, E. Post, P. Convey, et al., "Ecological responses to recent climate change," *Nature*, vol. 416, no. 6879, pp. 389–395, 2002.

[18] C. D. Harvell, C. E. Mitchell, J. R. Ward, et al., "Climate warming and disease risks for terrestrial and marine biota," *Science*, vol. 296, no. 5576, pp. 2158–2162, 2002.

[19] S. R. Val, *Incidencia de los embalses en el comportamiento térmico del río. Caso del sistema de embalses Mequinenza-Ribarroja-Flix en el río Ebro*, Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 2003.

[20] J. Prats, R. Val, J. Armengol, and J. Dolz, "A methodological approach to the reconstruction of the 1949–2000 water temperature series in the Ebro River at Escatrón," *Limnetica*, vol. 26, no. 2, pp. 293–306, 2007.

[21] M. L. Arganis, S. R. Val, V. K. Rodríguez, M. R. Domínguez, and R. J. Dolz, "Comparación de curvas de ajuste a la Temperatura del Agua de un río usando programación genética," in *Congreso Mexicano de Computación Evolutiva (COMEV '05)*, pp. 1–8, Universidad Nacional Autónoma de Aguascalientes, May 2005.

[22] M. L. Arganis, S. R. Val, M. R. Domínguez, V. K. Rodríguez, R. J. Dolz, and J. Eaton, "Comparison between equations obtained by means of multiple linear regression and genetic programming to approach measured climatic data in a river," in *IWA Watermex*, pp. 1–8, Washington, DC, USA, May 2007.

[23] T. Bäck, *Evolutionary Algorithms in Theory and Practice*, Oxford University Press, Oxford, UK, 1996.

[24] N. L. Cramer, "A representation for the adaptive generation of simple sequential programs," in *Proceedings of the International Conference on Genetic Algorithms and the Applications*, J. J. Grefenstette, Ed., pp. 183–187, 1985.

[25] J. R. Koza, "Hierarchical genetic algorithms operating on populations of computer programs," in *Proceeding of the 11th International Joint Conference on Artificial Intelligence*, vol. 1, pp. 768–774, Morgan Kaufmann, 1989.

[26] J. E. Edinger, D. K. Brady, and J. C. Geyer, "Heat exchange and transport in the environment," Tech. Rep. 14, Electric Power Research Institute, Palo Alto, Calif, USA, 1974.

[27] The MathWorks, "MATLAB Reference Guide," The MathWorks, Inc., 1992.