Water Resources Management **16:** 431–445, 2002.

River Flow Modelling Using Fuzzy Decision Trees

D. HAN[1], I. D. CLUCKIE[1], D. KARBASSIOUN[1], J. LAWRY[2] and B. KRAUSKOPF[2]

[1] *Water and Environmental Management Research Centre, Department of Civil Engineering, University of Bristol, U.K.;* [2] *Department of Engineering Mathematics, University of Bristol, U.K.*
(author for correspondence, e-mail: d.han@bristol.ac.uk)*

(Received: 15 March 2001; in final form: 7 January 2002)

Abstract. A modern real time flood forecasting system requires its mathematical model(s) to handle highly complex rainfall runoff processes. Uncertainty in real time flood forecasting will involve a variety of components such as measurement noise from telemetry systems, inadequacy of the models, insufficiency of catchment conditions, etc. Probabilistic forecasting is becoming more and more important in this field. This article describes a novel attempt to use a Fuzzy Logic approach for river flow modelling based on fuzzy decision trees. These trees are learnt from data using the MA-ID3 algorithm. This is an extension of Quinlan's ID3 and is based on mass assignments. MA-ID3 allows for the incorporation of fuzzy attribute and class values into decision trees aiding generalisation and providing a framework for representing linguistic rules. The article showed that with only five fuzzy labels, the FDT model performed reasonably well and a comparison with a Neural Network model (Back Propagation) was carried out. Furthermore, the FDT model indicated that the rainfall values of four or five days before the prediction time are regarded as more informative to the prediction than the more recent ones. Although its performance is not as good as the neural network model in the test case, its glass box nature could provide some useful insight about the hydrological processes.

Key words: flood forecasting, fuzzy decision trees, hydroinformatics, river flow

1. Introduction

The most influential publication in the area of fuzzy set theory appeared in 1965 (Zadeh, 1965) and the development of this theory for almost 20 yr remained in the academic realm. Over the last two decades, however, a wide range of applications have emerged. From a methodological point of view, the application areas for Fuzzy Technology (FT) can be classified into the following categories (Zimmermann, 1996):

(1) Algorithmic Applications, e.g.
    Fuzzy Mathematical Programming; Fuzzy Planning Methods (CPM, Graphs);
    Fuzzy Petri Nets; Fuzzy Clustering, etc.

(2) Information Processing
Fuzzy Data Bank System; Fuzzy Query Languages; Fuzzy Languages, etc.
(3) Knowledge Based Applications
Expert Systems; Fuzzy Control; Knowledge Based Diagnosis, etc.
(4) Hybrid Application Areas
Fuzzy Data Analysis; Fuzzy Supervisory Control, etc.

Nowadays, the combination of FT with neural nets and genetic algorithms is rapidly increasing giving rise to the paradigm of soft computing. Attempts have been made in the past to introduce this new technology into water engineering. It has been found that modelling water movement in the unsaturated soil matrix was suitable for adopting fuzzy rules and in comparison with the numerical solution from the Richards equation, the fuzzy model performed quite well (Bardossy *et al*., 1995). Fuzzy algorithms have also been applied in urban drainage system modelling that was able to incorporate the GIS and remotely sensed thermal map to estimate the runoff potential (Campana, *et al*., 1995). Bardossy (1996) tried to use fuzzy rules to model infiltration, surface runoff and unsaturated flow. It was found that fuzzy systems provide a robust tool which can handle non-linearities, without requiring a prescribed functional structure. Furthermore fuzzy rule-based models can easily be coupled; for example, a model for flow in porous media may be coupled with a bacteriological growth model. They are capable of combining physical laws, expert knowledge and measurement data. Efforts were made to combine fuzzy theory with Genetic Algorithms to simulate the infiltration process and it was found that more accurate simulations could be achieved than with the standard Newton-Raphson method (Zeigler, *et al*., 1996). In flood control, Cheng (1999) has proposed a fuzzy optimal model for the flood control system of the upper and middle reaches of the Yangtze River and found the model effective and flexible when it was validated with three typical historical floods. Panigrahi and Mujumdar (2000) developed a fuzzy rule based model for the operation of a single purpose reservoir. The model operates on an 'if – then' principle, where the 'if' is a vector of fuzzy premises and the 'then' is a vector of fuzzy consequences.

It is clear that fuzzy technology is playing a more important role in modern day water engineering. One area deserving further investigation is real time flood forecasting. Stuber and Gemmar (1997) proposed an approach for data analysis and forecasting with neural fuzzy systems. In their article, two different system approaches are discussed: (a) a neural network for supervised learning of the functional behaviour of time series data and its approximation, and (b) a fuzzy system for modelling of the system behaviour with possibilities to exploit expert information and for systematic optimisation. Recently, See and Openshaw (1999) also combined a fuzzy logic model with neural networks. They split the forecasting data set into subsets before training with a series of neural networks. These networks were then recombined via a rule-based fuzzy logic model that has been optimized using a genetic algorithm. The overall results indicate that this methodology may provide a good performance, low-cost solution, which may be readily integrated into existing operational flood forecasting and warning systems.

It is known that a modern real time flood forecasting system demands that its mathematical model(s) handle highly complex rainfall runoff processes. Uncertainty in real time flood forecasting will involve a variety of components such as measurement noise from telemetry systems, inadequacy of the models, insufficiency of catchment conditions, etc. Fuzzy technology has a great potential to tackle the uncertainty problems in this field. This article will adopt a novel approach using a fuzzy machine learning algorithm MA-ID3 implemented in the logic programming language FRIL (Baldwin *et al.*, 1995).

2. Fuzziness in Hydrological Systems

Hydrological systems are dynamic and fuzzy. The contributing factors are difficult to evaluate and measure. Traditionally, rainfall records were obtained by raingauges that can produce up to 10% measurement errors due to a variety of causes, such as wind, splashing, trees, surrounding areas, etc. (Nespor and Sevruk, 1999). Habib *et al.* (2001) found that the commonly used tipping bucket raingauges suffer from significant errors if the measurements are based on time scales less than 10 to 15 min. In addition, raingauges also suffered from malfunctioning of the tipping-bucket frequently caused by biological and mechanical fouling, and human interference during a thirty major storms experiment carried out in Goodwin Creek, a small research watershed in northern Mississippi (Steiner *et al.*, 1999). The catchment average rainfall is usually calculated by the Thiessen polygon method, or more complicated mathematical processes (such as polynomial/Spline surface fitting or Kriging, etc). In real time flood forecasting situations, the telemetry raingauge network is usually quite sparse (e.g., about 1 in 200 km^2 in SW England, Han *et al.*, 2000) and the derived catchment average rainfall is an uncertain number due to the spatial sampling problems (Cluckie and Han, 2000).

To overcome the spatial sampling problems caused by sparse raingauge networks, more gauges are required but in reality, for economic constraints, it is impractical to set up a dense raingauge network for operational real time flood forecasting purposes. One solution is to use remote sensing technology such as weather radars to provide high resolution rainfall measurements in space and in time. Some modern hydrological radars can achieve very high definition rainfall measurements up to 250 by 250 m in space over a large area (50 to 76 km in radius) (Han *et al.*, 2000). However, radar measurement of rainfall is far from perfect and could still suffer from a variety of detrimental factors, such as variations in the relationship between back scattered energy and rainfall rate, effects of variation in precipitation with height, anomalous propagation of the beam. For example, WSR-88D radars used by the US National Weather Service could underestimate rainfall by about 25% in stratiform rain and overestimate it by about 33% in thunderstorm rain. Yet it is commonly observed that some WSR-88D radars systematically underestimate rainfall by a factor of 2 or more in stratiform rain (Ulbrich and Lee, 1999). On the other hand, stream flow records are usually derived from hydraulic measuring structures, such as weirs. These flow records are more accurate than rainfall measurements, however, discrepancies still exist and uncertainties in the records are inevitable (Herschy, 1999).

Traditional hydrological modelling tends to treat the input and output as definite values and lacks effective means to tackle the uncertainties in the rainfall and flow measurements. Nowadays more and more attention has been focused on uncertainty issues in real time flood forecasting. Fuzzy technology could take a more active role in this field and in this research, especially when combined with machine learning algorithms. In the following we show how a fuzzy decision tree induction algorithm MA-ID3 can be used to learn fuzzy IF-THEN rules which enhance the performance of operational real time flood forecasting systems (Han, *et al*., 2000). This algorithm has been implemented in the language FRIL that has built in capabilities for handling both fuzzy and probabilistic uncertainty.

3. The Catchment

The data used for this article were collected in a region called Bird Creek in the U.S.A. The data formed part of a real-time hydrological model intercomparison exercise conducted in Vancouver, Canada in 1987 and reported by WMO (WMO, 1992). The data set is divided into two parts: a calibration (training) period and a verification (testing) period. The daily rainfall values were derived from 12 rain gauges situated in/near the catchment area. The river flow values were obtained from a continuous stage recorder. The period used for model calibration spanned some eight years from October 1955 to September 1963, and the verification period ranged from November 1972 to November 1974. During the calibration period the discharge at the basin outlet ranged from 0 to 2540 $m^3\ s^{-1}$ and rainfall up to 153.8 mm day^{-1}. The highest recorded discharge during the verification period was 1506 $m^3\ s^{-1}$ (Hajjam, 1997).

The bird Creek catchment covers an area of 2344 km^2 and is located in Oklahoma close to the northern state border with Kansas. The outlet of the basin is near Sperry about ten kilometers north of Tulsa. The catchment is relatively low lying with altitudes ranging from 175 m up to 390 m above the mean sea level. There are no mountains or large water surfaces to influence local climatic conditions. Some twenty percent of the catchment surface is covered by the forest while the main vegetative cover is grassland. The storage capacity of the soil is very high (Georgakakos and Smith, 1990). The river basin and the stream network are shown in Figure 1.

The catchment receives significant rainfall in most years, and the catchment can be classified as humid although extended periods with very low rainfall can occur. Well defined rainy seasons occur in the spring and summer with rain in the form of

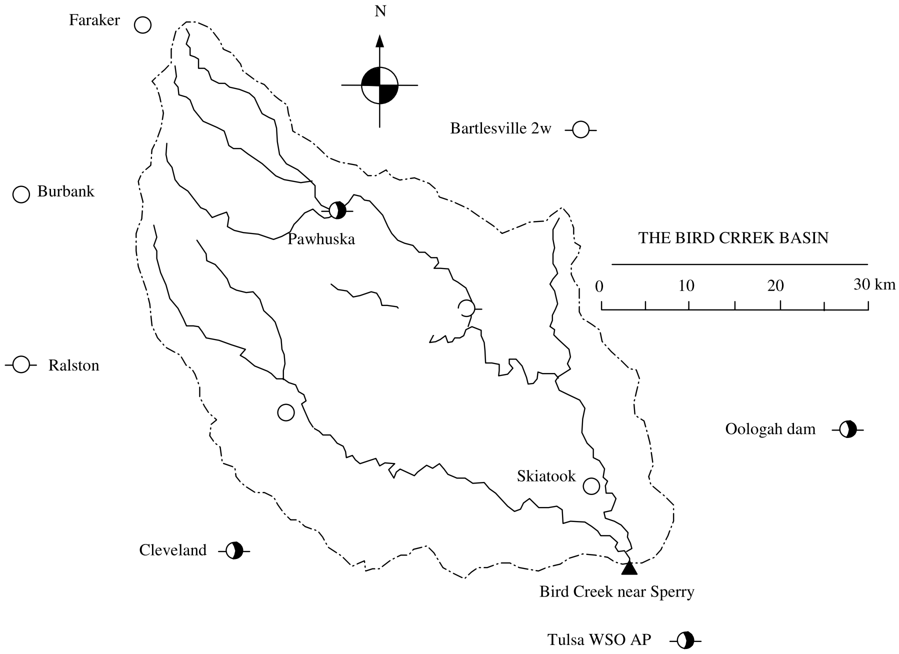

Figure 1. The Bird Creek drainage basin (Source WNO, 1992).

showers and thundershowers of convective origin. Snowfall remains on the ground for only a very short time. From the latter part of July to September air temperatures are high (38 °C is common) and as a result significant evapotranspiration occurs during this time. At the same time, relative humidity is low accompanied and southerly breezes are common (Georgakakos *et al.*, 1988).

4. Fuzzy Modelling

In this work the semantics for fuzzy membership functions are based on the voting model (Lawry, 1998). The membership value of a value x in a fuzzy set representing a term w is then taken to be the proportion of voters from some population who accept w as a suitable label for x. In practice voting experiments are not actually carried out and instead the membership functions are predefined and the voting model is used to justify the operations used on fuzzy sets. This framework can then be used to learn and represent fuzzy rules for classifying the data of both rainfall and river flow into representing fuzzy sets (Figure 2). For example, rainfall values are split up into five overlapping fuzzy sets described by labels such as: very low, low, medium, high and very high. Both the number of fuzzy sets used and the degree of overlap can be varied noting that an increase generally affects the level of interpolation (usually increasing complexity). The fuzzy sets used take the shape of trapezoid but can be formatted to preference.

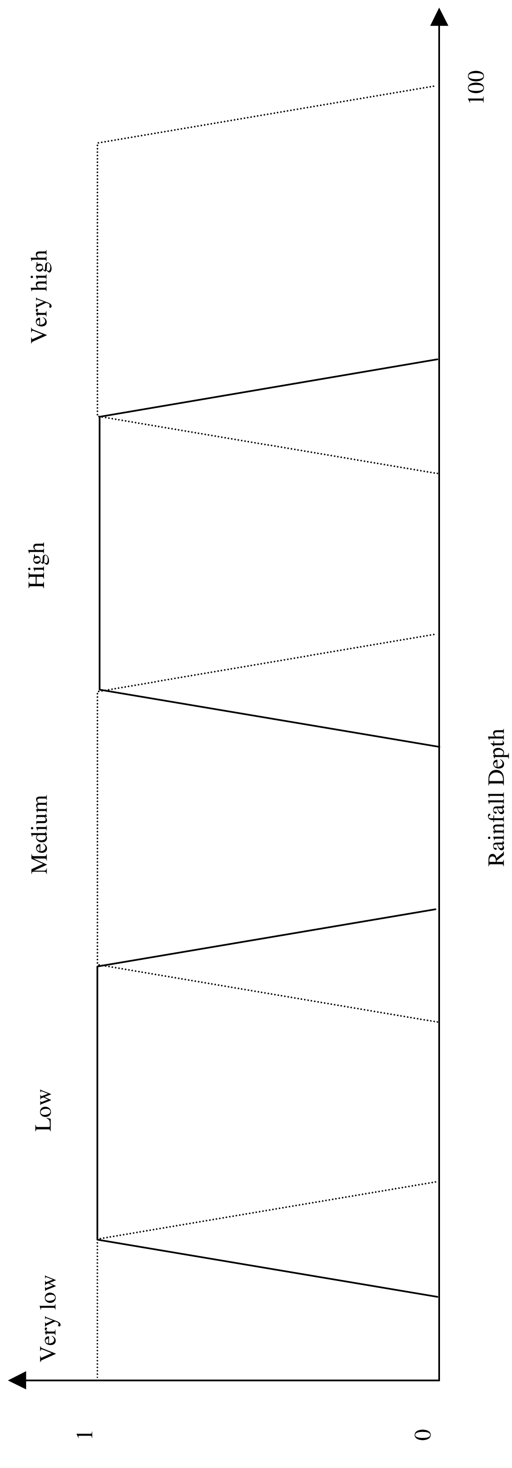

Figure 2. Rainfall fuzzy set classification.

The modelling process is based on the fuzzy ID3 algorithm as proposed by Baldwin *et al*. (1998). This process provides a method which generates a decision tree based on conditional probabilities. The conditional probabilities are relative to each particular branch of the decision trees, which are linked through IF-THEN type statements: i.e.

IF the rainfall today is *medium*

AND IF the rainfall yesterday was *very high*

THEN it is *highly likely* that the river flow today will be *high*

The above expression may be obtained from a branch of the decision tree with a conditional probability such as:

Pr (River flow today *high*| Rainfall today *medium*
AND Rainfall yesterday *very high*) = 0.95

Essentially both the river flow output and the conditional probabilities are fuzzy concepts where a probability of 0.95 has a high membership value in the 'highly likely' fuzzy set.

Since the universe of the predicted output (river flow) is fuzzy, it is necessary to aggregate the results based on the relative conditional probabilities since any one output may fall in a number or even all of the fuzzy set. This aspect is also a component of the approach adopted, which offers to weigh values by proportioning the probability that the particular output is chosen multiplied by the probability that the respective fuzzy set labels are appropriate (Baldwin *et al*., 1998). For the prediction to produce an output, a process of 'defuzzification' is necessary. This is done simply by weighting the river flow values according to their respective probabilities and the average fuzzy set value for each distinctive fuzzy set (i.e. the centroidal value which best suits the linguistic description).

In this research, the MA-ID3 algorithm developed by Baldwin *et al*. (1998) is adopted to generate a set of classification rules from the data, based on Shannon's measure of entropy, in the form of a traditional decision tree for a collection of data related to a number of classes. Features are selected on the basis of maximising the expected information gain of evaluating them as quantified by Shannon's measure. The decision tree is a combination of nodes (corresponding to features) and branches (corresponding to particular feature values). The prime aim of the MA-ID3 algorithm is to construct a tree giving the most concise rules for checking features for classification. An iterative method is then performed ranking the features according to their effectiveness in partitioning the set of defined classes.

Therefore, the fuzzy sets for rainfall and stream flow are:

For rainfall: (over a domain from 0 to 50)

$\mu_{very\ low}$ = [10:1, 16.5:0], μ_{low} = [3.5:0,10:1,20:1,26.5:0], μ_{medium} = [13.5:0, 20:1, 30:1, 36.5:0], μ_{high} = [23.5:0,30:1, 40:1, 46.5:0], $\mu_{very\ high}$ = [33.5:0, 40:1].

For streamflow: (over a domain from 0 to 6500)

$\mu_{very\ low}$ = [73:1, 105.15:0], μ_{low} = [0:0, 73:1, 137.3:1, 189.05:0], μ_{medium} = , μ_{high} [189.05:0, 240.8:1, 919.7:1, 6500:0], $\mu_{very\ high}$ = [580.25:0, 919.7:1].

Note that here we are using the Fril notation for piecewise linear functions where $[x_i : y_1, \ldots, x_n : y_n]$ denotes a function $F(x)$ such that

$$\forall x \in [x_i, x_{i+1}] F(x) = (y_i - y_{i+1}/x_i - x_{i+1})x + (x_i y_{i+1} - y_i x_{i+1}/x_i - x_{i+1})$$

$$\text{for} \quad i = 1, \ldots, n \, .$$

The rainfall runoff data records are divided into two parts: a model calibration data set (700 days) and a model validation data set (150 days). Initial attempts were made using the ‘exponential time steps’ method. The calibration data set are classified through five fuzzy sets at a degree of overlap of 0.65. A decision tree is created according to the fuzzy set classification and the conditional probability of the stream flow output. Examing the full decision tree, it is noticeable that points tend to cluster in the very low fuzzy sets which is due to the high occurrence of zero values of rainfall. The problem is that zero is not providing the system with any information to form a basis and thus the system tends to bias towards zero. Figure 3 displays the actual results during the model calibration process with a 13.28% difference between the simulated flow and measured flow. Despite the uniform displacement of simulated flow, the model is able to map the actual peaks reasonably well. A correction to this displacement is made as shown in Figure 4.

For real time flood forecasting systems, on-line measured flow records are usually available to the modelling process and it would be useful if the model is able to absorb the newly measured data records. Therefore, a feedback term is incorporated to allow the system to see the actual result for stream flow at time t (after the prediction is made) and use this actual value to predict the stream flow at $t + 1$ and so on. The improvements by introducing the feedback term are very obviously illustrated by Figure 5 during the calibration process with a much reduced error between the measured flow and simulated flow (down to 4.55%). The model validation results are presented in Figure 5 and the fuzzy logic model is able to simulate the flood peaks quite well.

The simulation results are summarised in Table 1. Utilising the MA-ID3 algorithm, the system selects the four rainfall values found most significant on building a powerful decision tree. Hydrologically it is noted that when the program selects the four most influential inputs from ten preceding rainfall values, the model

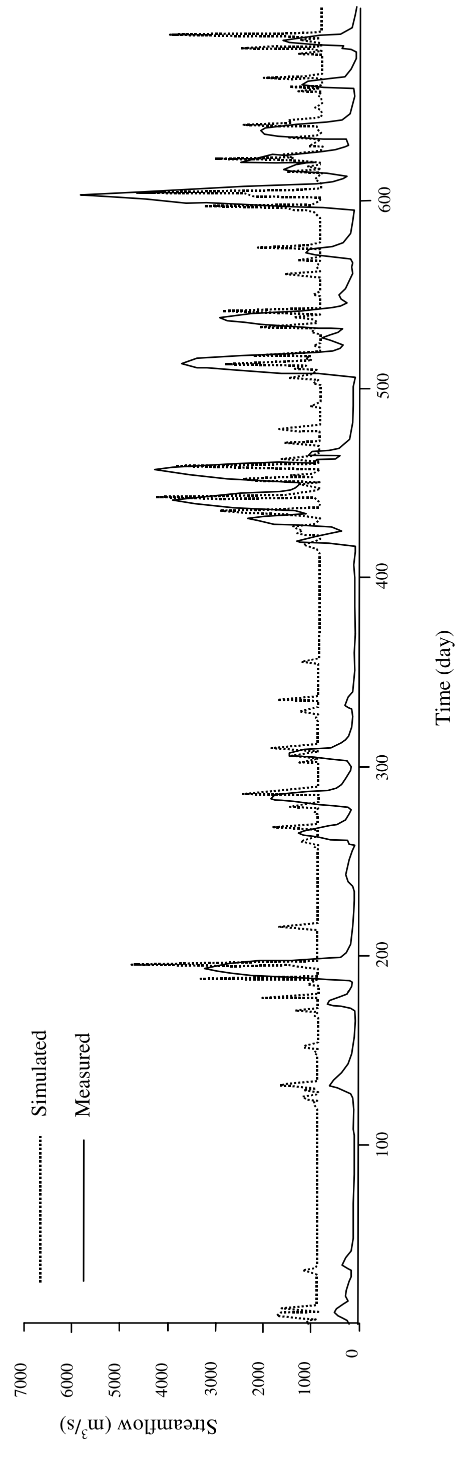

Figure 3. Model training with offset.

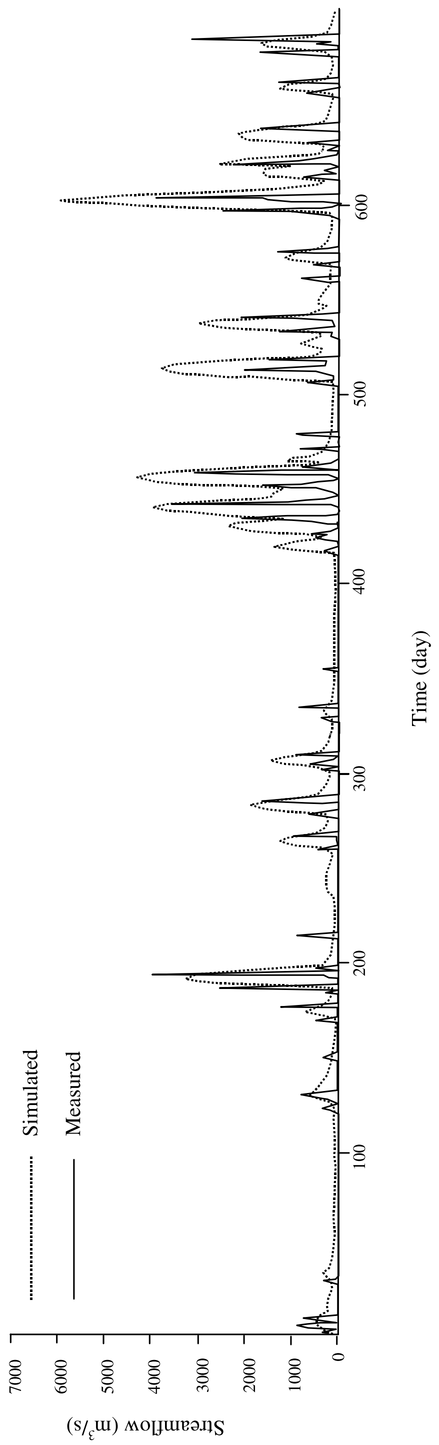

Figure 4. Model training with offset correction.

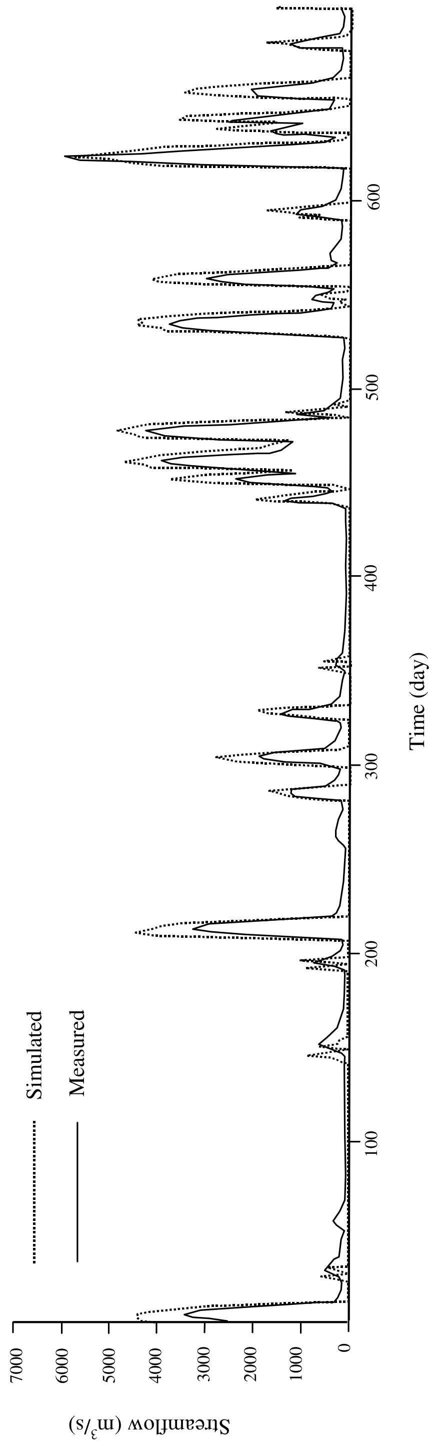

Figure 5. Model training with feedback.

Table I. Statistical comparison of the three main approaches

Bird Creek data	Format 1	Format 2
Training: 700 day Testing: 150 days	Exponential time steps	Incorporation of a feedback term
Approximate prediction time interval	1 day	1 day
Percentage error of output universe on training	12.3%	4.55%
Percentage error of output universe on testing	13.8%	7.51%

regards the values of four, five or six days ago as providing more information than the more recent rainfall values. This feature illustrates the glass box nature of fuzzy logic methods where the model provides feedback on what method is undertaken in classification. This information can thus highlight the values relevant to an ideal training set to optimise further the accuracy of prediction. The above comparison shows that the introduction of a feedback function does increase the accuracy but further work is needed to increase the warning lead time beyond a one day period. One solution may be to incorporate a feedback term for stream flow two or three days before rather than the day before. The extra stream flow information gives the system the classification attributes it needs to cope with an abundance of zero rainfall data.

Comparison with traditional Neural Network model is made using 230 days test data as shown in Figure 6. The NN model was built using a basic Back Propagation method with 10 rainfall inputs and 10 flow inputs. There were 20 nodes in the hidden layer. This structure was selected after a selection process using the training data. A major flood event with several peaks was included in the test data set. The highest peak was beyond any peaks seen by NN model and Fuzzy Decision Trees model during their training process. Overall, NN model had a superior performance although both models underestimated the highest peaks and overestimated small peaks. It is likely that more information is required to improve the modelling accuracy in addition to just rainfall and river flow data. Fuzzy Decision Trees model may be better suited when spatial rainfall data are introduced.

As mentioned above, the modelling system is coded in FRIL. FRIL is a logic programming language in which a program takes the form of a set of first-order predicate logic statements and rules (Baldwin *et al.*, 1995). More specifically, as with all such languages the rules are a restricted type of logic formula called Horn clauses. Limiting the language in this way means that the problem of deciding

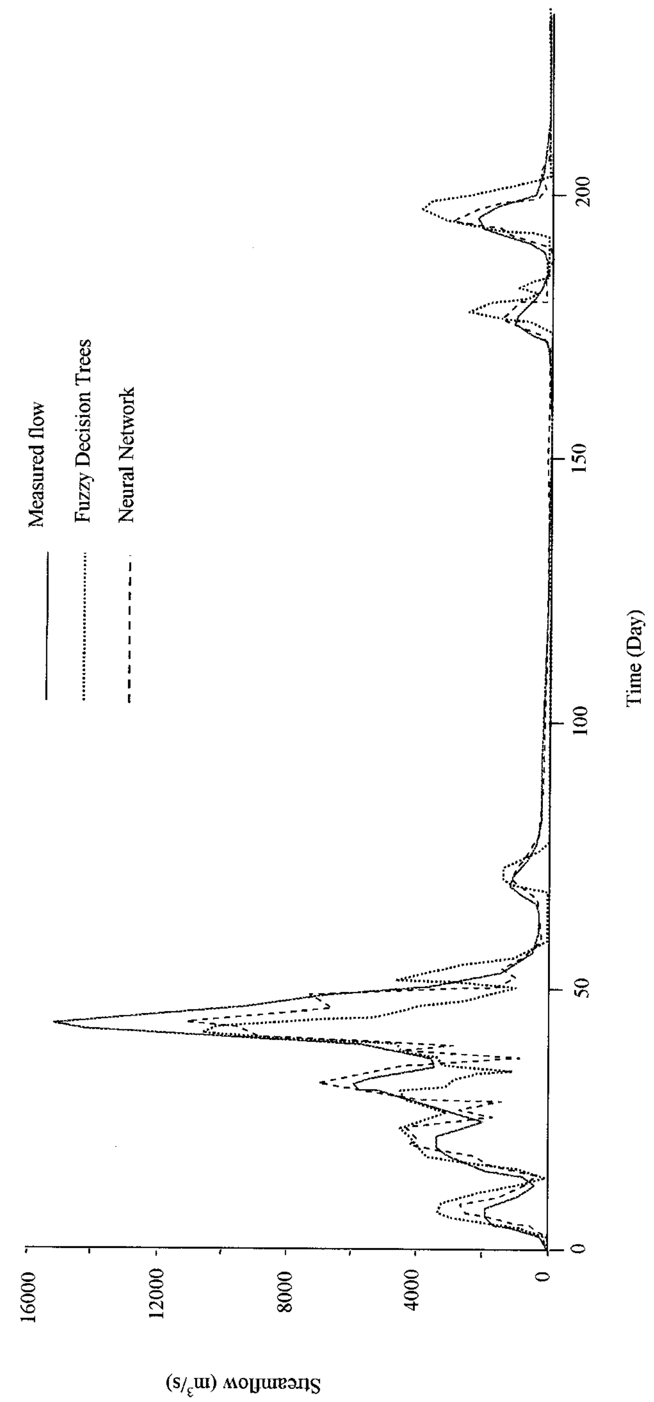

Figure 6. Comparison of fuzzy decision trees and neural network models.

whether a certain hypothesis follows from a particular program can be made tractable. Unlike other logic programming languages such as PROLOG, FRIL has a list based syntax which provides a very flexible knowledge representation framework. In addition, FRIL has built in capabilities for handling uncertainty which include mechanisms for representing and manipulating fuzzy sets as well as a calculus for inference based on interval probabilities.

5. Conclusions

This article explored the possibility of using fuzzy technology to model river flows for real time flood forecasting. There are a variety of uncertainties in rainfall and stream flow measurements, in spite of the advances of modern telemetry technology, and it is difficulty to treat these uncertainties using traditional deterministic methods. In this article we have shown that fuzzy decision trees model has its potential usage in flood forecasting. With only five fuzzy labels, the FDT model performed reasonably well. The FDT model indicated that the rainfall values of four or five days before the prediction time are regarded as more informative to the prediction than the more recent ones. Although its performance is not as good as the neural network model in test case, its glass box nature could provide some useful insight about the hydrological processes. On the other hand, NN models are black box models which lack the see-through ability. However, results from both NN and FDT models illustrated a common problem in hydrological modelling, i.e., lack of hydrological measurements, such as spatial rainfall, soil moisture, etc. It is unlikely that model prediction accuracy for both FDT and NN could be further improved significantly without extra hydrological information.

References

Baldwin, J. F., Martin, T. P. and Pilsworth, B. W.: 1995, *Fril: Fuzzy and Evidential Reasoning in Artificial Intelligence, Taunton*, Research Studies Press.

Baldwin, J. F., Lawry, J. and Martin, T. P.: 1998, 'Mass Assignment Based Induction of Decision Trees on Words', *Proceedings of IPMU 1998*, Vol. 1, pp. 524–531.

Bardossy, A, Bronstrert, A. and Merz, B.: 1995, '1 Dimensional, 2 Dimensional and 3 Dimensional Modelling of Water Movement in the Unsaturated Soil Matrix using a Fuzzy Approach', *Advances in Water* **18**(4), 237.

Bardossy, A.: 1996, 'The Use of Fuzzy Rules for the Description of Elements of the Hydrological Cycle', *Ecol. Modelling* **85**(1), 59–65.

Campana, N. A., Mendiondo, E. M. and Tucci, C. E. M.: 1995, 'A Multisource Approach to Hydrologic Parameter Estimation in Urban Basins', *Water Sci. Technol.* **32**(1), 233–239.

Cheng, C.T.: 1999, 'Fuzzy Optimal Model for the Flood Control System of the Upper and Middle Reaches of the Yangtze River', *Hydrol. Sci. J.* **44**(4), 573–582.

Cluckie, I. D. and Han, D.: 2000, 'Fluvial Flood Forecasting', *Water Environ. Manage.* **14**(4), pp. 270–276.

Georgakakos, K. P. and Smith, G. F.: 1990, 'On Improved Hydrological Forecasting Results from a WMO Real-time Forecasting Experiment', *J. Hydrol.* **114**, pp. 17–45.

Georgakakos, K. P., Rajaram, H. and Li, S. G.: 1988, 'On Improved Operational Hydrological Forecasting of Stream Flows', *IAHR Report*, No. 325.

Habib, E., Krajewski, W. F. and Kruger: 2001, 'Sampling Errors of Tipping-bucket Rain Gauge Measurements', *J. Hydrol. Engineer.* **6**(2), 159–166.

Hajjam, S.: 1997, 'Real Time Flood Forecasting Model Intercomparison and Parameter Updating using Rain Gauge and Weather Radar Data', *Ph.D. Thesis*, Telford Research Institute, University of Salford, U.K.

Han, D., Cluckie, I. D., Griffith, R. J., and Austin, G.: 2000, 'Rainfall Measurement over Urban Catchment using Weather Radars', *J. Urban Technol.* **7**(1), 85–102.

Han, D., Cluckie, I. D., Karbassioun, D. and Lawry, J.: 2000, 'A Fuzzy Logic Approach to River Flow Modelling', *Proceedings of the Eighth International Symposium on Stochastic Hydraulics*, Beijing, July 25–28, 2000, ISBN 90 5809 166X, pp. 853–860.

Herschy, R. W.: 1999, *Hydrometry Principles and Practice*, 2nd ed., Wiley, Chichester, 376 p.

Lawry, J.: 1998, 'A Voting Mechanism for Fuzzy Logic', *Internat. J. Approx. Reasoning* **19**(3–4), 315–333.

Nespor, V. and Sevruk, B.: 1999, 'Estimation of Wind-induced Error of Rainfall Gauge Measurements, using a Numerical Simulation, *J. Atmosph. Oceanic Technol.* **16**(4), 450–464.

Panigrahi, D. P. and Mujumdar, P. P.: 2000, 'Reservoir Operation Modelling with Fuzzy Logic', *Water Recour. Manage.* **14**(2), 89–109.

See, L and Openshaw, S.: 1999, 'Applying Soft Computing Approaches to River Level Forecasting', *Hydrol. Sci. J.* **44**(5), 763–778.

Steiner, M., Smith, J. A., Burges, S. J., Alonso, C. V. and Darden, R. W.: 1999, 'Effect of Bias Adjustment and Rain Gauge Data Quality Control on Radar Rainfall Estimation', *Water Resour. Res.* **35**(8), 2487–2503.

Stuber, M. and Gemmar, P.: 1997, 'An Approach for Data Analysis and Forecasting with Neuro Fuzzy Systems – Demonstrated on Flood Events at River Mosel', *Comput. Intell.* **1226**, 468–477.

Yager, R. R. and Zadeh, L. A.: 1994, *Fuzzy Sets, Neural Networks, and Soft Computing Imprint*, Van Nostrand Reinhold, New York.

Zadeh, L. A.: 1965, 'Fuzzy Sets', *Information and Control* **8**, 338–353.

Zeigler, B. P., Moon, Y., Lopes, V. L. and Kim, J.: 1996, 'DEVS Approximation of Infiltration using Genetic Algorithm Optimization of a Fuzzy System', *Math. Computer Modell.* **23**(11–12), 215–228.

Zimmermann, H. J.: 1996, *Recent Developments in Fuzzy Logic and Intelligent Technology, Fuzzy Logic*, J. F. Baldwin (ed.), John Wiley & Sons Ltd.

Ulbrich, C. W. and Lee, L. G.: 1999, 'Rainfall Measurement Error by WSR-88D Radars Due to Variations in Z-R Law Parameters and the Radar Constant, *J. Atmosph. Oceanic Technol.* **16**(8), 1017–1024.

WMO: 1992, 'Simulated Real-time Intercomparison of Hydrological Models', *WMO report 779*, 241 pp. ISBN 92-63-10779-3.