Hydrology and Earth System Sciences, 7(5), *693–706* (2003) © EGU

Over-parameterisation, a major obstacle to the use of artificial neural networks in hydrology ?

Eric Gaume and Raphael Gosset

Ecole Nationale des Ponts et Chaussées, CEREVE, 6 et 8 avenue Blaise Pascal, 77455 Marne la Vallée Cedex 2, France

Email for corresponding author: gaume@cereve.enpc.fr

Abstract

Recently Feed-Forward Artificial Neural Networks (FNN) have been gaining popularity for stream flow forecasting. However, despite the promising results presented in recent papers, their use is questionable. In theory, their "universal approximator" property guarantees that, if a sufficient number of neurons is selected, good performance of the models for interpolation purposes can be achieved. But the choice of a more complex model does not ensure a better prediction. Models with many parameters have a high capacity to fit the noise and the particularities of the calibration dataset, at the cost of diminishing their generalisation capacity. In support of the principle of model parsimony, a model selection method based on the validation performance of the models, 'traditionally' used in the context of conceptual rainfall-runoff modelling, was adapted to the choice of a FFN structure. This method was applied to two different case studies : river flow prediction based on knowledge of upstream flows, and rainfall-runoff modelling. The predictive powers of the neural networks selected are compared to the results obtained with a linear model and a conceptual model (GR4j). In both case studies, the method leads to the selection of neural network structures with a limited number of neurons in the hidden layer (two or three). Moreover, the validation results of the selected FNN and of the linear model are very close. The conceptual model, specifically dedicated to rainfall-runoff modelling, appears to outperform the other two approaches. These conclusions, drawn on specific case studies using a particular evaluation method, add to the debate on the usefulness of Artificial Neural Networks in hydrology.

Keywords: forecasting; stream-flow; rainfall-runoff; Artificial Neural Networks

Introduction

Accurate stream flow predictions are required in many hydrological applications, e.g. water resources management, flood warning systems and optimisation of water treatment plants. Among the models for stream flow forecasting, feed-forward artificial neural networks (FNN) have attracted much interest recently (Karunanithi *et al*., 1994; Hsu *et al*., 1995; Minns and Hall, 1996; Dimopoulos *et al*., 1996; Shamseldin, 1997; Zealand *et al*., 1999; Sajikumar and Thandaveswara, 1999; Jain *et al*., 1999; Coulibaly *et al*., 2000). The main reason for this popularity is that, as is the case with other black-box models, they seem to be easy to use. Unlike conceptual modelling , no hydrological expertise is needed to derive functional relationships between the independent and the dependent variables; these are determined automatically in the calibration phase. Moreover, like polynomials and in contrast to linear models, the FNNs are universal approximators (Hornick *et al*., 1989); with a sufficient level of complexity, they can approximate any continuous function, to any degree of accuracy. Therefore, the FNNs appear to be good candidates for reproducing hydrological processes, which are well known to be non-linear. However, despite the promising results published recently, the use of FNNs for stream flow prediction is questionable.

If a sufficient number of neurons is selected, the "universal approximator" property of the FNN guarantees good performance of the model for interpolation purposes i.e. during calibration. This property does not mean that extrapolations using the model will be accurate. As noted by many environmental model users (Perrin *et al*., 2001; Jorgensen, 1988) but also by neural network model developers (Amari *et al*., 1997), increasing the complexity of a model does not ensure a better predicting power: beyond a certain level of complexity (i.e. a certain number of parameters), additional parameters may reduce the predicting power of the model. This difficulty can be addressed from two different points of view and solved in two ways. On the one hand, any decrease in performance of the model on a validation data set can be considered as indicative that the FNN has been 'over-fitted'. Beyond a certain number of parameters in an FNN, its calibration (i.e. training) must not be carried too far. The stopping criteria has, then, to be adjusted through a calibration-validation (i.e. cross-validation) procedure, as proposed by Maier and Dandy, (2000), Coulibaly *et al*. (1999) and Amari *et al*. (1997). The effectiveness of cross-validation methods is still discussed (Hassoun, 1995). Alternatively, the model can be considered as over-parameterised: i.e. the available data set is inadequate to calibrate the parameters of the proposed model. The complexity of the model required for satisfactory calibration results must be balanced against its subsequent performances when applied in extrapolation mode and the number of parameters may have to be adjusted.

The principle of parameter parsimony, well known to modellers (Box and Jenkins, 1970), leads to a plethora of model identification criteria (Akaike, 1973; Jakeman and Hornberger, 1993) and can also be applied when designing a FNN structure.

Neural network applications usually involve many parameters; in hydrology, this number may exceed one hundred (Coulibaly *et al*., 1999), in contrast to the usual hydrological conceptual models which depend on as few as two or three. Without a direct way of determining the optimal number of neurons (corresponding to an optimal number of parameters) of the FNN network, there is a risk of selecting 'over-parameterised' models: i.e. models so complex that they cannot be considered reliable when applied beyond the range of the calibration data-set. In the present study, selection of the model was based on the validation performance of the models tested, 'traditionally' used in the context of conceptual rainfall-runoff modelling, was adapted to detect evidence of 'over-parameterisation' and to determine the optimal number of neurons in a network. It was applied to two different case studies, one involving river flow prediction based on upstream mean daily flows (i.e. a routing problem), and the other involving rainfall-runoff prediction. The predictive power of the FNNs selected was compared to the results obtained with a linear model and with a conceptual model referred to as GR4j.

The selected forecasting models

FEED-FORWARD ARTIFICIAL NEURAL NETWORKS

Three-layer feed-forward neural networks used in this study (Fig. 1) have been widely used for hydrological modelling,

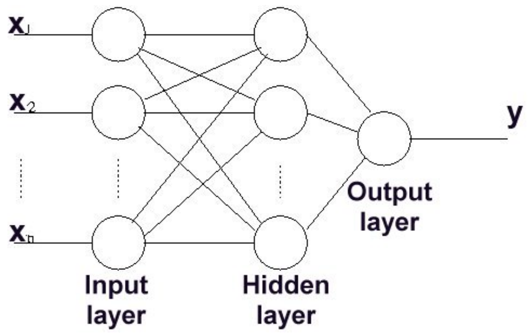

Fig. 1. *The n-input, single-output, three-layer feed-forward neural network*

because three layers are sufficient to generate arbitrarily complex output signals (Lippmann, 1987). Each of these three layers has a precise role. The first input or passive layer is dedicated to the capture of the external inputs and to their delivery to each of the neurons of the next layer. The second, also called the hidden layer, performs complex non-linear mapping of the input data, to simulate the relationship between inputs and outputs of the model. The outputs of the hidden layer are gathered and processed by the last or output layer, which delivers the final output of the network. The FNNs are only one example of the many possible structures of Artificial Neural Networks (ANNs).

A neuron is a processing unit with n inputs $(x_1, x_2, \ldots, x_n)$, and only one output y, with

$$y = f(x_1, x_2, \ldots, x_n) = A\left[\left(\sum_{i=1}^{n} w_i x_i\right) + b\right]. \qquad (1)$$

where the w_i are the weights of the neuron, b is the constant bias, and A is the so called activation or transfer function. In a FNN, the outputs of the neurons of a layer are the inputs of the neurons of the next layer. A sigmoid activation function (Fig. 2) was chosen for the neurons of the hidden layer, the identity function being used for the input and output neurons. Hereafter, the notation FNN(j,m,l) will define a three layer FNN structure with j neurons in the first layer, m neurons in the second layer, and l neurons in the last, with l=1 in the present case study.

As is the case with any empirical model, the FNNs need to be calibrated, an operation called "training" in the neural network terminology. The root mean square error (RMSE) is the chosen objective function or criterion. The calibration is performed through the Levenberg-Marquardt algorithm, which is remarkable for its accuracy and its high convergence speed, particularly when it is applied to calibrate FNNs of moderate size (Hagan and Menhaj, 1994). The Levenberg-Marquardt algorithm uses the following updating procedure at each optimisation step k:

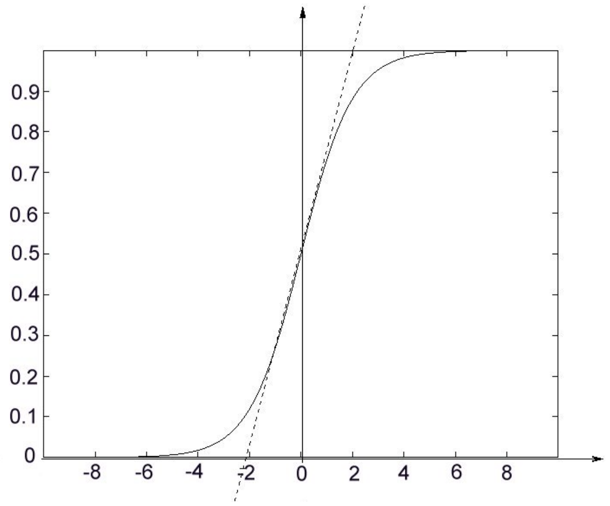

Fig. 2. *The sigmoid (or logistic) function - almost linear on the range [–1,1]*

$$\theta_{k+1} = \theta_k - \left[J_k^t J_k + \mu_k I\right] J_k^t E(\theta_k) \tag{2}$$

where θ_k is the vector of the parameters (weights and biases) of the network neurons at the step k of the optimisation, μ_k is a parameter, E is the vector of the network errors, with :

$$E(\theta_k) = X - F(\theta_k) \tag{3}$$

X is the vector of the measured data and $F(\theta_k)$ is the vector of the simulated data. J is the Jacobian matrix, which contains the first derivatives of the function F with respect to the weights and biases, i.e.

$$J_{k,i,j} = \left[\frac{\partial F_i}{\partial \theta_j}\right]_{\theta = \theta_k} . \tag{4}$$

When μ_k is large, Eqn. (2) becomes:

$$\theta_{k+1} = \theta_k - \frac{1}{\mu_k} J_k^t E(\theta_k) \tag{5}$$

or

$$\theta_{k+1} = \theta_k - \frac{1}{\mu_k} \nabla E(\theta_k) \tag{6}$$

which is equivalent to the well known back-propagation algorithm, with $1/\mu_k$ as the learning rate. When μ_k is small, Eqn. (2) can be approximated by :

$$\theta_{k+1} = \theta_k - \left[J_k^t J_k\right]^{-1} J_k^t E(\theta_k) \tag{7}$$

If the calibration criterion is the sum of squares, this is an approximation of the powerful Newton algorithm, with the assumption that :

$$H_k \approx J_k^t J_k \tag{8}$$

where H_k is the Hessian matrix, which contains the second derivatives of the function F with respect to the weights and biases. The value of μ_k is decreased after each successful step corresponding to a reduction in the error function, and is increased otherwise. The goal is to switch as quickly as possible from the back-propagation algorithm to Newton's method which is faster and more accurate near a minimum of the criterion function.

Because of its speed, the use of this algorithm allowed a large number of calibrations.

LINEAR MODEL

The linear model used in this study consists of a weighted sum of all its inputs $(x_1, x_2, ..., x_n)$, and adds a bias to compute the output y :

$$y = f(x_1, x_2, ..., x_n) = \left(\sum_{i=1}^{n} w_i x_i\right) + b \tag{9}$$

Equation 9 is equivalent to Eqn. (1) if any linear function is chosen as the activation function A. This implies that a neural network for which all the activation functions are linear is a linear model (thanks to the stability of the linear functions ensemble).

A CONCEPTUAL MODEL : GR4J

GR4j is a four-parameter lumped conceptual model. This reflects a drastic parsimony, limiting parameter interdependence which tends to plague the calibration of conceptual models (Edijatno *et al*., 1999). The flow chart of the model is shown in Fig. 3.

Let P and E denote precipitation and potential evapotranspiration, respectively in mm day^{-1}. E is a climatic average over several years, derived from 10-day data. The first transformation is done by an interception reservoir of zero-value capacity :

if $P \geq E$ then $P_n = P - E$ and $E_n = 0$ (10)

otherwise $P_n = 0$ and $E_n = E - P$ (11)

A moisture accounting reservoir with capacity X_2, executes the next operation. Let S be its current storage value. Depending on S, fluxes into P_s and out E_s of this reservoir occur when P_n and E_n are positive, respectively.

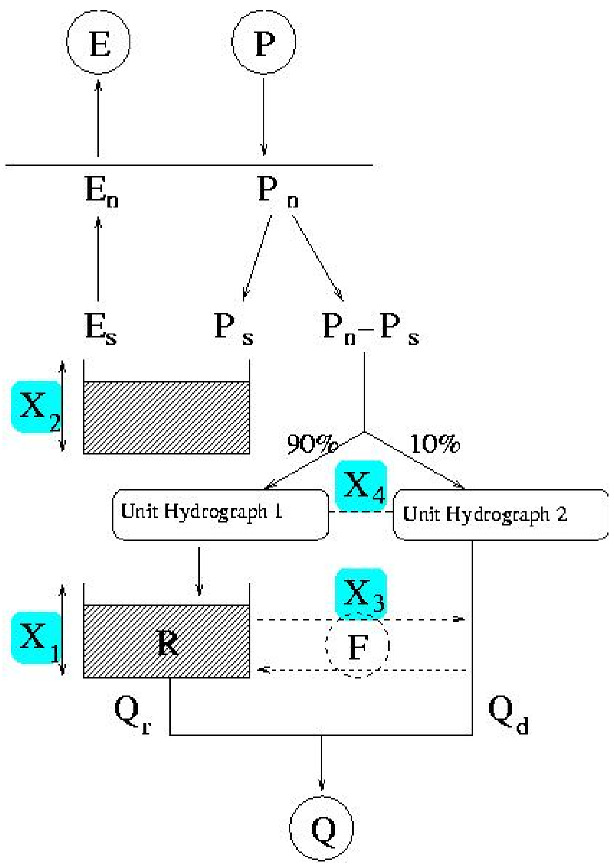

Fig. 3. *Schematic representation of the GR4j model*

$$\text{If} \quad E > P \quad \text{then} \quad P_s = 0 \quad \text{and} \quad E_s = E_n \frac{S}{E_n + \dfrac{(X_2)^2}{2X_2 - S}}. \tag{12}$$

$$\text{If} \quad E \leq P \quad \text{then} \quad E_s = 0 \quad \text{and} \quad P_s = \frac{P_n(X_2 - S)}{P_n + \dfrac{(X_2)^2}{X_2 + S}} \ldots \tag{13}$$

The new storage value of the reservoir is computed at the end of each time step.

$$S = S + P_s + E_s \tag{14}$$

The effective rainfall $(P_n - P_s)$ is then divided into two parts: a direct runoff (10%) and a delayed flow (90%). Both are transformed by a unit hydrograph, of which the time to peak is X_4. The shapes of the unit hydrographs are given by the following equations:

for the unit hydrograph 1

$$\text{If} \ldots 0 < j \leq X_4, \quad SH1(j) = \left(\frac{j}{X_4}\right)^3 \tag{15}$$

$$\text{If} \ldots j > X_4, \quad SH1(j) = 1 \tag{16}$$

$$UH1(j) = SH1(j) - SH1(j-1) \tag{17}$$

for the unit hydrograph 2

$$\text{if} \ldots 0 < j \leq X_4, \quad SH2(j) = \frac{1}{2}\left(\frac{j}{X_4}\right)^3 \tag{18}$$

$$\text{if} \ldots X_4 < j \leq 2X_4, \quad SH2(j) = 1 - \frac{1}{2}\left(2 - \frac{j}{X_4}\right)^3 \tag{19}$$

$$\text{if} \ldots j > 2X_4, \quad SH2(j) = 1 \tag{20}$$

$$UH2(j) = SH2(j) - SH2(j-1) \tag{21}$$

where j is the time step index and *UH1(j)* and *UH2(j)* are the values of the unit hydrographs 1 and 2 for the j^{th} time step.

Exchanges with groundwater are also simulated. The exchange rate F (mm day^{-1}) depends on the value of the storage value R of the second reservoir and on a parameter X_3.

$$F = X_3\left(\frac{R}{X_1}\right)^4 \tag{22}$$

The delayed flow is routed through a second reservoir, from which the outflow Q_r (*mm/day*) depends on its current storage content R and on the parameter X_1.

$$R_* = \max(0, R + F + UH1) \tag{23}$$

$$Q_r = R_* - (R_*^{-4} + X_1^{-4})^{-1/4} \tag{24}$$

$$R = R_* - Q_r \tag{25}$$

The final output of the model is the sum of the direct and delayed components of the flow. The four parameters that have to be calibrated are X_1, X_2, X_3 and X_4 (Rakem, 1999).

Unlike the FNN or linear models, conceptual rainfall-runoff models do not generally incorporate the last measured discharges. Their mean error value can be much larger than the day-to-day discharge variations. Therefore, they must be adapted for use as stream-flow prediction tools: the rainfall-runoff model can be combined with an auto-regressive (AR) model of its forecast errors or its parameter values updated at each time step to take into account the most recent forecast errors.

If properly implemented, updating methods seem to perform slightly better than the combination with an AR model (Yang and Michel, 2000). Nevertheless, for the sake of clarity, the combination with an AR model has been selected in the present case study. In both cases, the adaptation involves additional parameters: the parameters of the AR model (generally the one or two previous time steps are considered), or the number of previous time steps considered in the updating procedure and the relative weight of the optimisation criteria computed on the whole available series and on the previous time steps which have to be adjusted for each case study. The version of GR4j tested herein comprises six parameters; the four original parameters and the two parameters of the AR model with which it is combined.

Methodology

ON FNN MODELS IDENTIFICATION AND PARAMETER CALIBRATION

The initial values of the parameters are chosen randomly during calibration. Therefore, depending on the initial parameter values selected and on the form of the calibration criterion (objective function), several calibration trials of one FNN structure can lead to different sets of parameters.

The minimum values of the objective function (i.e. the local optima) obtained during the calibration trials on measurements appear to be very close to one another and the best calibration parameter set obtained is not necessarily the best validation set. Therefore, it is not really possible to choose sensibly between the various minima of the objective function that occur whenever a FNN has more than one hidden neuron.

Most previous papers present the validation performance obtained with one of the possible FNN structures identified through calibration — generally after multiple random initialisation of the network parameters — (Zealand *et al*., 1999; Imrie *et al*., 2000) or through a cross-validation approach (Coulibaly *et al*., 2000). In both cases, the choice of the parameter values of the FNN is partly due to chance, and the performance of the FNNs may appear case-dependent or variable (Shamseldin, 1997). It is the opinion of the authors that, with such an approach, the comparison of the performance of various types of models and neural networks is difficult, if not impossible. Therefore, to reduce the possible effect of chance in the estimation of the efficiency of the FNNs, several plausible parameter sets have been considered in the present study, rather than a single 'best' set.

THE CALIBRATION-VALIDATION PROCEDURE

A 'constructive' approach (Kwok and Yeung, 1997) is used to determine the optimal structure (i.e. the optimal number of neurons in the hidden layer) of the FNN. The validation performance of the FNN is compared as the number of neurons in the hidden layer is increased systematically. In view of the need to consider several equally plausible optimum sets of parameters, 20 different calibration trials of each tested network structure were performed, followed by the 20 corresponding validations. Twenty trials represent a compromise between a good representation of the diversity of plausible parameter sets and an acceptable computing time. The twenty root mean square error (RMSE) values, or rather the performance criteria based on the RMSE values , are represented on two box-and-whisker plots (see Figs. 4, 7, 8, 11 and 13), one for the calibration errors and one for the validation errors.

The line in the boxes is the median value. The boxes end at the data quartiles, and the whiskers extend to the most extreme data points which are no more than 1.5 times the interquartile range from the box. Outliers are represented as circles. In the interpretation of the results, reference will be made to the median validation performances of the FNN; these correspond to the median prediction performance that can be expected if the FNN is chosen on the basis of one calibration trial. It could be that, by chance, a selected FNN leads to a better validation result. But it could just as well lead to a worse result.

While it is expected that the median calibration criterion of the FNN will decrease as the number of neurons is

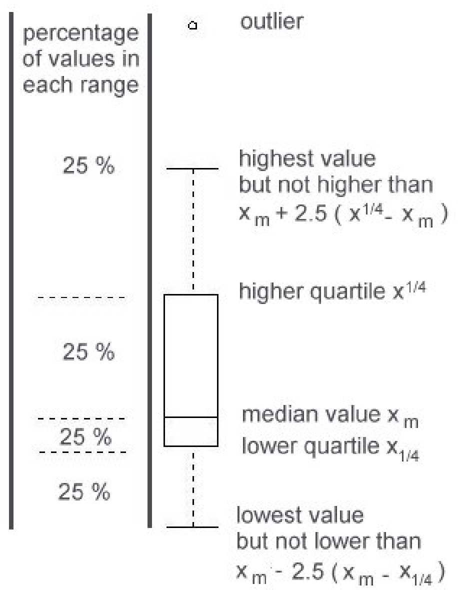

Fig. 4. *Meaning of a box-and-whisker plot*

increased, after a certain point, the addition of new neurons will raise the median criterion of the FNN on the validation data-set. The optimum FNN structure adopted is that having the lowest median validation criterion.

Applications

MEAN DAILY RIVER FLOW PREDICTIONS BASED ON UPSTREAM MEASURED FLOWS

To predict the mean daily flow at the Noisiel station, the data used were ten years (1989–1998) of measurements of daily mean flows at three gauging stations on the river Marne watershed in France: Noisiel and Chalons-sur-Marne on the river Marne, and Pommeuse on the Grand Morin. The Grand Morin is a tributary stream of the river Marne (Fig. 5). The 1989–1994 period was used for calibration and the 1995–1998 period for the validation of the model. The two data sets have very close characteristics (Fig. 6 and Table 1). The flood propagation time between Chalons-sur-Marne and Noisiel is about three days for medium floods and five days for larger events (Gaume and Tassin, 1999). Hence, the measurements of the five previous days at Chalons-sur-Marne were used as input data for the FNNs and for the linear model. For simplicity, the last five measurements at the two other gauging stations were used, giving a total amount of 15 inputs for both types of models. In this configuration, the number of parameters of each neural network is 16 times the number of neurons (15 input variables and one constant per neuron).

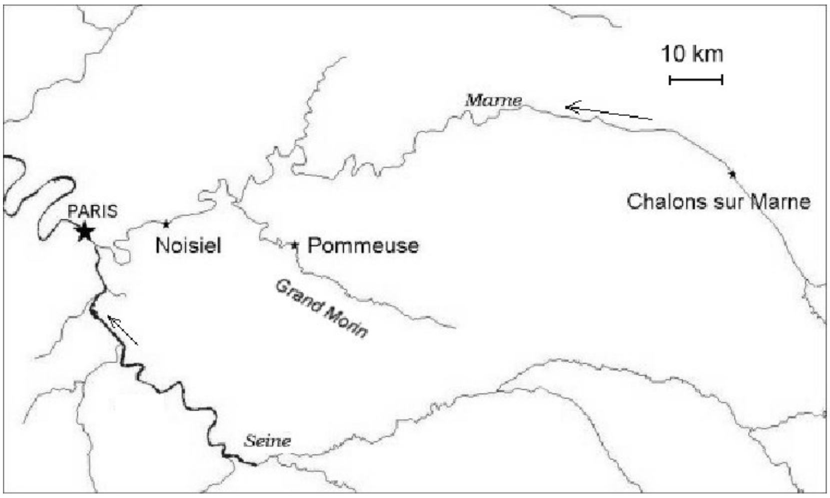

Fig. 5. *The Marne river and the three gauging stations*

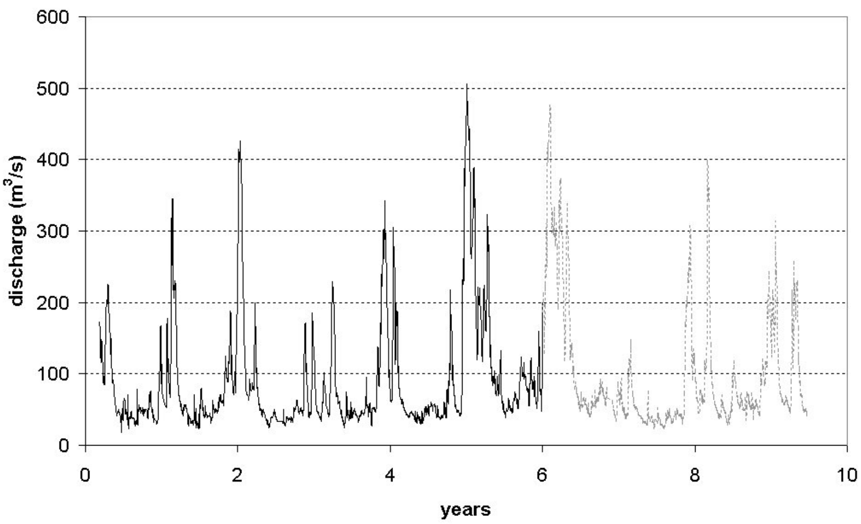

Fig. 6. *Discharges measured at the Noisiel gauging station: calibration (black line) and validation (grey line) data sets.*

Table 1. Descriptive statistics of the calibration and validation data sets at the Noisiel gauging station.

	Calibration set	*Validation set*
Mean daily flow	90.4 $m^3 s^{-1}$	100.2 $m^3 s^{-1}$
Median daily flow	56.2 $m^3 s^{-1}$	51.7 $m^3 s^{-1}$
Minimum daily flow	17.8 $m^3 s^{-1}$	23.1 $m^3 s^{-1}$
Maximum daily flow	506 $m^3 s^{-1}$	477 $m^3 s^{-1}$
% days over 100 $m^3 s^{-1}$	24	28.5
% days over 50 $m^3 s^{-1}$	59	68

Non-informative or redundant inputs may reduce the efficiency of neural networks (Shamseldin, 1997) and some authors suggest that the number of neurons in the passive layer should also be optimised (Abrahart *et al*., 1999). The optimal selection of the number of external inputs is beyond the range of the present paper. Nevertheless, to indicate the possible influence of the number of input data on the results obtained, a series of tests was affected with a smaller number of neurons in the input layer. Six input data were selected : the measured mean daily flow at the Noisiel station during the two preceding days, the mean discharge measured at the Pommeuse station the day before, and the discharge measured at Chalons between three and five days before. In view of the estimated flood propagation times between the three gauging stations (three to five days between Chalons and Noisiel and about one day from Pommeuse to Noisiel), this input data set represents a bottom set for a flow prediction model at Noisiel based on the data of the three gauging stations. The neural networks tested in this second trial have a number of parameters equal to 7 times the number of their neurons in the hidden layer.

The proposed calibration-validation method is applied to the Marne example. Figures 7 and 8 summarise the results. The mean square error is not plotted directly on these and the following figures but a prediction efficiency criterion, the value of which is less dependent on the case study and on the set of measurements used. The selected criterion has the following form :

$$CRIT = 1 - \frac{\sum_i \left(Q_i - Q_i^*\right)^2}{\sum_i \left(Q_i - Q_{i-1}\right)^2} \qquad (26)$$

Where Q_i is the discharge measured at Noisiel at the i^{th} time step, Q_i^* is the predicted discharge for the same time step. An efficiency criterion value of 1 corresponds to a perfect prediction. A negative value of the criterion indicates that the mean square error of the model is higher than the variance of the day to day fluctuation in discharge: i.e. the

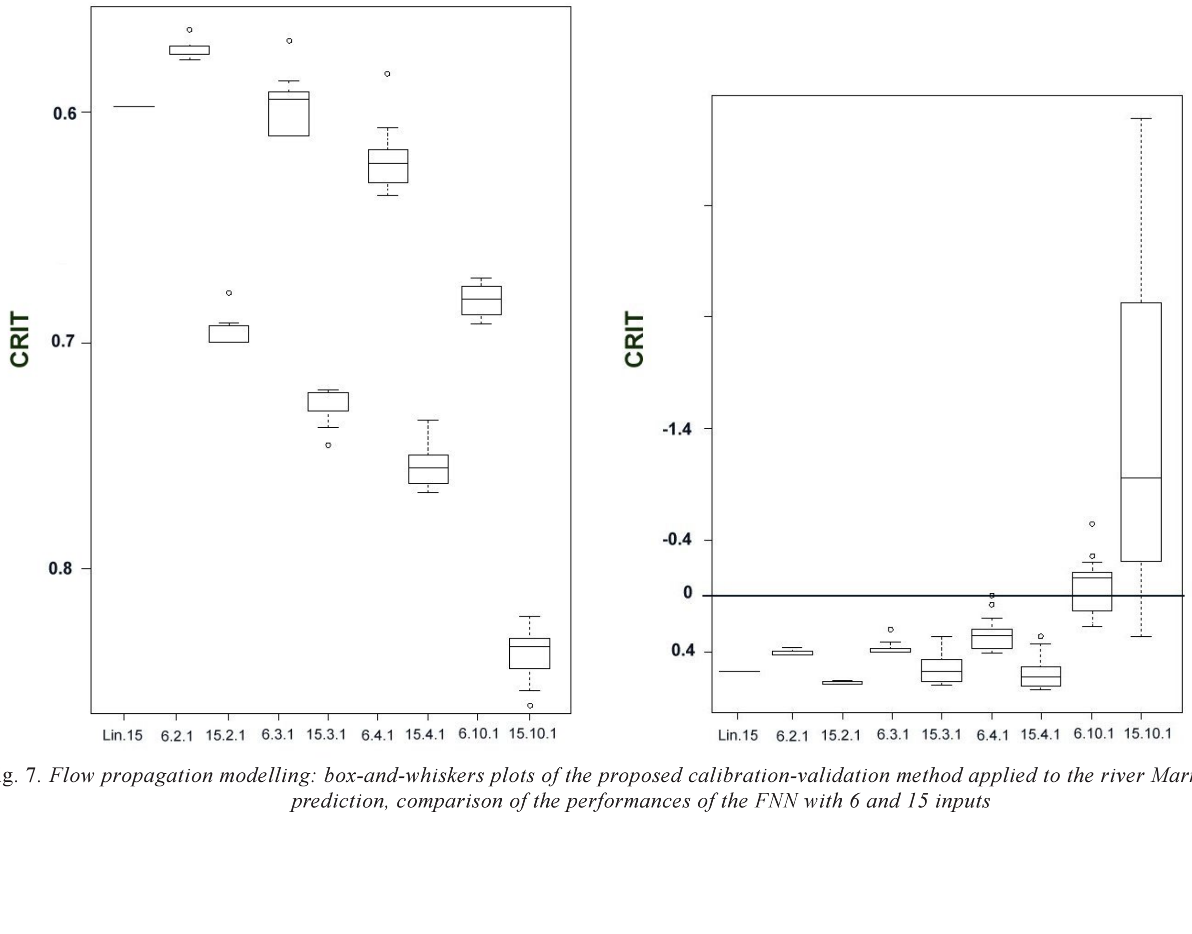

Fig. 7. *Flow propagation modelling: box-and-whiskers plots of the proposed calibration-validation method applied to the river Marne flow prediction, comparison of the performances of the FNN with 6 and 15 inputs*

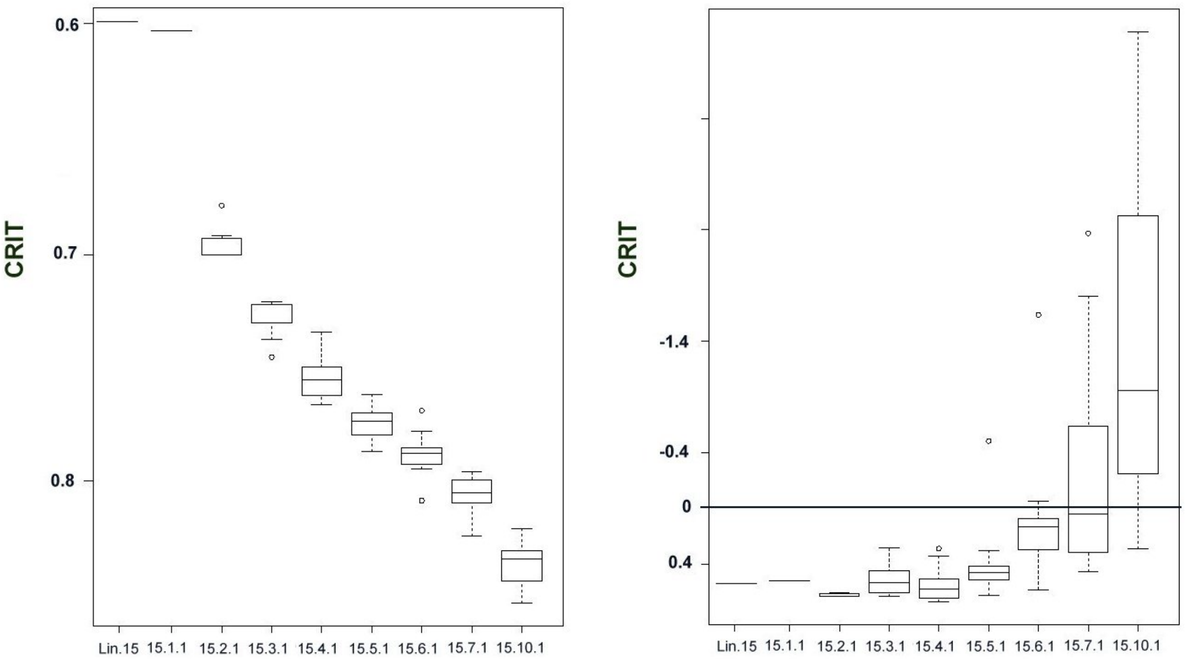

Fig. 8. *Flow propagation modelling: box-and-whiskers plots of the proposed calibration-validation method applied to the river Marne flow prediction, results obtained for the FNN with 15 inputs*

simple prediction model repeating the previous observations would then have a higher prediction efficiency.

The prediction efficiency in calibration (Figs. 7 and 8 (a)) tends to increase with the number of neurons but the calibration becomes more and more difficult. This is reflected in the growing size of the boxes, which denotes a scattering of the calibration trajectories, resulting in different values of the optimum criterion.

The validation criterion (Figs. 7 and 8 (b)) behaves differently. After a small increase corresponding to FNN(15,2,1), the median criterion value begins to decrease. The very high error values found for FNN(15,3,1) and FNN(15,4,1) are signs of over-parameterisation. Consequently, FNN(15,2,1) presents the smallest median error (i.e. the best median efficiency criterion value) on the validation data-set.

The performances of the FNN with 15 inputs are higher than those ones with six inputs during calibration but also during validation. The shapes of the box-plots are similar. No clear evidence of a perturbing effect of eventually superfluous input data in the FNN structure with 15 inputs appears on the basis of this comparison. The lower validation performances of the FNN with 6 inputs indicate that the linear model as well as the FNN with 15 inputs take advantage of input data which could appear of secondary importance on a first analysis.

The criterion obtained with the linear model on the validation period (0.5 corresponding to a mean square error of 9.2 $m^3 s^{-1}$) is close to the median value found for the FNN selected (0.56 corresponding to a mean square error of 8.5 $m^3 s^{-1}$).

The analysis of the results of both models (Fig. 9) shows that the predictions during (b) rising and (c) falling limbs of flood hydrographs are very close. For other events, such as (a) floods and (b) low water periods, the FNN approach is a marginal improvement on the predictions made by the linear model. This suggests that the smaller value of the mean error found for the FNN approach is due to better predictions made during low water periods, which outnumber those of flood events.

The FNN approach and the linear model exhibit very close behaviours : they show the same inability to anticipate the increasing flow (Fig. 9 (a)) and their predicted hydrographs are similar, even if the FNN predictions are usually closer to the measurements than the corresponding linear predictions.

The near-linearity of the modelled process (mainly small hydrograph shifts and tributary flow additions) can explain why the results of the FNN are only marginally better than those of the linear model. This led the present authors to study a rainfall-runoff process, well known for its non-linearity; this should exploit the potential of the FNNs to adapt to such non-linear relationships.

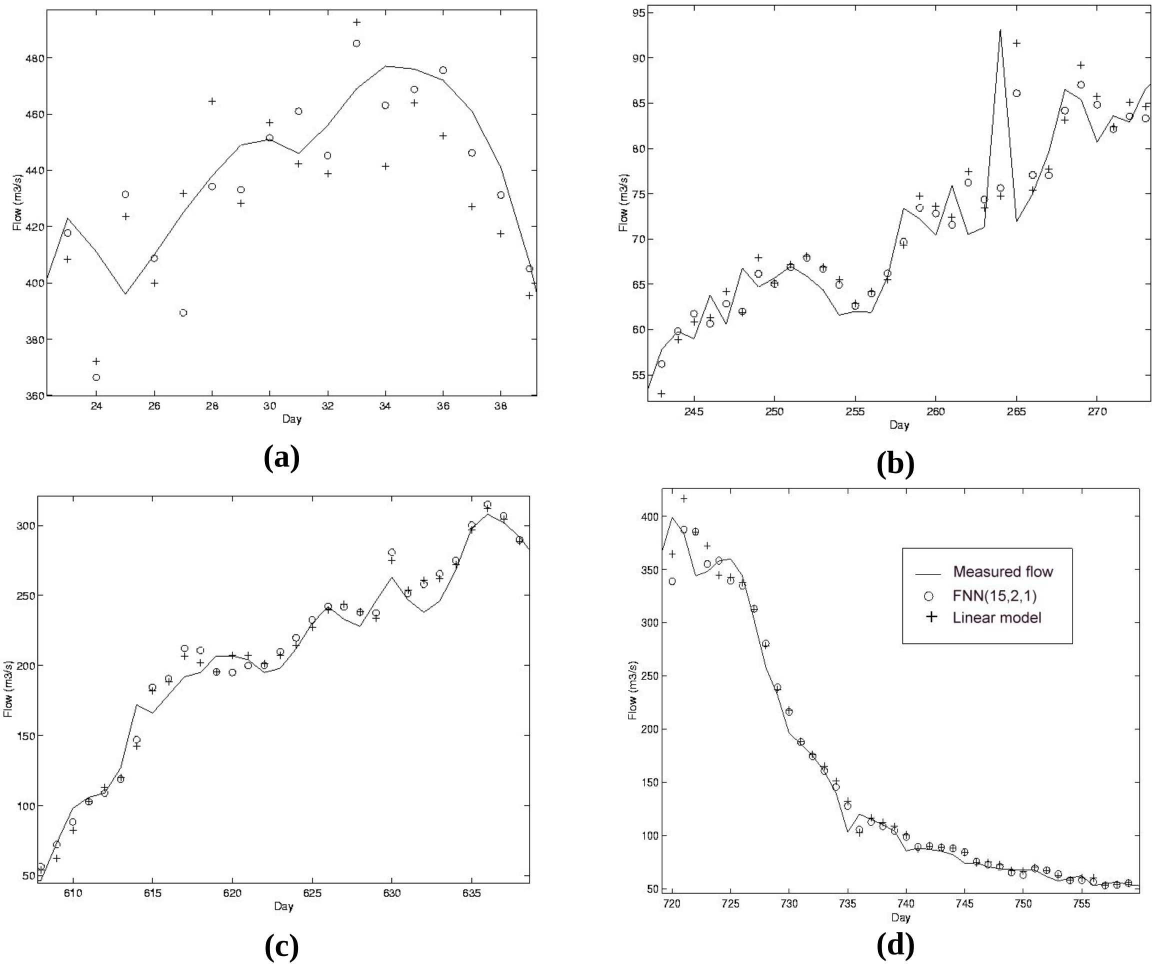

Fig. 9. *Flow propagation modelling : forecasted and observed hydrographs for (a) a flood event (21 Jan. 1995), (b) a low water period (30 Aug. 1995), (c) a rising (8 Nov. 1996), and (d) a decreasing (4 Mar. 1997) limb of flood hydrographs of the river Marne.*

MEAN DAILY RIVER FLOW PREDICTIONS BASED ON METEOROLOGICAL DATA (PRECIPITATION AND EVAPOTRANSPIRATION)

The rainfall-runoff process is often described as being highly non-linear, particularly in small catchment areas (Bras, 1990). Rainfall and flow measurements from the river Le Sauzay in France — having an upstream catchment area of 81 km^2 — have been used. The data consist of the daily mean flows at the outlet of Le Sauzay's catchment area, mean 10-day potential evapotranspiration computed each day and daily precipitation from a nearby meteorological station. The 1986-1990 period was used for calibration and the 1991-1995 period was used for validation of the models. The two data sets appear comparable (Fig.10 and Table 2). The data of the previous five days were used to predict the flow on the following day. This amount of input data represents a compromise. The calibrated linear model takes into account only the data of the two preceding days: the parameters differ significantly from zero only for the two preceding days. In this configuration, the number of parameters of each neural network is 16 times the number of neurons (15 input variables and one constant per neuron). As previously, further tests were conducted with a FNN structure with a reduced number of input neurons to assess the possible perturbing effect of eventually superfluous input data. The selected input data were the measured river flows for the two preceding days, the daily precipitation of the three preceding days and the mean 10-day potential evapotranspiration computed for the previous day. The number of parameters of the neural networks is then seven times the number of neurons.

The application of the calibration-validation procedure exhibits the same characteristics as in the previous example (Fig. 11). Contrary to expectation, evidence of over-parameterisation appears as quickly as previously as the number of parameters is increased. FNN(15,2,1) is selected

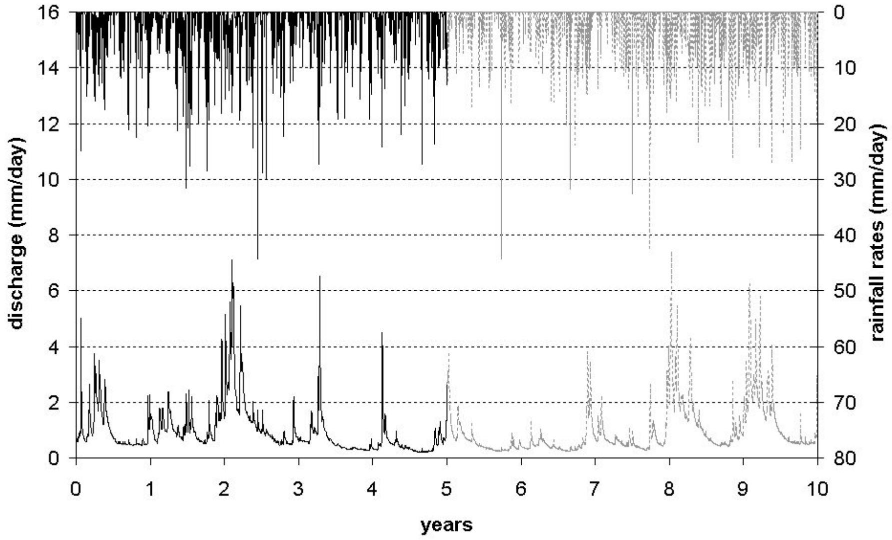

Fig. 10. *Discharges and rainfall rates measured on the Sauzay catchment: calibration (black lines) and validation (grey lines) data sets.*

Table 2. Descriptive satistics of the calibration and validation data sets available for the Sauzay catchment.

	Calibration set	*Validation set*
Mean annual rainfall	792 mm	798 mm
Mean annual flow	356 mm	348 mm
Mean annual number of rainfall days	188	189
Maximum rainfall rate	44 mm/j	44 mm/j
Maximum mean daily discharge	7.1 mm/j	7.4 mm/j
Minimum mean daily discharge	0.19 mm/j	0.23 mm/j

here also because it presents the smallest median error on the validation dataset. Larger FNN structures show evident signs of over-parameterisation.

The comparison of the results obtained with the linear model and the FNN with 15 and 6 inputs show more or less the same characteristics as before. The best performances during the validation are obtained with the FNN(15,2,1). The relative validation performances of the FNN models 15 and 6 inputs appear erratic: no systematic tendency indicates a noticeable 'perturbing' effect of the superfluous data possibly included in the input data-set of the FNN with 15 inputs.

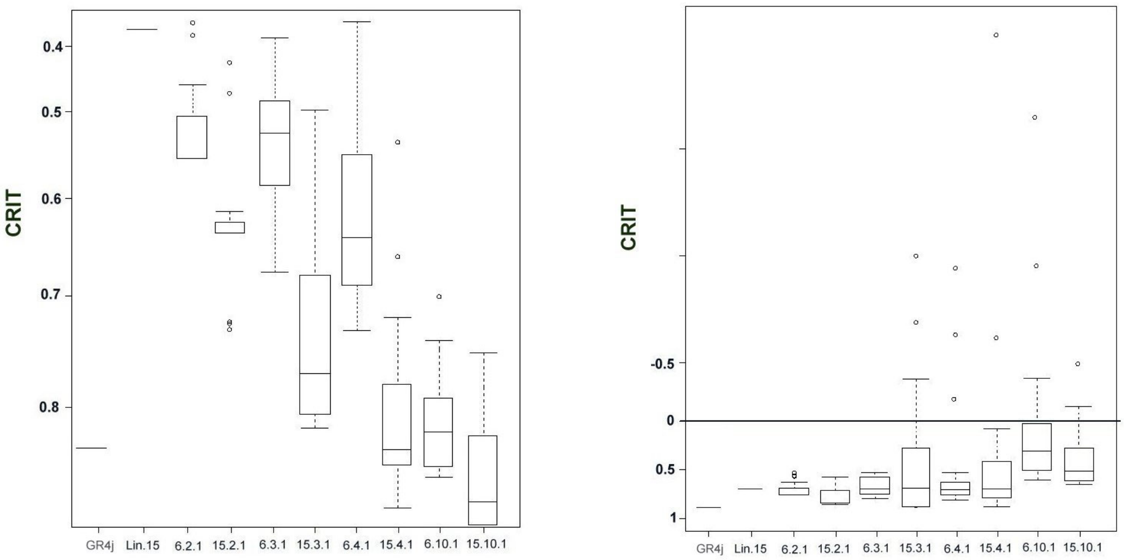

Fig. 11. *Rainfall-runoff modelling : box-and-whiskers plots of the calibration-validation method applied to the river Le Sauzay flow prediction, comparison of the performances of the FNN with 6 and 15 inputs and of the GR4j rainfall-runoff model*

Moreover, the difference between the smallest validation root mean square error found for the FNN and the error of the linear model is relatively small (0.14 $m^3 s^{-1}$ against 0.16 $m^3 s^{-1}$ respectively). For comparison, the GR4j model was also used with the same rainfall-runoff data-set; the root mean square error on the validation dataset dropped to about 0.09 $m^3 s^{-1}$.
Again, it is interesting to analyse in which cases each model outperforms the others. The hydrographs (Fig. 12) illustrate these performance variations.

It appears that the predictions made during flood events (d), rising (b) and decreasing (c) limbs of flood hydrographs are approximately equivalent for the linear model and the selected FNN structure. This confirms that, in esssence, the FNN approach is more accurate than the linear model during low water periods (a).

GR4j also performs particularly well on the low water periods (Fig. 12(a)). Performances in other situations are difficult to evaluate, as the three models seem to be equivalent elsewhere but GR4j is the only model capable of anticipating the rising of hydrographs, which is an important feature for flood forecasting issues.

Discussion

The results obtained in the two case studies are somewhat disappointing as far as the FNNs are concerned. When compared to the linear model, the predictions are only marginally improved and, in the second case study, both simulation approaches are outperformed by the conceptual model. On the one hand, the noise present in the hydrological measurements can explain the difficulties encountered during the calibration of the FNN models and their poorer validation results. The first part of the discussion will be devoted to the influence of noise on the efficiency of the FNNs. On the other hand, the comparison of the prediction performance the FNNs and of the conceptual model leads to another question which will be discussed in the second part: are black-box models, like the FNNs, really suitable for forecasting issues?

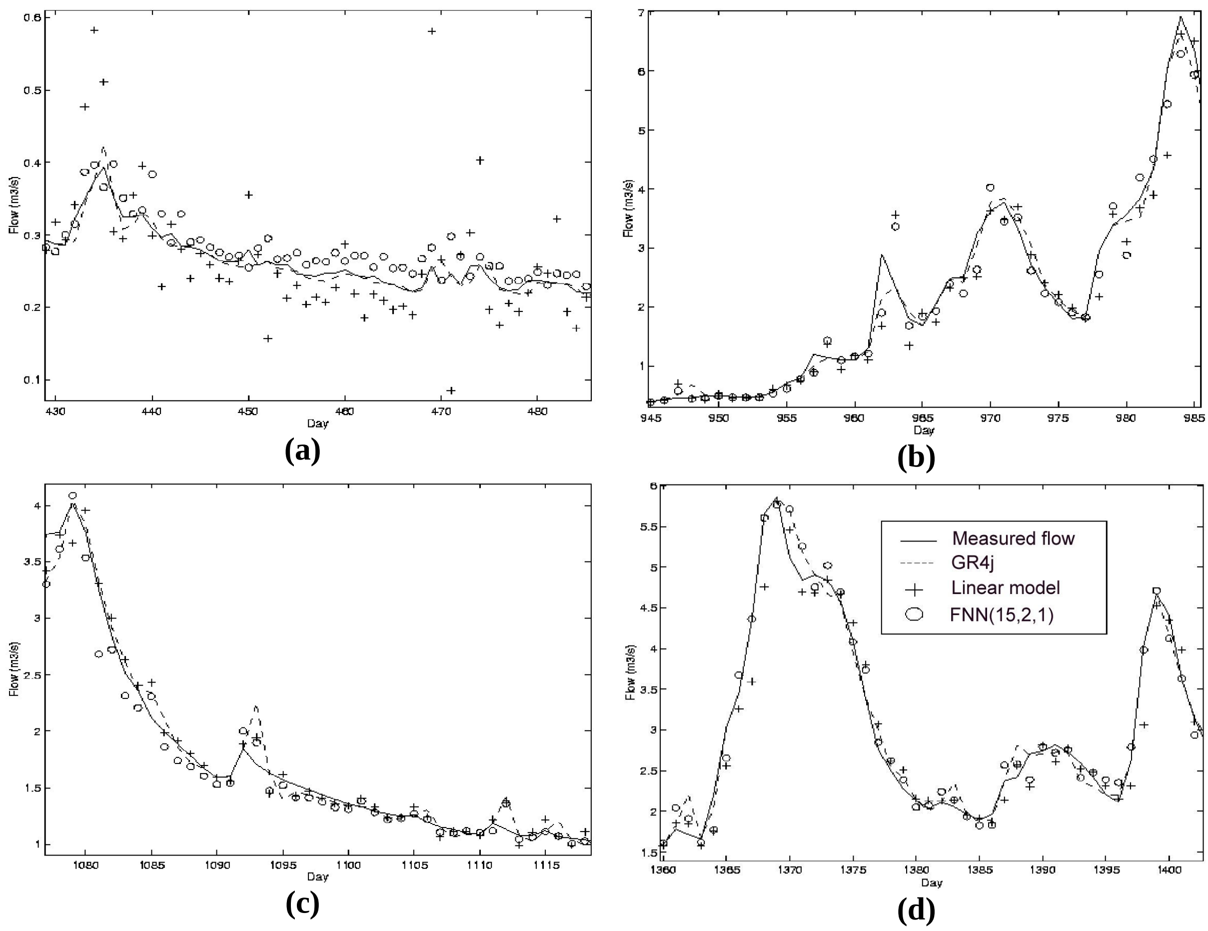

Fig. 12. *Rainfall-runoff modelling : forecasted and observed hydrographs for (a) a low water period (17 June 1992), (b) a rising (22 Nov. 1993), (c) a decreasing (2 Apr. 1997) limb of flood hydrographs, and (d) a flood (10 Jan. 1995) of the river Le Sauzay.*

Noise is a recurrent problem in hydrological time series. In particular, flow measurements obtained by measuring stage (water-surface elevation) are subject to many errors. Approximations made in determining the stage-discharge relationship, or changes in the cross section of the stream after a flood event, for example, may affect the accuracy of the measurements. In fact, the definition of noise also includes the processes which are not taken into account in the model. Noise levels in hydrological data usually range between 5 and 15% (Sivakumar *et al*., 1999).

To test the possible effect of noise on the prediction efficiency of the FNNs, the same calibration-validation method was used to select an optimal FNN for predicting synthetic flow series simulated using the GR4j model on Le Sauzay catchment. The data on which the FNNs are calibrated are, in this case, produced by a purely deterministic non-linear process. A Gaussian noise was added to the input data and to the simulated dataset, with various relative standard deviations (0, 5, 10 and 15%). The results are summarized in Fig. 13. The criterion values obtained using the GR4j model for the calibration and validation procedure are also shown.

Clearly, the noise level has an effect, but does not in itself explain the shape of the validation box-plots: the optimum FNN model remains a simple one even with the 'perfect' or 'no noise' data-set. The 'no noise' example indicates that calibration results are, in fact very good with the FNN with two neurons in the hidden layer (root mean square error of 0.02 $m^3\ s^{-1}$); they are improved only slightly by the addition of new neurons. A perfect adjustment would certainly require very many neurons. The 'universal approximator' property of the FNN ensures that it is theoretically possible to find a network capable of reproducing the simulated process with any desired precision. It will, nevertheless, be difficult if

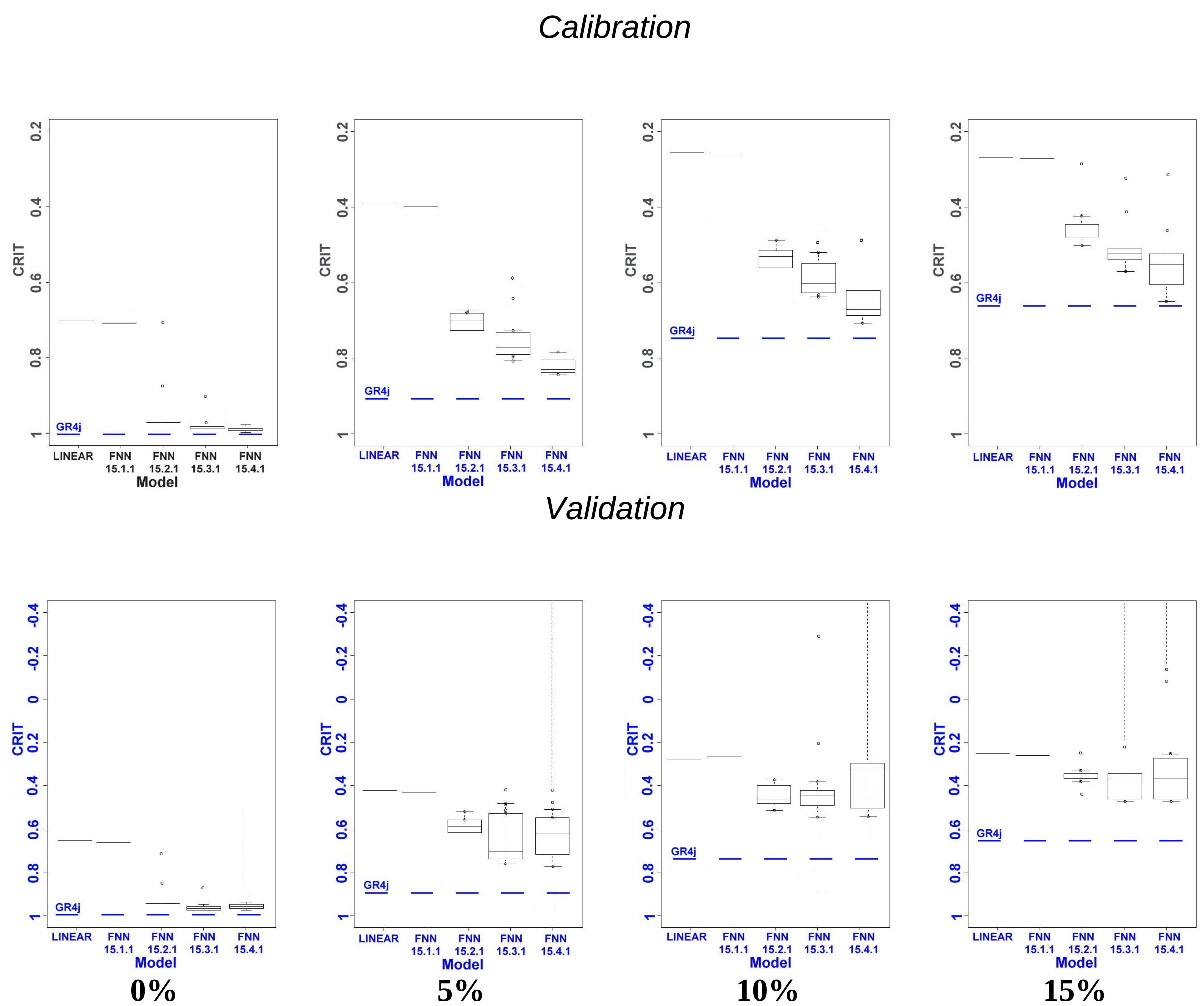

Fig. 13. *Evolution of the error values on the calibration and the validation datasets with different levels of noise perturbation. Calibration and validation on series of discharge produced by the non-linear GR4j rainfall-runoff model. The dotted lines indicate the criterion values obtained with the GR4j model*

not quite impossible to find this network through calibration in some cases.

Figure 13 also indicates that, as the noise level increases, it becomes more difficult to improve the prediction performance of the FNNs during calibration by increasing the number of neurons; indeed the distance between the validation results of the FNNs and of the GR4j model, by construction suited to the modelled process, increases progressively. As in the rainfall-runoff application on the Sauzay catchment (Fig. 11), the calibration and validation criteria of the GR4j model are comparable while the median calibration and validation criterion values of the FNNs move apart as the number of neurons in the hidden layer is increased. The calibration improvements obtained with more complex FNN structures may reflect a better fit to the noise rather than a more precise simulation of the underlying process. Black-box models, like the FNN approach, cannot distinguish between noise and process. Therefore, increasing the number of parameters may be inefficient as far as the predicting power of the model is concerned when the measurements used are subject to noise.

Neural network simulation consists of a decomposition of a signal in a base of given functions (the sum of activation functions in the present case). The calibration difficulties, especially when the data are affected by noise, reveal that the logistic function is not particularly well suited for simulating hydrological processes. The search for efficient activation functions was beyond the scope of this paper, but some elements of discussion are published elsewhere (Imrie *et al*., 2000). Nevertheless, the authors consider that testing different types of activation functions selected *a priori* with no insight into the dynamics of the process being studied may not improve the forecasting performance of the FNNs. If the functions used for interpolating measurements are not relevant to the process being observed (which will often be the case for black-box models in which the function types are selected *a priori*), the model can lead to 'absurd' results when used for predictions. Hence, it is easier to trust predictions made with conceptual models incorporating functions chosen on the basis of hydrological expertise than black-box models such as FNNs.

Conclusion

The proposed calibration-validation procedure leads to the choice of very simple FNN structures with two or three hidden neurons in both case studies. If the number of parameters is optimised, the FNN models appear to be better forecasting tools than linear models. However, the benefit is limited. The FNN approach improved the predictions marginally, essentially during low flow periods. This conclusion differs from the findings of previous reported studies (Sajikumar and Thandaveswara, 1999; Coulibaly *et al*., 2000). The various lengths of the data-sets used may possibly explain this disagreement. In the present study, only five year datasets were used to calibrate the models. While short, this corresponds to many operational situations with which engineers can be confronted. Too short a calibration set may not provide the variety of hydrological situations necessary for the model to 'learn' properly.

Nevertheless, these conclusions agree with previous reports (Jakeman and Hornberger, 1993; Gaume *et al*., 1998), that the data available in hydrology can support only the development of models with limited complexity.

A conceptual model, with only six parameters, specifically dedicated to rainfall-runoff simulation, outperformed the FNN and linear simulation approaches. Cross-validation methods, mentioned in the introduction, were not tested. A different implementation of the FNN may improve their forecasting performances in these two case studies. The data used in this paper can be downloaded at the following address for additional trials : http:\\www.enpc.fr\cereve\HomePages\gaume\hess2003.zip.

Nevertheless, in the opinion of the authors, the result presented here is not surprising. Models based on classes of functions, like FNN models, can be efficient only if those functions are suitable for the process to be simulated, which is generally not the case. Therefore, it is easier to trust predictions made with conceptual models including functions chosen on the basis of hydrological expertise than black-box models like FNNs. The quest for a universal model requiring no hydrological expertise may be hopeless.

Acknowledgments

The present work benefited from the financial support of the Syndicat des Eau d'Ile de France. The authors thank Kieran M. O'Connor (National University of Ireland, Galway), Mickael Cailas (University of Illinois at Chicago), Bruno Tassin (CEREVE) and the reviewers for their very helpful comments.

References

Abrahart, R.J., See, L. and Kneale, P.E., 1999. Using pruning algorithms and genetic algorithms to optimise network architectures and forecasting inputs in a neural network rainfall-runoff model. *J. Hydroinform.*, **1**, 103–114.

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: *Second international conference on Information Theory, Budapest*, B. Petrov and F. Csaki (Eds.) 267–281.

Amari, S-I., Murata, N., Muller, K-R., Finke, M. and Yang, H.H., 1997. Asymptotic statistical theory of overtraining and cross-validation, *IEEE Trans. Neural Networks*, **8**, 985–996.
Box, G.E.P. and Jenkins, G.M., 1970. *Time series analysis, forecasting and control*. Holden Day, San Francisco, USA.
Bras, R.L., 1990. *Hydrology - an introduction to hydrologic science*. Addison-Wesley.
Coulibaly, P., Anctil, F. and Bobée, B., 1999. Prévision hydrologique par réseaux de neurones artificiels: état de l'art, *Can. J. Civil Eng.*, **26**, 293–304.
Coulibaly, P., Anctil, F., and Bobée, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *J. Hydrol*., **230**, 244–257.
Dimopoulos, I., Lek, S. and Lauga, J., 1996. Modélisation de la relation pluie-débit par les réseaux connexionnistes et le filtre de Kalman. *Hydrolog. Sci. J.*, **41**, 179–193.
Edijatno, M., Nascimento, N.O., Yang, X., Makhlouf, Z. and Michel, C., 1999. GR3j: a daily watershed model with three free parameters, *Hydrolog. Sci. J.*, **44**, 263–277.
Gaume, E. and Tassin, B., 1999. *Prévision du débit et de la température de l'eau aux prises d'eau du syndicat des eaux d'Ile de France, prévision des débits de la Marne a Noisiel.* Technical report, Ecole Nationale des Ponts et Chaussées, Champs-sur-Marne, France.
Gaume, E., Villeneuve, J.-P. and Desbordes, M., 1998. Uncertainty assessment and analysis of the calibrated parameter values of an urban storm water quality model. *J. Hydrol*., **210**, 38–50.
Hagan, M.T. and Menhaj, M., 1994. Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Networks*, **5**, 989–993.
Hassoun, M.H., 1995. *Fundamentals of artificial neural networks*. MIT Press, Cambridge, USA.
Hornick, K., Stinchcombe, M. and White, H., 1989. Multilayer feed forward networks are universal approximators. *Neural Networks*, **2**, 359–366.
Hsu, K.L., Gupta, H.V. and Sorooshian, S., 1995. Artificial neural network modeling of the rainfall-runoff process, *Water Resour. Res.*, **31**, 2517–2530.
Imrie, C.E., Durucan, S. and Korre, A., 2000. River flow prediction using artificial neural networks : generalisation beyond the calibration range. *J. Hydrol*., **233**, 138–153.
Jain, S.K., Das, A. and Srivastava, D.K., 1999. Application of ANN for reservoir inflow prediction and operation. *J. Water Resour. Plan. Man. ASCE*, **125**, 263–271.
Jakeman, A.J. and Hornberger, G.M., 1993. How much complexity warranted in a rainfall-runoff model? *Water Resour. Res.*, **29**, 2637–2649.
Jorgensen, S.E., 1988. *Fundamentals of ecological modeling : developments in environmental modeling*. Elsevier, Amsterdam, The Netherlands.
Karunanithi, N., Grenney, W.J., Whitley, D. and Bovee, K., 1994. Neural network for river flow prediction. *J. Comput.Civil Eng.*, **8**, 201–220.
Kwok, T.-Y. and Yeung, D.-Y, 1997. Constructive algorithms for structure learning in feedforward neural networks for regression problems. *IEEE Trans. Neural Networks*, **8**, 630–645.
Lippmann, R.P., 1987. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4–22.
Maier, H.R. and Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Model. Software*, **15**, 101–124.
Minns, A.W. and Hall, M.J, 1996. Artificial neural networks as rainfall-runoff models. *Hydrolog. Sci. J.*, **41**, 399–417.
Perrin, C., Michel, C. and Andreassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *J. Hydrol*., **242**, 275–301.
Rakem, Y., 1999. Analyse critique et reformulation mathématique d'un modèle pluie-débit empirique (GR4J). PhD thesis, Ecole Nationale des Ponts et Chaussées, Champs-sur-Marne, France.
Sajikumar, N. and Thandaveswara, B.S., 1999. A non-linear rainfall-runoff model using an artificial neural network, *J. Hydrol*., **216**, 32–55.
Shamseldin, A.Y., 1997. Application of neural network technique to rainfall-runoff modeling. *J. Hydrol*., **199**, 272–294.
Sivakumar, B., Phoon, K.K., Liong S.-Y. and Liaw C.-Y., 1999. A systematic approach to noise reduction in chaotic hydrological time series. *J. Hydrol*., **219**, 103–135.
Yang, X. and Michel, C., 2000. Flood forecasting with a watershed model : a new method of parameter updating. *Hydrolog. Sci. J.*, **45**, 537–-546.
Zealand, C.M., Burn, D.H. and Simonovic, S.P., 1999. Short term streamflow forecasting using artificial neural networks. *J. Hydrol*., **214**, 32–48.