ELSEVIER

Ecological Modelling 120 (1999) 141–156

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

pH modelling by neural networks. Application of control and validation data series in the Middle Loire river

Florentina Moatar [a,*], Françoise Fessant [b], Alain Poirel [c]

[a] *LTHE, UMR 5564, CNRS-INPG-ORSTOM-UJF, BP 53, 38041, Grenoble Cedex 9, France*
[b] *INRETS-MAIA, 2 avenue du General Malleret Joinville, 94114, Arcueil, France*
[c] *EDF-Division Technique Générale, 21, avenue de l'Europe, BP 41, 38 040, Grenoble Cedex 9, France*

Abstract

Artificial neural networks (ANNs) are applied as a new type of model to estimate the daily pH of the Middle Loire river. The model is used for pH measurement screening, error detection (abnormal values, discontinuities and recording drifts) and validating the collected data. The measured values of pH are compared with the values estimated by the ANN model using statistical tests to verify homogeneity and stationarity. River water pH is affected by numerous processes: biological, physical and geochemical. Examples are: CO_2 pressure equilibrium with the atmosphere, photosynthesis, respiration of plants, organic matter degradation, geological and mineral background, pollution etc. Inter-relationships between these processes and pH values are complex, non-linear and not well understood. As a neural network provides a non-linear function mapping of a set of input variables into the corresponding network output, without the requirement of having to specify the actual mathematical form of the relation between the input and output variables, it has the versatility for modelling a wide range of complex non-linear phenomena. For this reason the neural network approach has been selected and tested for pH modelling. We used the classical multilayer perceptron model (MLP).

River discharge and solar radiation variables are used as inputs to the MLP model. The choice of these variables is dictated by what are perceived to be the predominant processes that control pH in the Middle Loire river, which is typically eutrophic during the low-flow summer period. The influence of the previous day's flows and radiation has been evaluated in the calibration and verification test. The best network found to simulate pH was one with two input nodes and three hidden nodes. The inputs are: daily discharge and a variable called 'Index of anterior radiation', i.e. calculated as an exponential smoothing of the daily radiation variable. When calibrated over 4 years of data and tested (i.e. verified) for a one-year independent set of data, the model proved satisfactory on pH simulations, with accuracies in the order of 86%. After elaborating the pH model, the Student test and the cumulative Page–Hinkley test were applied for automatic detection of changes in the mean of the residuals from the ANN pH model. This analysis has shown that such tests are capable of detecting a measurement error occurring over a short period of time (1–4 days).

Keywords: Artificial neural network; Middle Loire river; pH; River discharge; Solar radiation

* Corresponding author. Fax: + 33-476-825-286..
E-mail address: moatar@hmg.inpg.fr (F. Moatar)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00098-8

1. Introduction

French environmental regulations impose continuous monitoring of the aquatic environment at every river-site equipped with a nuclear power plant. Therefore 'Electricité de France' performs continuous data acquisition of four parameters: temperature, electrical conductivity, dissolved oxygen and pH, on an hourly basis. Field measurements do not always give a perfect view of reality. The sensor may have a bad contact due to fouling, clogging or lack of maintenance. The measurement can be influenced by external factors: humidity, temperature extremes or electromagnetic fields. The calibration of the measuring instrument may also give rise to problems. Experience has demonstrated the need to verify measurements in order to be able to distinguish between the different reasons for an anomaly: brief and unexpected though real fluctuations, systematic or progressive error in a sensor or progressive evolution of the parameter being measured. A method to critically analyse these data has been developed (Moatar, 1997). The method combines modelling and statistical evaluation. The modelling facilitates the estimation of the pH parameter values and the statistical decision tests allow the verification of the coherence of the measurements to detect inherent errors.

In this paper, the modelling of pH using neural networks and details on how to use this technique for the critical analysis of data are presented. In water, the pH is affected by the water's chemistry, particularly the concentration of some of the CO_2-system components (CO_2, $H_2CO_3^-$ and CO_3^{2-}) according to the equilibria reactions (Stumm and Morgan, 1981). The concentration of CO_2 is a function of the CO_2 pressure equilibrium with the atmosphere, as well as photosynthesis, respiration of plants and the degradation of organic matter. Under acidic conditions, where water chemistry is predominant, the pH is directly related to the flow. Several authors have modelled this relation after linearisation using regression or Box and Jenkins (1976) transfer functions (Whitehead et al., 1986; Fisher et al., 1988; Hirst, 1992). Under alkaline conditions, the CO_2 concentration which affects the pH is principally related to photosynthesis. Photosynthesis is driven predominantly by solar radiation, nutrients, temperature and algal biomass. In the eutrophic Slapy reservoir (Nesmerak and Straskraba, 1985), methods of time series analysis (Box and Jenkins, 1976) have been used to identify relationships between automatic measurements of major driving (i.e. input) variables and changes of pH as an expression of photosynthesis. These analyses have suggested that daily changes of pH are closely related to changes in solar radiation and water temperature.

The site selected for this study is the Dampierre power plant, which is located in the Middle Loire River (Fig. 1). At this location, the stream is typically eutrophic (the amount of chlorophyll-A being up to 150–250 mg/m^3) during summer low-flows. The high level of phytoplankton photosynthetic activity (>0.6 mg C/h during summer) controls the physical–chemical characteristics of the water body at this period, notably the pH. Compared with lake and reservoir studies, the strong variation in the hydrological regimes throughout the year makes river discharge a predominant parameter in determining algal biomass (Recknagel et al., 1997) and other physical and chemical variables, including pH. This was illustrated for the Dampierre site by the Principal Component Analysis run on 104 data series over 13 years (Lair and Sargos, 1993). For this site, the pH can be considered as a function of the flow and the variables characteristic of photosynthetic activity which are themselves related to the hydraulic regime and energy exchanges between the water body and the atmosphere. The purpose of the model is to quickly furnish probable pH values to validate the measured values. The calculation is based on reliable variables which are measured on a daily basis. For this reason we excluded from our model algal biomass, nutrients and carbonates which are not reliable measurements and are only measured one or twice monthly. Only the discharge and solar radiation data were used in the model. The water temperature is measured by the same monitoring system as the pH. We choose to use only those parameters which are measured independently. We did, however, test the sensitivity of the influence of temperature on the model.

A preliminary study of the daily pH-daily discharge relationship at the Dampierre station suggested that it has a non-linear and complex shape (Fig. 2). By segmenting the data after solar radiation ($S(t) < 200$ W/m^2 and $S(t) > 200$ W/m^2) we can improve the correlation of the relationship between the daily discharge and the daily pH. Moreover, the data series are nonstationary, i.e. the basic statistical characteristics such as mean and standard deviation of the process change with the time. The interannual mean and standard deviation of pH present a complex periodic behaviour (Moatar, 1997). The standard deviation of pH displays a strong annual variability not directly related to the absolute level of the pH. Transformations of the pH data usually used for modelling water resources time series (Box and Jenkins, 1976; Salas et al., 1980) do not induce complete stationarity. Discharge data series do not follow a normal distribution. In this case the Box–Cox transform (Box and Cox, 1964) is usually used to obtain normally distributed data (Lemke, 1991). For instance, the linear time series models such as ARMAX (auto-regressive moving average with exogenous inputs) models developed by Box and Jenkins (1976) are not applicable.

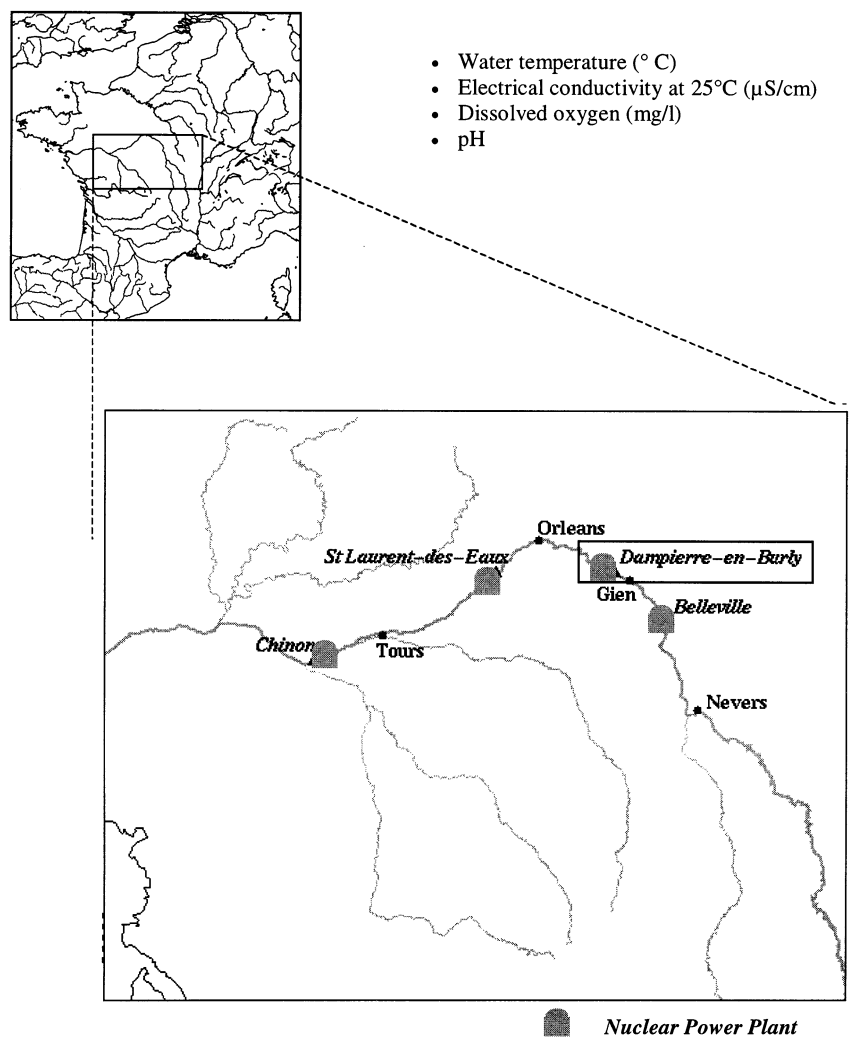

Fig. 1. Location and equipment of the Dampierre en Burly study site.

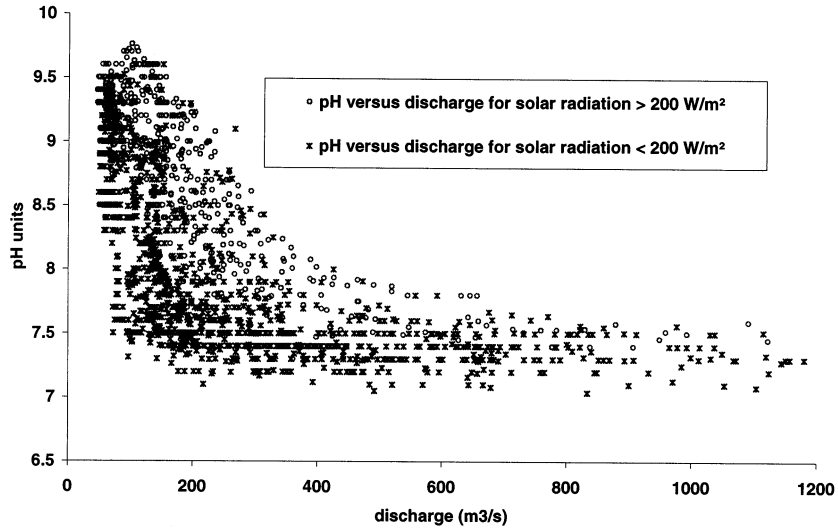

Fig. 2. Daily pH versus daily discharge.

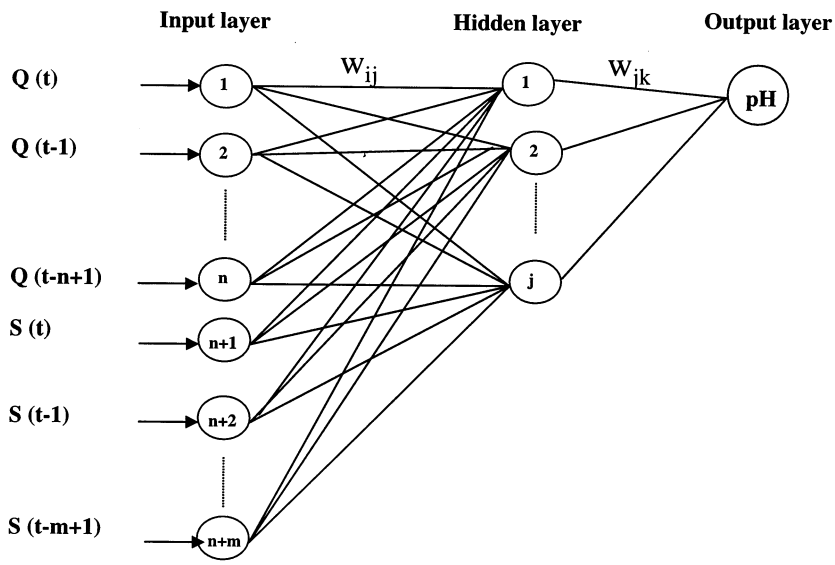

Fig. 3. Structure of the three-layer feed-forward artificial neural network used in this study.

However, when developing ANN models, the nonstationarities in the data are accounted for by the hidden layer nodes and the statistical distribution of the data does not need to be known (Maier and Dandy, 1996). Neural network models have been largely studied for the last 15 years. Although they first proceeded from physical, biological or psychological works about modelling, their use has broadly spread out to many different scientific areas. Neural networks are usually used as a particular type of non parametrical statistical model (Thiria et al., 1997). The most important and interesting characteristics shared by most neural networks models may be summarised as follows: non linear modelling capacity, generic modelling capacity, robustness to noisy data and ability to deal with high dimensional data. In the analysis of water resource phenomena, ANNs are

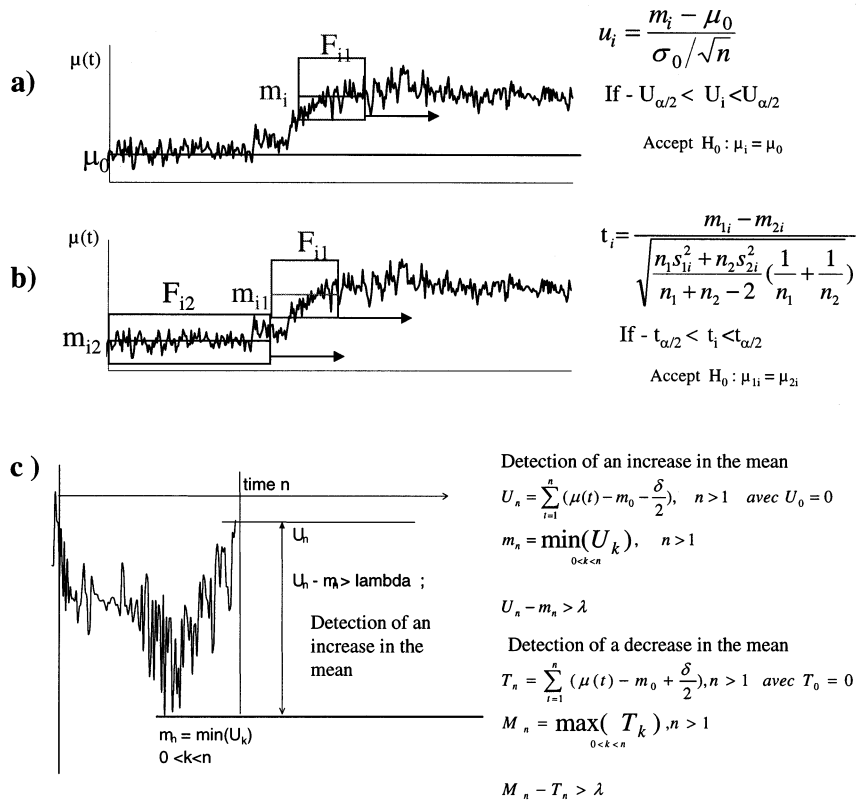

Fig. 4. Statistical tests.

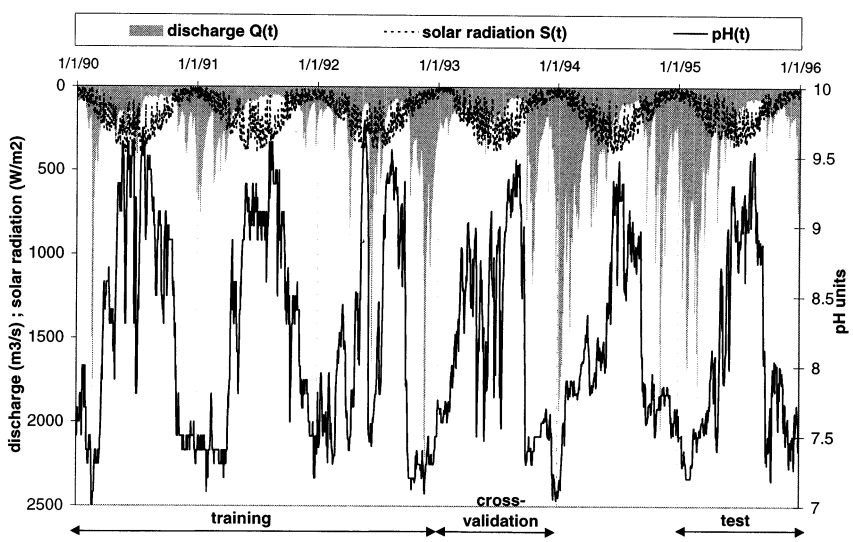

Fig. 5. Plot of the flow, solar radiation and pH time-series under study (1990–1995).

Table 1
Comparison of regression and ANN for single inputs variables

Single inputs variables	Regression		ANN	
	E criterion	S.D.* (pH units)	*E* criterion	S.D.* (pH units)
$Q(t)$	0.45	0.50	0.72	0.34
Log $Q(t)$	0.69	0.39	0.71	0.34
$S(t)$	0.37	0.53	0.42	0.53
$T(t)$	0.33	0.54	0.26	0.55

* S.D. = standard deviation of residuals.

Table 2
Comparison of regression and ANN for multiple inputs variables

Multiple inputs variables	Regression		ANN	
	E criterion	S.D.* (pH units)	*E* criterion	S.D.* (pH units)
$Q(t)$ S (t)	0.60	0.41	0.77	0.30
$Q(t)$ $T(t)$	0.61	0.41	0.73	0.34
Log $Q(t)$ $S(t)$	0.73	0.31	0.73	0.33
Log $Q(t)$ $T(t)$	0.74	0.33	0.77	0.31
$Q(t)$ $S(t)$ $T(t)$	0.62	0.41	0.71	0.35
Log $Q(t)$ $S(t)$ $T(t)$	0.76	0.32	0.74	0.33

* S.D. = standard deviation of residuals.

typically used to model the relation between rainfall and runoff (Dimopoulos et al., 1996; Minns and Hall, 1996). The ANN is shown to provide a better representation of the rainfall-runoff relationship than the linear ARMAX time series approach (Hsu et al., 1995; Lek et al., 1996b). In the domain of ecological modelling successful results have been obtained. For instance, Recknagel et al. (1997) studied the relationship between different species of algae and several limiting factors such as: solar radiation, nutrient concentrations, density and composition of zooplankton. Lek et al. (1996a) applied ANNs to modelling fish diversity with respect to riverine habitat characteristics.

2. The data base and the methods used in this study

2.1. Site and monitoring system description

The Loire river has a length of $\approx 1\,012$ km and a drainage area covering 115 000 km^2 of the centre and the west of France (Fig. 1). The Dampierre site, considered in this study, is situated 550 km from the source and drains 35 500 km^2 of watershed. It has the longest available record (1990–1995) of water quality parameters measurements. The monitoring system consists of a floating platform including a temperature sensor for direct measurement of water temperature (at a depth of 20 cm) in the river course and a pumping device sending a small flow of water (approx. 0.5 l/s) to the three following electrodes: pH (range: 0–14 pH unit; accuracy: $\pm 0.2\%$), Dissolved Oxygen (DO) (range: 0–20 mg/l; accuracy $\pm 1\%$) and electrical conductivity at 25°C (range: 0–1000 μS/cm; accuracy $\pm 1\%$). The pH accuracy given above is for instantaneous values and is that estimated by the manufacturer. However, the corresponding accuracy of pH (including electrode, transmission, and calibration) estimated *in situ* by comparison with laboratory measurements for the maintenance department of the Dampierre site are closer to ± 0.3 pH units. Accu-

racy is defined as two times the standard deviation (S.D.) of the check-sample readings. Furthermore, these instantaneous values, taken every 5 s are not archived as such, but as an hourly mean, which in fact is the average over 50 min (the remaining 10 min being used for the circuit cleaning cycle). In this study daily pH values were used. The hydrometeorological data used in connection with the pH data are the discharges at the Dampierre site (obtained from water-level records and a rating curve; range 46–2900 m^3/s; accuracy 8–10%) and daily solar radiation (W/m^2) measured at the meteorological station of the city of Tours located 185 km from the study site.

2.2. pH modelling by artificial neural networks

In this study, the neural network model used is the classical multilayer perceptron (MLP) with

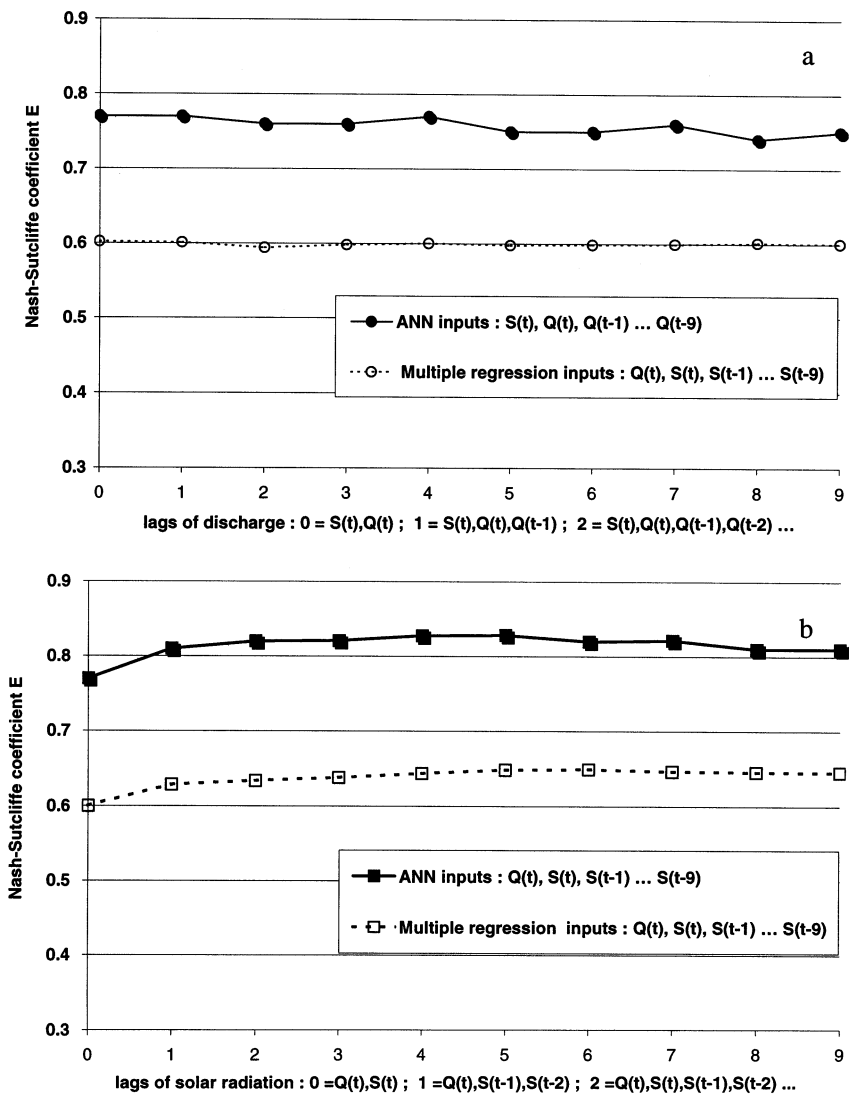

Fig. 6. Nash–Sutcliffe coefficient: (a) lags of discharge; (b) lags of solar radiation.

Table 3
Comparison of regression and ANN for multiple inputs variables: $Q(t)$ and $IS(t)$ for different values of the weighting parameter β

Multiple inputs variables	Regression		ANN	
	E criterion	S.D.* (pH units)	E criterion	S.D.* (pH units)
$Q(t)$ $IS(t)$; $\beta = 0.1$	0.61	0.41	0.79	0.29
$Q(t)$ IS (t); $\beta = 0.2$	0.61	0.41	0.79	0.28
$Q(t)$ IS (t); $\beta = 0.3$	0.62	0.40	0.80	0.28
$Q(t)$ IS (t); $\beta = 0.4$	0.63	0.40	0.81	0.28
$Q(t)$ IS (t); $\beta = 0.5$	0.63	0.39	0.82	0.27
$Q(t)$ IS (t); $\beta = 0.6$	0.64	0.39	0.82	0.27
$Q(t)$ IS (t); $\beta = 0.7$	0.65	0.38	0.83	0.26
$Q(t)$ IS (t); $\beta = 0.8$	0.64	0.38	0.79	0.29
$Q(t)$ IS (t); $\beta = 0.9$	0.63	0.39	0.79	0.29

* S.D. = standard deviation of residuals.

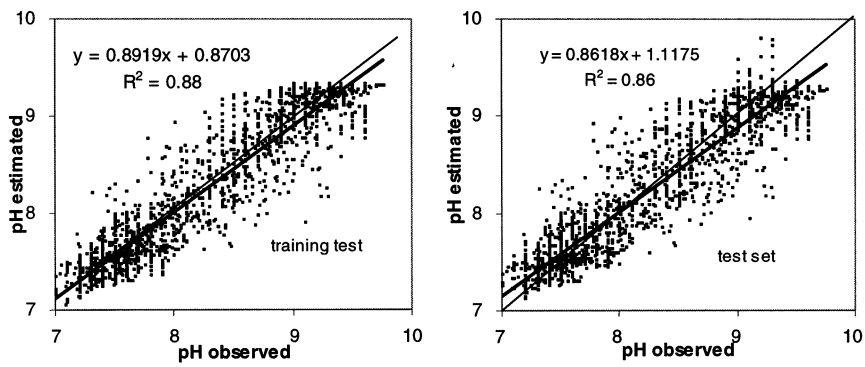

Fig. 7. Estimated versus observed pH values for the 5 years calibration period (left) and for 5 years verification (right).

one layer of hidden neurons (Fig. 3). It was developed using the commercially available software package Matlab-Neural Network Toolbox (The MathWorks Inc., 1998). The MLP consists of a large number of highly connected non-linear simple neurons. We can differentiate three types of neurons: input, output and hidden neurons. The input neurons receive information to be processed, in our case the discharge $Q(t)$ and solar radiation $S(t)$ (eventually incorporating also the discharge and solar radiation from previous days). The output neurons give the results of the neural network. In this case we have only one neuron which should return the result of the dependent variable pH (t). The hidden neurons which are neither input nor output neurons are used to keep an internal representation of the problem. The parameters associated with each of these connections are called weights. Knowledge of the network is kept in these weights. Each hidden and output unit computes its value as the weighted sum of its inputs, passed through a nonlinear function. For a given network architecture, the model calculates the weights that minimize a cost function (generally the mean square error function). Given a cost function, a network architecture and some data, the next step is to find the appropriate weights which minimize the cost function. This is usually done using an iterative procedure. The best known learning mechanism for neural networks is the backpropagation (BPA) rule of Rumelhart et al. (1986). It is a simple

gradient descent technique, which minimizes the cost function in weight space by modifying the weights in the opposite direction of the gradient error with respect to the weights. The BPA is often too slow for practical problems. Since 1986, a variety of improvements have been proposed (introduction of a momentum term, use of conjugate gradient techniques, use of second order information, etc.) (Hertz et al., 1991). We used the Levenberg–Marquardt algorithm, an alternative to the conjugate gradient techniques for fast optimization.

One of the most important features of learning systems is their ability to generalize to new situations. An early stopping procedure to stop the learning process was used for improving generalization. In this technique the available data were divided into three subsets. The first subset is the training subset which is used for computing the gradient and updating the network weights. The second subset is the validation set. The error on the validation set is monitored during the training process. The validation error will normally decrease during the initial phase of training, as does the training set error. However, when the network begins to overfit the data, the error on the validation set will typically begin to rise. When the validation error increases, the training is stopped, and the weights at the minimum of validation error are returned. The verification test subset is a set of independent data used to verify the consistency of the efficiency of the model.

The right number of hidden neurons cannot be achieved from a universal formula. Networks with too many parameters tend to memorize the

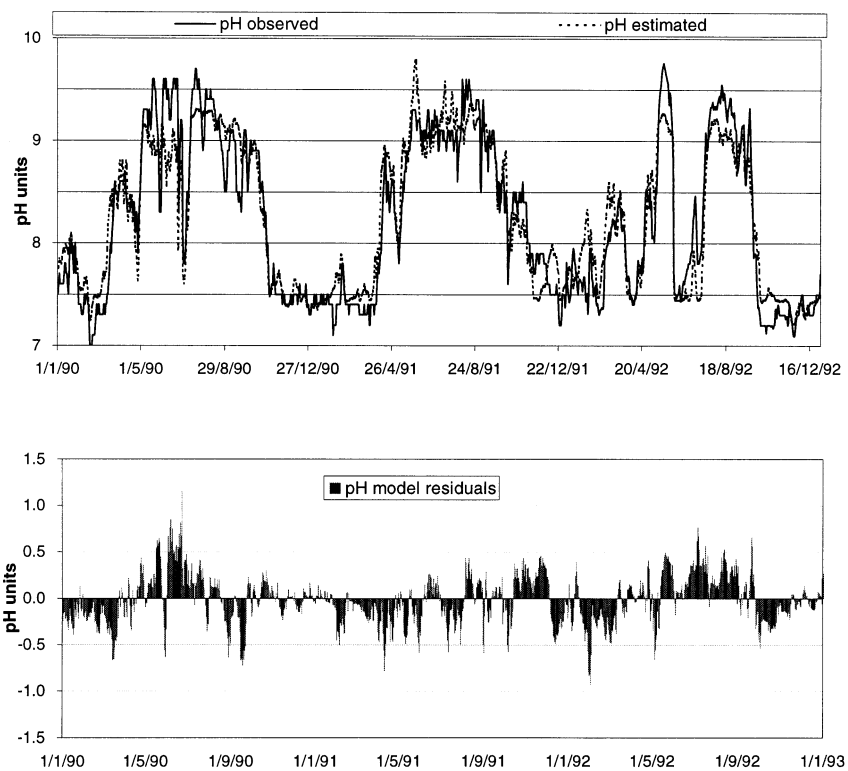

Fig. 8. Observed and estimated pH values for the period 1990–1992 inclusive (upper). Residual pH values for the same period (lower).

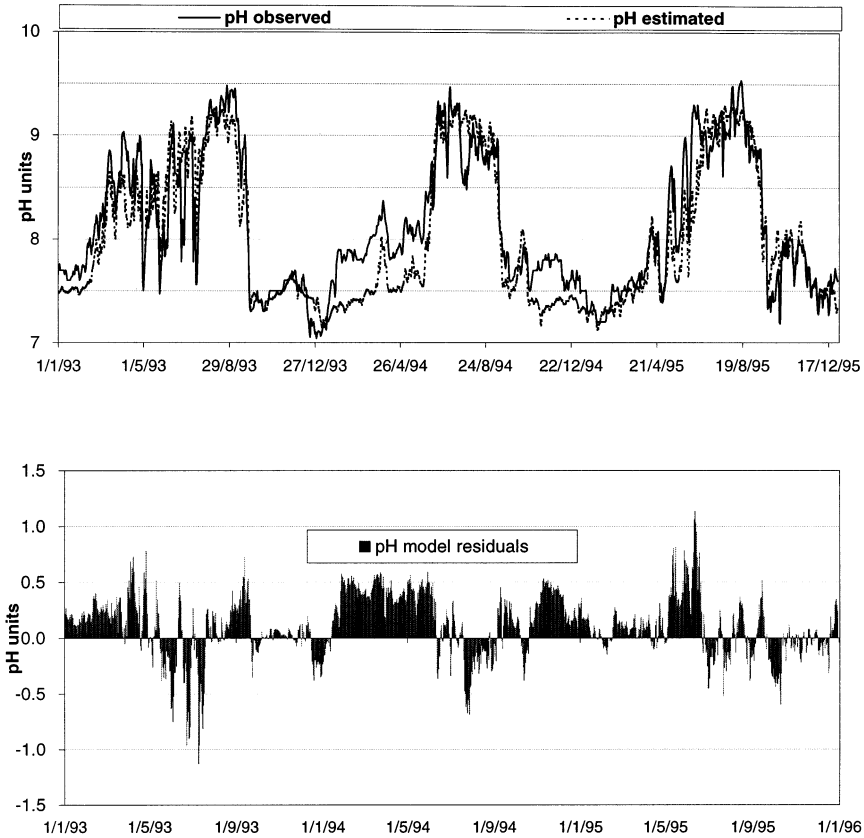

Fig. 9. Observed and estimated pH values for the period 1993–1995 inclusive (upper). Residual pH values for the same period (lower).

Table 4
Statistical evaluation of estimated pH values (verification period)

Subset test year	Observed pH		Estimated pH		pH model residuals	
	Mean	S.D.*	Mean	S.D.	Mean	S.D.
1990	8.35	0.78	8.36	0.67	−0.01	0.28
1991	8.27	0.74	8.33	0.73	−0.06	0.23
1992	8.04	0.77	8.04	0.65	0.00	0.27
1993	8.19	0.66	8.13	0.62	0.06	0.28
1994	8.11	0.55	7.92	0.69	0.20	0.27
1995	8.05	0.69	7.98	0.67	0.07	0.26

* S.D. = standard deviation of residuals.

input patterns, while those with too few hidden parameters may not be able to simulate a complex system at all. We applied a trial-and-error approach to select the best ANN architecture. Our initial model had few parameters, we gradually added hidden neurons during learning until the optimal result is achieved in the test subset.

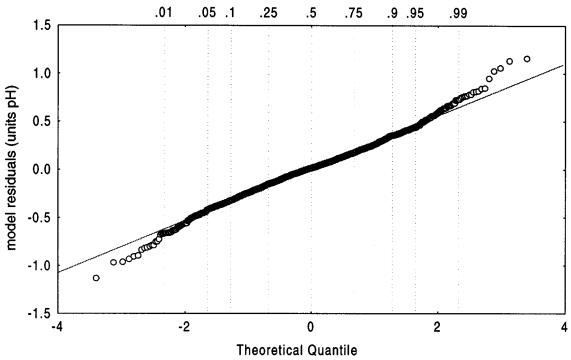

Fig. 10. Empirical distribution of residuals on Gauss paper.

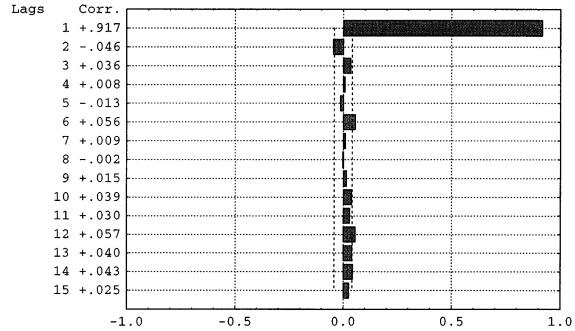

Fig. 11. Partial autocorrelation function of ANNs pH residuals.

2.3. Description of the pH data control method

A method for pH data control was developed after the pH model was built. The measured pH values were compared with those estimated by the ANN pH model using statistical tests in order to verify the homogeneity and the stationarity of the residual error series. These tests are performed for normal variables having independent observations. The series of residuals $\varepsilon(t)$ from the ANN pH model are normal (cf. Fig. 10) but have, in this case, a temporal structure (cf. Fig. 11). The modelling of this series using an autoregressive AR model allows the extraction of the independent residual series $\mu(t)$.

$$\varepsilon(t) = a_1\varepsilon(t-1) + a_2\varepsilon(t-2) + \cdots a_n\varepsilon(t-2) + \mu(t) \quad (1)$$

The order n of the AR model was estimated after analysis of the auto (ACF) and partial (PACF) autocorrelation functions. Finally, two types of statistical test are applied for automatic detection of changes in the mean of the signal $\mu(t)$:

1. The Student test comparing the mean of the values within a sliding window F_{1i} and either a reference mean μ_0 (Fig. 4(a)), or the mean of the values within an anterior sliding window F_{2i} (cf. Fig. 4(b)).
2. The Page–Hynkley test (Basseville, 1986) was performed as a cumulative sum test, where jumps in the mean occur at unknown time instants (Fig. 4(c).

The details of how the tests are applied are presented below:

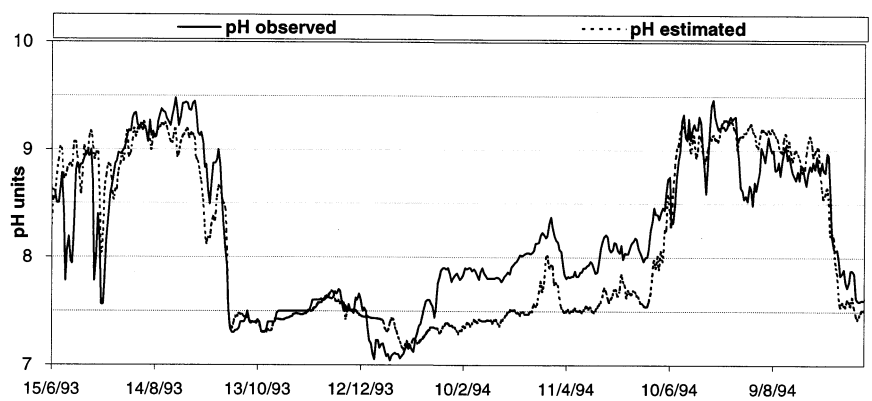

Fig. 12. Estimated and observed pH values for the period 15/06/–15/07/1993 inclusive.

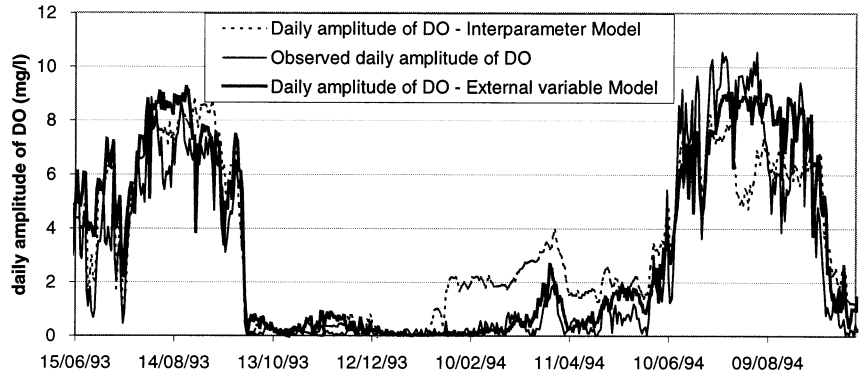

Fig. 13. Daily amplitude of DO measured and estimated with the two models: the 'Interparameter Model' and the 'External Variable Model' for the period 15/06/–15/07/1993 inclusive.

1. For each window, the statistical tests which must evolve according to known probability distribution laws (assuming the hypotheses that we are trying to prove are true) are carried out.
   1.1. To detect a change in the mean of the signal $\mu(t)$, we calculate the mean m_i within the current mobile window F_{1i} and the statistic test u_i The values μ_0 and σ_0 are the mean and S.D. calculated from independent samples known to be free of error measurements. The statistic test u_i follows, assuming no changes, a normal, centred and reduced law. The test is used to verify the hypothesis: $\mu_i = \mu_0$. If this hypothesis is confirmed, the difference between m_i and μ_0 is uniquely due to errors of estimation of the true mean population μ_i by the mean of the sample m_i.
   1.2. If no reference values to test the calculated magnitudes are available, they will be calculated in two windows to allow comparison. The comparison of the mean of the two windows (m_{1i} and m_{2i}) of size n_1 and n_2, is performed for small samples ($n_1 < 30$ and/or $n_2 < 30$), sampled independently from a normal population from unknown variance but assumed to be equal to a common variance value ($\sigma_{1i}^2 = \sigma_{2i}^2 = \sigma^2$). If we assume that hypothesis H_0: $\mu_{1i} = \mu_{2i}$ is true, the statistic test follows a Student law with $n_1 + n_2 - 2$ degrees of freedom (d.f.).
2. The Page–Hinkley test (Fig. 4(c)) consists in fixing a priori a minimum jump magnitude δ to be detected, and running two tests in parallel, because the 'direction' of the jump is not known a priori (increasing or decreasing mean). The detector will set the alarm at the first time n at which $U_n - m_n > \lambda$ (cf. Eq. (2)) for detecting an increase in the mean and at the first time n at which $M_n - T_n > \lambda$ (cf. Eq. (3)), for detecting a decrease in the mean.

$$U_n = \sum_{i=1}^{n} \left(\mu(t) - m_0 - \frac{\delta}{2} \right); \quad n > 0 \text{ and } U_0 = 0$$

$$m_n = \min_{0 \le k \le n} (U_k) \tag{2}$$

$$T_n = \sum_{i=1}^{n} \left(\mu(t) - m_0 + \frac{\delta}{2} \right); \quad n > 0 \text{ and } T_0 = 0$$

$$M_n = \max_{0 < k < n} (T_k) \tag{3}$$

The limit λ is determined by learning. The initial value is calculated by the expression: $\lambda = 2*h*\sigma/\delta$ where $h = 2$ for normal distributions and σ is the standard deviation of the signal (Ragot et al., 1990).

3. Case study

3.1. Determination of appropriate ANNs model parameters

The daily pH, discharge and solar radiation values from the period 1990 to 1995 were used (cf. Section 1). For these series, data sets for the

network training, cross-validation and verification steps were prepared (Fig. 5). Data from 3 years were used for training, 1 year of data was used for cross-validation and 1 year of data was used for verification. The measurements for 1994 were not taken into account in the final calibration and the cross-validation process because of a lack of confidence in the measurements (as explained later). Each of the 5 years was chosen, one at a time, as the verification period, the other 4 years being used as the training and cross-validation data periods. The performance of the model was therefore verified using five different test samples.

For each input variable, the performance of the ANN was compared with the linear regression. The inputs variables tested was discharge $Q\ (t)$, natural logarithm of discharge Log $Q\ (t)$, solar radiation $S\ (t)$ and water temperature $T\ (t)$. Table 1 summarises the results in terms of the Nash and Sutcliffe (1970) efficiency criterion (E criterion) and the S.D. of the residuals in the verification subset. A plain improvement of regression ($E = 0.45$, S.D. $= 0.50$ for $Q\ (t)$ and $E = 0.69$, S.D. $= 0.36$ for Log $Q\ (t)$) is indicated for discharge by

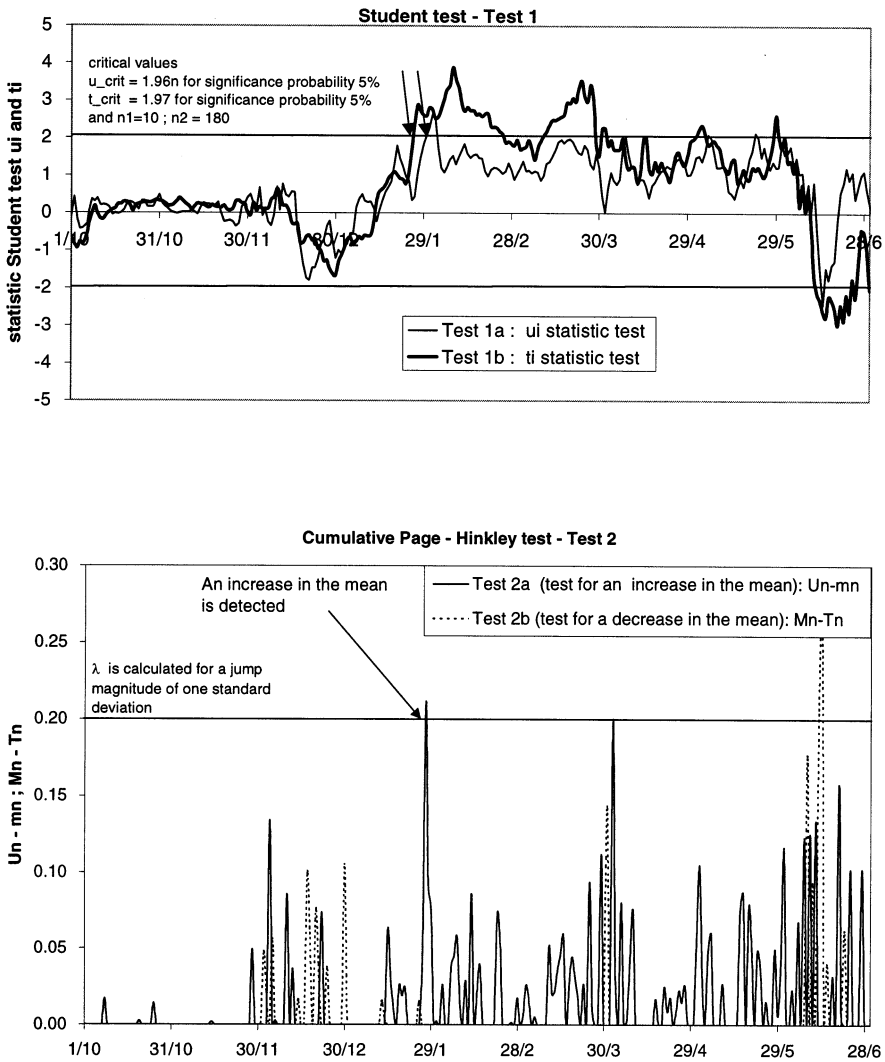

Fig. 14. Control charts: (a) Student's t test; (b) Page–Hinkley test.

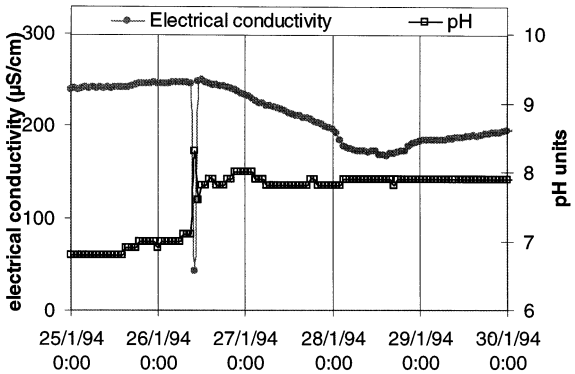

Fig. 15. pH and electrical conductivity (25/01–30/01/1994).

the ANN ($E = 0.72$, S.D. $= 0.34$), that confirms the nonlinear relationship between pH and discharge. In contrast, the relationship between solar radiation and pH appears to be linear, because no improvement of the regression model ($E = 0.37$, S.D. $= 0.53$) is obtained with an ANN model ($E = 0.42$, S.D. $= 0.53$). The same result is obtained if we enter as input the water temperature.

Table 2 presents the results for multiple inputs variable. Both the ANN and the regression model have a better estimation of the pH when the discharge and solar radiation and/or temperature are considered together. The best result for the multiple regression ($E = 0.76$, S.D. $= 0.32$) is obtained when Log Q (t), S (t), T (t) are taken into account. For the ANN, the results are similar for the diverse combination tested with a slight amelioration for the case Q (t) and S (t) ($E = 0.77$, S.D. $= 0.30$) or Log Q (t) and T (t) ($E = 0.77$, S.D. $= 0.31$). The influence of previous day's flows and previous day's radiation was then investigated. The efficiency E was calculated for the regression model and the ANN model (in the verification sets) as follows: initially the coefficient was calculated with the daily radiation S (t) and flows of N previous days Q ($t - N$), N varying from 0 to 9, as the input variables (Fig. 6(a)). It was then calculated for the case where the daily flow Q (t) and solar radiation of the N preceding days S ($t - N$) were used as the input variables (Fig. 6(b)).

Fig. 6(a) shows that in the case of the radiation, the preceding day's flows do not give better results as compared to the flow of the considered day. On the contrary, in the case of the flow, the preceding day's radiation does improve the pH estimation. Thus E, which was 0.77 with two input variables considered (flow and daily radiation), becomes 0.83 for the case of five input variables (flow and daily solar radiation and solar radiation at lag times 1, 2 . . 3, days: $t - 1$, $t - 2$, $t - 3$). To decrease the number of input variables, without losing the influence of the previous day's radiation, an exponential smoothing was applied. This variable has been called 'index of anterior radiation', IS, and has been calculated for a given day in the following manner:

$$IS(t) = \beta\, IS(t-1) + (1-\beta)\, S(t) \quad (4)$$

When applied recursively to each successive observation in the series, each new smoothed value is computed as the weighted average of the current observation S (t) and the previous smoothed observation IS ($t - 1$) depending on the value of the weighting parameter β. The optimal value of β in terms of the Nash–Sutcliffe coefficient, during both calibration and verification, was 0.7 (cf. Table 3). Finally the model has two inputs: Q (t) and IS (t).

We used the tan–sigmoid transfer function on the hidden layer and a linear transfer function on the output layer. In order to select the optimal number of hidden neurons, tests were performed by varying the number of neurons between 1 and 10. The optimal result of the test set is obtained for three neurons in the hidden layer, a choice that is justified by the absence of improvement of the model beyond this value. The data were standardised (zero mean and unity S.D.).

3.2. Results

3.2.1. pH modelling by ANNs

Finally the best model found has two inputs (Q (t) and IS (t)), three hidden neurons and one output for pH (t). The model fits the data well and explains 86% of the pH variance. The correlation coefficient is high in the calibration set ($R^2 = 0.88$) as well as in the verification set ($R^2 = 0.86$), indicating a high consistency of the model efficiency. (cf. Fig. 7). The time series of observed

and estimated values as well as the corresponding series of the residuals for the period between 1990–1992 inclusive and 1993–1995 are presented respectively in Figs. 8 and 9. The model conserves the same mean as the mean of the data. The S.D. of the estimated values is slightly smaller than for the observed values (Table 4). The mean error is zero for each year, except 1994, for which the values are underestimated. The S.D. of the errors vary between 0.23 and 0.28 pH units.

The normality and temporal structure of the residuals were analysed on the test set for 1990–1993 and 1995. Fig. 10 shows that the sample of model residuals is normal in the central part of the distribution and for more than 90% of the data. However, the partial autocorrelation function shows the existence of a temporal (i.e. persistence) structure (the autocorrelation function at lag 1 being equal to 0.9) (Fig. 11).

As shown in Fig. 12, during 5 months in 1994 (from mid-January to mid-June), the difference between the estimated and observed pH values is systematically in the order of 0.5 units. To explain this difference, another parameter measured by the monitoring system was analysed—daily amplitudes of dissolved oxygen, for which two estimation models are available. The first is a linear stochastic model, using the pH from the monitoring system (the 'Interparameter' model). The second model is based on physical principles and has variables which are not measured by the Electricité de France (EDF) monitoring system—river discharge and solar radiation. This model is called the 'External variable model'. As indicated in Fig. 13, from mid-January to mid-June, 1994, the 'External variable model' reproduces the daily amplitudes of dissolved oxygen (DO) quite well while the 'Interparameter model' systematically over-estimates them. This comparison indicates that the pH measurement for this time period is false (calibration error) or that it is significantly influenced by an external phenomenon (e.g. pollution) which cannot readily be explained. The method of critical data analysis applied to the pH and described in the following section shows that this time period is indeed suspect.

3.2.2. Control and validation of pH data

In this section, the results of the statistics tests of detection for the period 1/10/1993–30/6/1994 are presented. Using the method described in Section 2.3, the residuals $\varepsilon(t)$ from the pH model we initially decorrelated. After analysis of the autocorrelation function and the partial autocorrelation function, a first order autorregressive model was used.

$$\varepsilon(t) = 0.86\,\varepsilon(t-1) + \mu(t) \quad (5)$$

The calibration of this model as well as the calculation of the reference values (mean and S.D.), were performed for the 1991 which appears to have the most reliable measurements based on the critical analysis and validation of the other parameters. This year presents the best correlation between the modelled and measured pH values. Fig. 14 presents the test variables calculated for the series $\mu(t)$:

- u_i for Test 1a: the mean of the values in the sliding window containing ten values compared to the mean reference $m_0 = 0$.
- t_i for Test 1b: the mean of the values in the sliding window containing ten values compared to the mean of the window from 6 anterior months.
- $U_n - m_n$ and $M_n - T_n$ for Test 2: cumulative sum of the values from 1/10/1993. This test is re-initialised after each detection.

It is observed that in each of the three tests, the first signal is detected between 27/01/1994 and 30/01/1994, after 4 months of error free measurements (Test 1a: 30/01; Test 1b: 28/01; Test 2b: 27/01). This period corresponds with the beginning of a pH series which was already considered suspect through using the 'Interparameter' model which calculates daily amplitudes of dissolved oxygen (DO) from the pH measurements. Analysing the raw pH data, a systematic difference of 0.5 units during 6 months (previously presented in Section 3.2.1) is noted. This difference corresponds with a discontinuity observed on the 26 /01 at 10:00 (Fig. 15). For the same time step, an 'abnormal' electrical conductivity value was measured. This analysis shows that such tests are capable of detecting a measurement error occurring over a short period of time (1–4 days).

4. Conclusion

The results presented in this paper indicate that ANN clearly give satisfactory responses in the modelling of pH as a function of hydrometeorological data such as discharge and solar radiation. The best network found ($R^2 = 0.86$) to simulate pH was one with two inputs and three hidden nodes. The inputs are daily discharge Q (t) and the IS (t), 'index of anterior radiation', i.e. calculated as an exponential smoothing of the daily radiation variable. The model, which was adopted for its generality and its simplicity, and also because of the availability and reliability of the significant input variables, was integrated into our system of modelling tools which facilitate the critical analyses and validation of physical–chemical measurements. This system of modelling tools is currently in the process of being put into service on-line by EDF to allow them to follow and critically evaluate water quality parameters with respect to hydrometeorological conditions.

References

Basseville, B., 1986. On line detection of jumps in mean. Lect. Notes Contr. Inf. Sci. 77, 12–26.

Box, G., Jenkins, G., 1976. Time Series Analysis; Forecasting and Control, Holden-Day, San Francisco.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc., Ser. B 26, 211–243.

Dimopoulos, I., Lek, S., Lauga, J., 1996. Modélisation de la relation pluie-débit par les réseaux connexionnistes et le filtre de Kalman. Hydrol. Sci. J. 41 (2), 179–193.

Fisher, F., Dickson, K., Rodgers, J., Anderson, K., Slocomb, J., 1988. A statistical approach to assess factors affecting water chemistry using monitoring data. Water Resour. Bull. 24 (5), 1017–1029.

Hertz, J., Krogh, A., Palmer, R.G., 1991. Introduction to the Theory of Neural Computation, Santa Fe Institute Studies in the Sciences of Complexity, Addison Wesley, Reading, MA, 327 pp.

Hirst, D., 1992. A new technique for the analysis of continuously monitored water-quality data. J. Hydrol. 134, 95–102.

Hsu, K-L., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modelling of the rainfall-runoff process. Water Resour. Res. 31 (10), 2517–2530.

Lair, N., Sargos, D., 1993. A 10-year study at four sites of the middle course of the River Loire. I-Patterns of change in hydrological, physical and chemical variables in relation to algal biomass. Hydroécol. Appl. 5 (1), 1–27.

Lek, S., Delacoste, M., Baran, Ph., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996a. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.

Lek, S., Dimopoulos, I., Derraz, M., Ghachtoul, Y.E., 1996b. Rainfall-runoff modelling using artificial neural networks. Rev. Sci. l'Eau 3, 319–331.

Lemke, K., 1991. Transfer function of suspended sediment concentration. Water Resour. Res. 27 (3), 293–305.

Maier, H.R., Dandy, G.C., 1996. The use of artificial neural networks for the prediction of water quality parameters. Water Resour. Res. 32 (4), 1013–1022.

Minns, A.W., Hall, M.J., 1996. Artificial neural networks as rainfall-runoff models. Hydro. Sci. 41 (3), 399–417.

Moatar, F., 1997. Modélisations statistiques et déterministes des paramètres physico-chimiques utilisés en surveillance des eaux des rivières: Application à la validation des séries de mesures en continu (Cas de la Loire Moyenne). Ph.D. thesis, INP Grenoble, 283 pp.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, 1: a discussion of principles. J. Hydrol. 10, 282–290.

Nesmerak, I., Straskraba, M., 1985. Spectral analysis of the automatically recorded data from Slapy Reservoir, Czechoslovakia. Int. Revue Hydrobiol. 70 (1), 27–46.

Ragot, J., Darouach, M., Maquin, D., Bloch, G., 1990. Validation de Données et Diagnostic, Hermès, Paris, 593 pp.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K.I., 1997. Artificial nerural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28.

Rumelhart, D., Hinton, G., Williams, R., 1986. Learning Internal Representations by Error Propagation, Parallel Distributed Processing, 1, MIT Press.

Salas, J.D., Delleur, J.W., Yevjevich, V., 1980. Applied Modelling of Hydrologic Time Series, Water Resources Publications, Book Crafters Inc., 484 pp.

Stumm, W., Morgan, J., 1981. Aquatic Chemistry, Wiley Interscience, New York, 781 pp.

The MathWorks Inc., 1998. Neural Network Toolbox User's guide version 3, The MathWorks Inc., 296 pp.

Thiria, S, Lechevalier, Y., Gascuel, O., Canu, S., 1997. Statistique et Méthodes Neuronales, Dunod, Paris, 311 pp.

Whitehead, P.G., Neal, C., Seden-Perriton, S., Christophersen, N., 1986. A time series approach to modelling stream acidity. J. Hydrol. 85, 281–303.