A fuzzy inference system for modelling streamflow: Case of Letaba River, South Africa

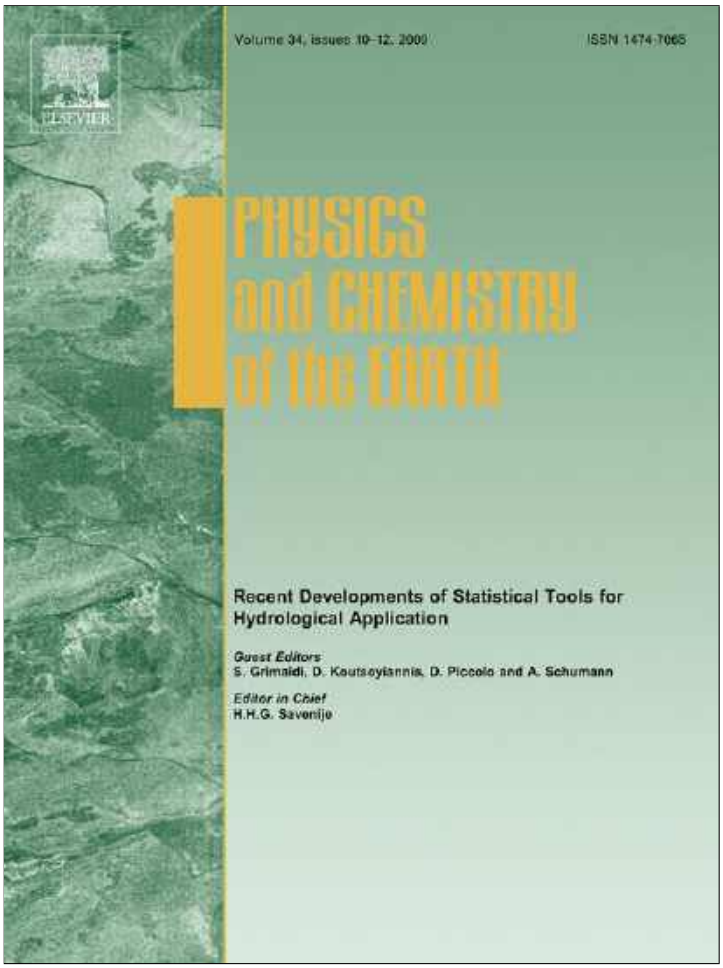
Article *in* Physics and Chemistry of the Earth Parts A/B/C · December 2009

DOI: 10.1016/j.pce.2009.06.001

2 authors:

Zacharia Katambara
Mbeya University of Science and Technology
31 PUBLICATIONS **293** CITATIONS
SEE PROFILE

John Ndiritu
University of the Witwatersrand
52 PUBLICATIONS **615** CITATIONS
SEE PROFILE



Physics and Chemistry of the Earth 34 (2009) 688–700

Contents lists available at ScienceDirect

Physics and Chemistry of the Earth

journal homepage: www.elsevier.com/locate/pce

A fuzzy inference system for modelling streamflow: Case of Letaba River, South Africa

Zacharia Katambara *, John Ndiritu

School of Civil and Environmental Engineering, University of the Witwatersrand, P Bag X3, Wits 2050, South Africa

ARTICLE INFO

Article history:
Received 10 July 2008
Received in revised form 22 May 2009
Accepted 2 June 2009
Available online 7 June 2009

Keywords:
Complex river system
Scale of monitoring
Takagi–Sugeno fuzzy inference system
Fuzzy streamflow modelling
Subtractive clustering

ABSTRACT

Streamflow modelling of Letaba River in South Africa is complicated by several factors including the existence of dams and other storage structures whose releases are intermittent and based on rules of thumb depending on the irrigation demands and the need to maintain the flow required in the Kruger National park (KNP). The KNP is located about a hundred kilometres downstream of the main storage and water flows through an alluvial aquifer where complex surface–groundwater interactions occur. Farmers abstract water intermittently along the route directly from the river or indirectly from the alluvial aquifer complicating the flow patterns even more. Consequently, the streamflow series in the river shows very little similarity to what would be considered as natural. The actual abstractions are not measured and only monthly estimates of the abstractions currently exist. Like in many other basins in South Africa, streamflow, groundwater level, rainfall and evaporation data in Letaba is sparse and not very reliable. The Takagi–Sugeno fuzzy inference system using subtractive clustering, an approach which are capable of dealing with vague and inadequate information and data has therefore been used to develop a daily streamflow model for Letaba River. In order to take into account the spatial variability and to maximize the use of the available data, the model is applied in a semi-distributed manner consisting of three river reaches. The shuffled complex evolution (SCE-UA) optimizer has been used to calibrate the model. Six years of data from March 2002 to April 2008 has been used for model calibration and verification. To maximize the Nash-Sutcliffe efficiency, the minimum number of clusters required was found to be 10 for 1000 data points in calibration. An analysis of the location of the cluster centers, the coefficients relating the inputs with the simulated streamflow, and the degrees of membership indicates that no single cluster can be associated to the simulation of a specific hydrologic process or component of the streamflow hydrograph (e.g. high flows or low flows). The fuzzy model does not therefore provide any evidence that it is not a pure black box.

The Nash–Sutcliffe efficiency results obtained in calibration and verification showed average values of 0.658 and 0.535 with poor values on the first river reach. Very low percent bias values averaging to −0.4% and −2.7% in calibration and verification are obtained highlighting the model's potential for applications where mass balance considerations are most important.

1. Introduction

In many river systems, dams are constructed across rivers for various purposes including managing demands downstream during periods of low flows. The demands typically include irrigation, municipal/domestic supply, industrial supply and ecological requirements. The common and the easiest mode of conveyance of water released from the reservoir is via the river as streamflow. In systems like the Letaba River in South Africa where the demand outstrips the supply, careful system operation that seeks to maximize on the resource is paramount. A valuable aid to such operation is a streamflow model that would inform how releases from the storage reservoirs, abstractions, meteorological conditions (rainfall and evaporation), etc. impact on the flow at various points of interest in the river. The appropriate streamflow modelling approaches would depend on several factors including the time step, the complexity of the problem, and the data available.

Traditionally, it was assumed that process model parameters could be obtained by field measurements but experience has shown that the difficulty of obtaining the required data accurately often forces these parameters to be calibrated. This effectively transforms process modelling into conceptual modelling. The calibration of conceptual models on the other hand requires data which in many cases are not available in sufficient quantities

* Corresponding author. Tel.: +27 82963201.
E-mail addresses: Zacharia.Katambara@students.wits.ac.za (Z. Katambara), John.Ndiritu@wits.ac.za (J. Ndiritu).

1474-7065/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pce.2009.06.001

(Hughes, 2004). Gőrgens (1983) suggested that long record lengths may be required in order to identify representative parameters for catchment modelling in semi-arid flow regimes. In response to the complexity of streamflow generation processes including non-linearity and imprecise and inadequate information, hydrologic modellers have pioneered new avenues. Approaches based on heuristic techniques like artificial neural networks (ANN) and fuzzy logic have been developed and applied. Fuzzy logic has the ability to describe knowledge in a descriptive human-like manner in the form of simple rules using linguistic variables. Further more, it offers a more flexible, less assumption dependent and self adaptive approach than the ANN and also provides a frame work to deal with vagueness in the data and uncertainty at various levels (Zimmermann, 2001; Xiong et al., 2001; Lohani et al., 2005, 2006). Other advantages include its capability to be linked to many numerical or analytical models and the fact that it is typically not computationally intensive (Xiong et al., 2001) particularly after the calibration process. Fuzzy logic may therefore offer an alternative to more commonly used hydrological modelling methods where the knowledge to enable the mathematical representation of the hydrological processes and the available information is significantly limiting (Altunkaynak and Sen, 2007; Katambara and Ndiritu, 2007).

However, it is recognized that for ungauged catchments or catchments with negligible amounts of flow data, it may be easier to obtain estimates of parameter values of process or conceptual models than black box models such as fuzzy logic based methods. This is because parameters of process and conceptual models generally relate more directly to the physical reality of the catchment than those of black box models. Process or conceptual models may therefore be more applicable in modelling ungauged catchments than fuzzy logic and other black box models. The possibility of transferring parameter values from a gauged to neighbouring ungauged catchments (that are similar) may however be applicable for black box models just as for process and conceptual models.

The use of fuzzy logic in hydrology is relatively new and includes the estimation of sediment transport from bare soil surface (Tayfur et al., 2003), rainfall–runoff modelling (Jacquin and Shamseldin, 2006; Vernieuwea et al., 2005), stage discharge relationships (Lohani et al., 2006) and flood forecasting (Xiong et al., 2001).

Without modelling the hydrological processes, fuzzy inference system maps the input to the output dataset. Hence, the fuzzy inference system can be considered to be a process of formulating and mapping a given input to an output using an environment based on fuzzy logic. Although most fuzzy logic applications in hydrology have applied this 'black box' approach, the possibility of including both fuzzy logic and conceptual modelling of the hydrological processes exists as Hundecha et al. (2001) demonstrated.

Streamflow in Letaba River (Fig. 1) is highly complex as a result of several features including: (i) irregular releases from the major storage (Tzaneen dam) and from storage weirs downstream of the dam, (ii) irrigation abstractions directly from the river and from the alluvial deposits next to the river bed, (iii) complex surface–groundwater interactions that are not monitored, (iv) evaporation from the river and evapotranspiration from riparian vegetation, and (v) unmonitored contributions from tributaries. As shown in Fig. 2 the streamflow time series' at various locations in the Letaba is highly impacted and has little similarity to a natural flow hydrograph. Daily irrigation abstraction data is not available and only monthly estimates aggregated for the whole river reach exist. Although daily releases from Tzaneen dam are available, releases from the storage weirs are not available and are difficult to estimate as the operation of the weirs is not based on actual measurements but a rule of thumb of how low or high the water depths at selected locations are in relation to the current irrigation demands. The application of fuzzy logic to modelling daily streamflow in the Letaba River was therefore more appealing than classical conceptual or process modelling. The first stage of the fuzzy modelling of the Letaba River applying a pure 'black box' approach is the

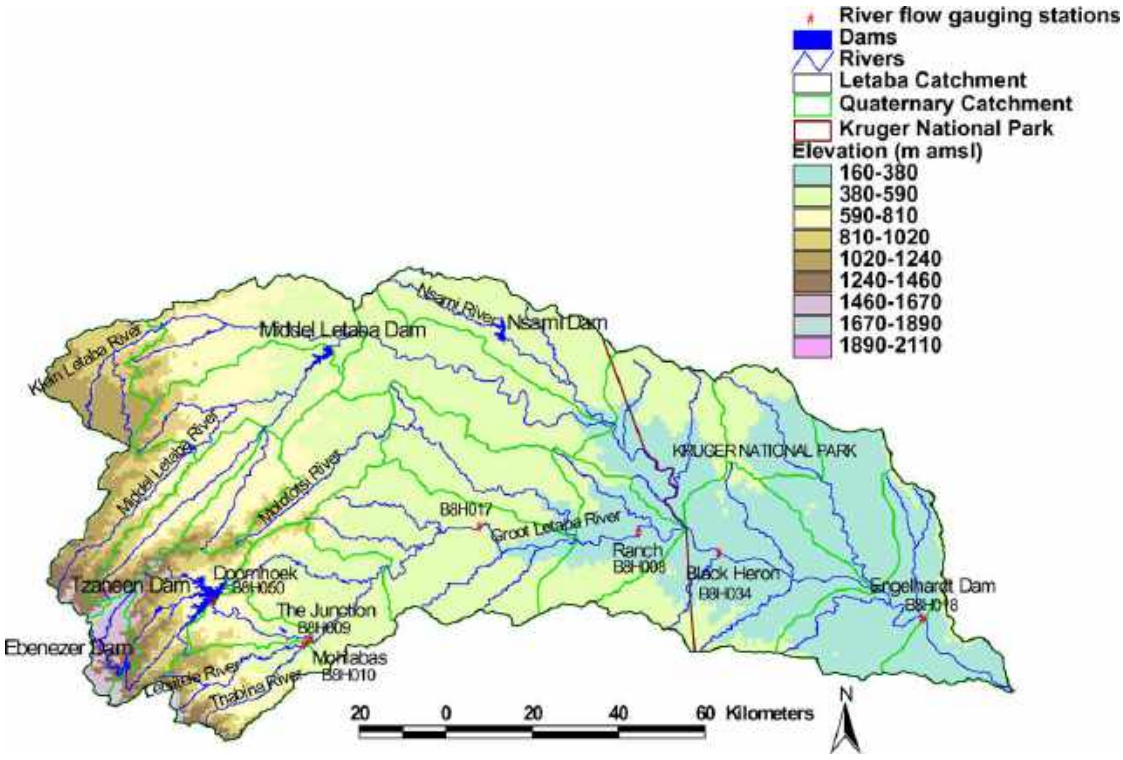


Fig. 1. Location of the Letaba River system.

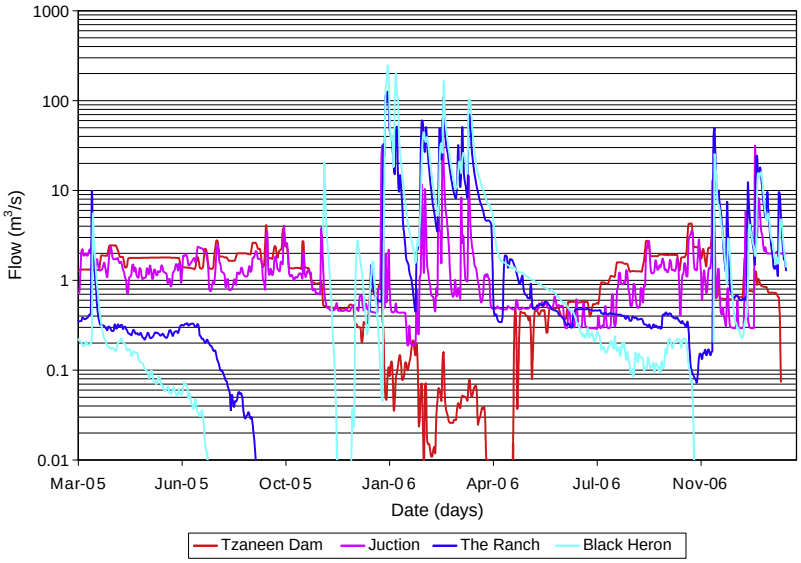


Fig. 2. Typical flow data for the four gauging stations along the Letaba River.

subject of this paper. Modelling that would incorporate known and inferred physical processes is the subject of additional work.

2. The Letaba River system

2.1. Location and physical characteristics

The Letaba catchment (Fig. 1) is within the Luvuvhu/Letaba Water Management Area of South Africa. The catchment is located in the north-eastern part of South Africa and covers an area of 13669 km^2. The catchment is in a semi-arid region with the western part of the catchment being mountainous with an altitude higher than 2000 m above mean sea level and decreasing gradually towards the eastern part of the catchment to slightly below 450 m above mean sea level. The main rivers (i.e. Groot Letaba River and Middle Letaba River) originate from this mountainous region and flow towards the lower eastern part of the catchment. Several reservoirs exist with Tzaneen Dam (capacity of 157.7×10^6 m^3) being the largest and the most downstream reservoir across the Letaba River (DWAF, 2006). Downstream of Tzaneen dam, several structures exist each of which serves a specific purpose which include stream flow monitoring (with gauging facilities), storage (weirs with retaining walls) and stream flow checking.

2.2. Climate

The rainfall is extremely seasonal with about 40–50% normally received in January and February (DWAF, 2004). The mean annual rainfall ranges from less than 450 mm to more than 2300 mm. The mean annual temperature ranges between 18 °C in the mountainous region to more than 28 °C in the eastern parts of the catchment with an average of about 25.5 °C. High and low temperatures occur in the months of January and July, respectively (DWAF, 2004).

In line with rainfall, the relative humidity is higher during wet months than during dry months. High values exist during the month of February ranging from about 70% in the west to above 72% in the east (DWAF, 2004).

2.3. Monitoring system

Several monitoring stations for stream flow, rainfall, evaporation, and groundwater levels exist in the catchment. In addition, surface water abstraction data is also recorded at some locations. Of the streamflow gauging stations, only a few operate in real-time. However, monitoring is still inadequate as the contribution from some of the tributaries as well as the releases made from storage weirs existing across the Letaba River are not measured. There are differences in the record length of the data as some monitoring stations were established earlier than others. Only concurrent record length has been used here in modelling. Typical flow data for the four streamflow gauging stations along the Letaba River are shown in Fig 2. It is clear that the flows are highly impacted by human activities.

2.4. Water demands

The water demands within the system are reflected in the annual water allocation which sums up to 130×10^6 m^3 (irrigation allocation is 103.9×10^6, the combined domestic and industrial water allocation is 7.13×10^6 m^3 and ecological requirement of 18.9×10^6 m^3). The annual historical yield (1:68 years) of Tzaneen Dam is 74.6×10^6 m^3 (DWAF, 2006). Therefore the various water demands within the system outstrip the supply from Tzaneen Dam.

2.5. Current system operation

The operation of the Letaba River system downstream of Tzaneen Dam is currently performed in manner that aims to balance the quantity of water flowing with the requirements while prioritising the reserve demand. When the available quantity of water does not meet the demands, releases are made from Tzaneen dam to supplement the flow. In line with this, releases are also made from unmonitored storage weirs in the system. In critical dry periods restrictions are imposed based on a predefined priority schedule to prevent total water unavailability to vital demands.

3. Fuzzy inference system

Fuzzy inference is defined as the process of mapping a set of input data sets into a set output data, using an approach based on fuzzy logic and falls under the category of black box models.

A fuzzy inference system has two main parts: the input (antecedence) and the output (consequence) part. The antecedence part could be either rule based or cluster based.

3.1. Rule based approach

For the rule based approach, each variable of the input and output dataset is grouped into subsets defined by linear or non-linear functions. The subset can be represented by linguistic variables (e.g. High, Medium, Low, and Very Low). Each input value can belong to one or more subsets and the likelihood of an input being in a subset is defined by a fuzzy value called degree of membership that ranges between 0 and 1. The process of obtaining subsets and determining the degree of membership is termed fuzzification. Fuzzy rules connect the input and output subsets together and the number of rules spans the whole range of the possible output. Typical fuzzy rules are simple logical IF–THEN statements and are stored in a fuzzy rule base. A typical rule is represented as IF-(antecedent part)–THEN-(consequence part). The central part of the inference system, the fuzzy inference engine, learns how to map the fuzzified input datasets to their corresponding output datasets by selecting appropriate rules from the fuzzy rule base. Depending on the type of the consequence part of the rule, the output from the inference system for each rule can be either a fuzzy value that needs to be transformed (deffuzzified) into a real number or a real number based on a linear or non-linear function attached to each rule. The former is referred to as Mamdani inference system (Mamdani and Assilian, 1975) and the latter the Takagi–Sugeno inference system (Takagi and Sugeno, 1985). Typically, the rule based fuzzy inference system has the disadvantage of requiring many rules in order to span the whole spectrum of the possible output – a disadvantage that the Takagi–Sugeno system is free from.

3.2. Clustering based approach

The clustering based fuzzy inference system establishes cluster centers and quantifies the fuzzyness of each data point based on its distance from the centers. A cluster center is defined as the point located at the shortest distance from the rest of the dataset and this distance is defined as the cluster potential. A number of clustering techniques exist and can be categorised in to two main groups; prior knowledge dependent and non-prior knowledge dependent approach. The non-prior knowledge dependent approach is often preferred as it does not require prior analysis of the dataset. Mountain clustering (Yager and Filev, 1994), a non-prior knowledge dependent approach establishes some grid points to cover the whole spectrum of dataset and assumes that the cluster centers are located on grid points. The potential for each grid point is determined and becomes the basis of clustering. The disadvantage of using the grids as the basis for clustering is the intense computation involved and large storage space requirement (that is exponentially proportional to the dimension of the data). Subtractive clustering (Chiu, 1994), another non-prior knowledge dependent approach assumes the location of the cluster centers coincides with specific data points and the intensity of computation involved only increases with the length of the data not the dimension.

4. Model development

Considering the advantages of subtractive clustering and the Takagi–Sugeno inference system, the two methods were selected in order to carry out the computational analysis efficiently. Considering the location of the streamflow gauging stations along the Letaba River (Fig. 1), the reliability of the measured information and the physical characteristics, the system has been demarcated into three river reaches with the first reach covering the section from Tzaneen Dam to the Junction weir, the second reach covering the section Junction – the Ranch weir and the third section covering the Ranch to Black Heron weir. The gauging stations located at the ends of these respective river reaches form the basis for the demarcation. Hence the outflow from the upper river reach is considered an inflow into the adjacent downstream river reach. This serves to incorporate spatial variability along the river including the non-existence of abstractions between the Ranch and Black Heron weir. The objective of the modelling is to come up with a model that estimates the daily streamflow ($q_{sim,i}$) at the downstream reach of the river given the rainfall (R_i), upstream inflow ($q_{in,i}$), evaporation (e_i), abstraction ($q_{abs,i}$) along the first and second reach and contribution from Letsitele tributary ($q_{tri,i}$) to the second river reach. Considering the current day's outflow to also depend on the previous day's outflow ($q_{out,i-1}$), rainfall (R_{i-1}), tributary contribution ($q_{tri,i-1}$), and inflow ($q_{in,i-1}$), the model takes the form:

$$q_{sim,i} = f(q_{in,1}, q_{in,i-1}, R_i, R_{i-1}, e_i, q_{abs,i}, q_{out,i-1}) \tag{1a}$$

$$q_{sim,i} = f(q_{in,1}, q_{in,i-1}, R_i, R_{i-1}, e_i, q_{abs,i}, q_{tri,i}, q_{tri,i-1}, q_{out,i-1}) \tag{1b}$$

$$q_{sim,i} = f(q_{in,1}, q_{in,i-1}, R_i, R_{i-1}, e_i, q_{out,i-1}) \tag{1c}$$

for the first, second and the third river reach, respectively.

The hydrology of the Letaba River is perceived to have changed significantly after the massive 2000 floods and the available abstraction estimates start from March 2002. Hence the records used for model calibration and verification range from March 2002 to April 2008.

From Eq. (1), the input set of variables considered can be denoted by a general variable z as:

$$\{z_{i,1}, z_{i,2}, \cdots z_{i,k} \cdots z_{i,q}\} = \{q_{in,i}, q_{in,i-1}, R_i, e_i, \cdots, q_{out,i-1}\} \tag{2}$$

where $z_{i,k}$ is the kth input variable, k (1 to q) is the counter for the dimension of the dataset, q is the number of input variables, i (1 to n) is the counter of the input data points and n is the number (days) of data points. The data points are normalized to range between 0 and 1 using the function

$$x_{i,k} = \frac{(z_{i,k} - z_{min,k})}{(z_{max,k} - z_{min,k})} \tag{3}$$

where $z_{min,k}$ and $z_{max,k}$ is the minimum and maximum value of the kth variable.

The potential of each data point is obtained as:

$$P_i = \sum_{j=1}^{n} \sum_{k=1}^{q} \exp(-\|x_{i,k} - x_{j,k}\|^2/(r_a/2)^2) \tag{4}$$

where P_i is the potential of the ith data point, whose center is at $x_{i,k}$, j is the counter for cluster center and r_a is a positive constant defining the neighbourhood range of the cluster (radius of the hypersphere cluster in data space).

The first cluster center x_{c1} is selected as the point with the largest potential of P_{ci}. The next cluster center is selected as the point with the highest potential after penalizing the initial cluster center and the points in the neighbourhood. This expression is given as:

$$P_{i+1} = P_i - P_{c1} \exp(-\|x_{i,k} - x_{j,k}\|^2/(r_b/s)^2) \tag{5}$$

where r_b is given as:

$$r_b = \eta \times r_a \tag{6}$$

where r_b is a positive constant which defines the efficient subtractive range and η is a quash factor which is set to a value greater than 1 (hence $r_b > r_a$) to prevent obtaining closely spaced clusters. The suggested values for η and r_a are $1.25 \leqslant \eta < 1.5$ and $0.15 \leqslant r_a \leqslant 0.30$ (Demirli et al., 2003). For the analysis reported here η and r_a were subjectively set to 1.25 and 0.15, respectively.The obtained cluster center is checked for the minimum distance given as (Demirli et al., 2003):

$$d_{min}/r_a + P_i/P_{c1} \geqslant 1 \tag{7}$$

where d_{min} is the minimum distance between the computed center with other centers. If the cluster center does not fulfil the above condition, its potential is set to zero and the data point with the next highest potential P_i is selected as the new possible cluster center. This data point is also checked for the same condition (Eq. (7)). Clustering ends when the following condition is fulfilled:

$$P_i < \varepsilon P_{ci} \tag{8}$$

where ε is the rejection ratio.

An alternative to the use of the rejection ratio that was used in this study was to increase the number of clusters (starting with a low number) until further improvement in model performance is not significant.

After determining the cluster centers, a Gaussian function is used to determine the degree of membership ($DOM_{i,m}$ ($m = 1$ *to* NC)) of every input data point given by:

$$DOM_{i,m} = \exp\left(-\frac{4}{r_a}\sum_{k=1}^{q}(x_{i,k} - c_{m,k})\right) \tag{9}$$

where $c_{m,k}$ is the cluster center. The sum of the degrees of membership of any given point for all the cluster centers is then obtained as:

$$DOMSUM_i = \sum_{m=1}^{NC} DOM_{i,j} \tag{10}$$

where NC is the number of clusters in consideration.

Each cluster center is associated with a function of the form:

$$y_{i,m} = a_{0,m} + a_{1,m}x_{1,i}^{b_1} + \cdots + a_{q,m}x_{q,m}^{b_q} \tag{11}$$

where $a_{0,j}$, $a_{1,j}, \ldots, a_{q,j}$ are the coefficients and $b_{1,j}, \ldots, b_{q,j}$ are the exponents which determine the degree of non-linearity of the function. A linear function with $b_{1,j}, \ldots, b_{q,j}$ is equal to 1 was applied in this study. It is the coefficients which transform the input datasets $x_{i,k}$ to the respective magnitude of streamflow based on the magnitude of the inputs. The simulated streamflow value is determined by using the weighted average method as follows:

$$q_{sim,i} = \sum_{m=1}^{NC}\left(y_{i,m} \times \frac{DOM_{i,m}}{DOMSUM_i}\right) \tag{12}$$

The simulated values are compared with the observed streamflows using the root mean square error (*RMSE*) objective function given as:

$$Minimize\ RMSE = \sqrt{\sum_{j=1}^{n}(q_{obs_i} - q_{sim_i})^2/n} \tag{13}$$

RMSE is minimized by varying the coefficients in Eq. (11). Although Eq. (11) was applied as a linear equation and Eq. (12) is also linear, the objective function, *RMSE* is non-linear and the optimization is still a catchment model calibration problem. The widely applied catchment model calibration method, the shuffled complex evolution (SCE-UA) (Duan et al., 1992) was therefore selected for the purpose of determining the optimum parameters of Eq. (11). This method combines the strength of the downhill simplex procedure with the concepts of controlled random search competitive evolution and complex shuffling. It commences by randomly selecting a population of feasible points that are sorted based on the value of the objective function. These points are then partitioned into complexes. Each individual complex is then evolved towards global improvement by applying competitive evolution techniques that are based on the downhill simplex method. Information sharing among the complexes is permitted through shuffling and reassigning to new complexes. The whole process is repeated until some termination criteria are met. Several researchers have observed the strength of SCE-UA with Ndiritu and Daniell (2001) suggesting that the SCE-UA should be the first preference when dealing with continuous problems including calibrating a rainfall-runoff model.

In this study, an iterative approach was used to determine the minimum number of clusters which will obtain the highest correlation coefficient and Nash–Sutcliff efficiency for each reach. Model performance improved with number of clusters up to some threshold. The minimum number of cluster centers which produced the best model results for the first, second and third river reach were 10, 10, and 5, respectively. The first and second river reach required twice as many cluster centers as the third river reach – an indication of the complexity resulting from more severe human impacts on the first two reaches.

5. Model evaluation

Several model evaluation techniques exist and the Nash–Sutcliffe efficiency (*NSE*), correlation coefficient (*CCoef*), percent bias (*PBIAS*), and the root mean square (*RSR*) were used. The Nash–Sutcliffe efficiency (*NSE*) (Nash and Sutcliffe, 1970) which is a normalized statistic determines the relative magnitude of the residual variance compared to the observed variance and it gives an indication of how well the observed and simulated results fit to 1:1 line. It is obtained as:

$$NSE = 1 - \left[\sum_{i=1}^{n}(Y_{obs_i} - Y_{sim_i})^2 \Big/ \sum_{i=1}^{n}(Y_{obs_i} - \overline{Y}_{obs_i})^2\right] \tag{14}$$

where Y_{obs_i} is the ith observation for the constituent being evaluated, Y_{sim_i} is the ith simulated value for the constituent being evaluated, $\overline{Y}_{obs}$ is the mean of the observed data for the constituent being evaluated and n is the total number of observations in consideration.

The correlation coefficient (*CCoef*) is based on the Pearson product moment correlation coefficient of the simulated and the observed flow series and is obtained as:

$$CCoef = \left[\sum_{i=1}^{n}(Y_{obs_i} - \overline{Y}_{obsi})(Y_{sim_i} - \overline{Y}_{sim_i}) \Big/ \sqrt{\textstyle\sum_{i=1}^{n}(Y_{obs_i} - \overline{Y}_{obsi})^2 \sum_{i=1}^{n}(Y_{sim_i} - \overline{Y}_{sim_i})^2}\right] \tag{15}$$

where $\overline{Y}_{sim}$ is the mean of the simulated values.

PBIAS measures the averaged tendency of the simulated series to be larger or smaller than their observed series. Positive values indicate model underestimation bias and negative values indicate model overestimation bias. *PBIAS* (Moriasi et al., 2007) is as obtained as:

$$PBIAS = \left[\textstyle\sum_{i=1}^{n}(Y_{obs_i} - Y_{sim_i})^* 100 \Big/ \sum_{i=1}^{n}(Y_{obs_i})\right] \tag{16}$$

The root mean square error observed standard deviation ratio (*RSR*) incorporates the benefits of error index statistics and includes a scaling factor of the standard deviation (Moriasi et al., 2007) of the observed series. The value varies from zero to a large positive value where zero indicates a perfect fit. The *RSR* is obtained as:

$$RSR = \left[\left[\sqrt{\textstyle\sum_{i=1}^{n}(Y_{obs_i} - Y_{sim_i})^2}\right] \Big/ \left[\sqrt{\textstyle\sum_{i=1}^{n}(Y_{obs_i} - \overline{Y}_{obs_i})^2}\right]\right] \tag{17}$$

6. Results and discussion

Several factors influence the quality of model simulations. In addition to the input dataset, the model structure and the intermediate output from the various model components also influence

the quality of the simulations. Although the model applied in this study is a black box, it is probable that the various components of the model and their outputs can be used to identify the main processes driving streamflow in the river. An attempt was therefore made to obtain a relationship between the intermediate output (e.g. variation of the radius of the hypersphere cluster in data space, location of the cluster centers, model coefficients and the degree of membership) and the known flow characteristics existing along the Letaba River. A discussion of the model results based on the model performance measures then follows.

6.1. Influence of the storage weirs and alluvial aquifer on the flows

The existence of storage weirs at various locations along the river particularly in the first and second river reach and also the underlying alluvial aquifer both impact the flows in the Letaba River. The extent of impact was found to decrease downstream. For instance, correlation values of rainfall to the outflow for each river reach were found to be 0.12, 0.15, and 0.12 for the first, second and third river reach, respectively (Fig. 3) whereas, the correlation between one day lag rainfall data to the outflow was found to be 0.48, 0.32, and 0.25 for the first, second and third river reach, respectively. This suggests that the storage weirs and alluvial aquifer retain the flow for one day. The one day lag of the inflows into the second river reach from the tributary, and also from the first river reach have slightly higher correlation values of 0.46 and 0.53, respectively, while values of 0.45 and 0.47 were obtained for the inflows without a one day lag (Fig. 3). Since there are no storage weirs and anthropogenic activities along the third river reach, this observation suggests that the alluvial aquifer may be responsible for this lag. Therefore, there is a significant impact of the storage weirs and the alluvial aquifer on the flows along the Letaba River.

Derived daily abstraction values and daily evaporation data have a less noticeable influence on the flows as compared to rainfall and inflows (Fig. 3). The known effect of evaporation on streamflow is not clearly manifested at the daily time step as the poor correlation values for all the three reaches suggest. However at a longer time scale (monthly or annual), higher correlations would be expected. As Fig. 3 also shows, poor correlations were obtained between streamflow and the abstraction estimates. It is worth noting, however, that the daily abstraction values were derived from monthly

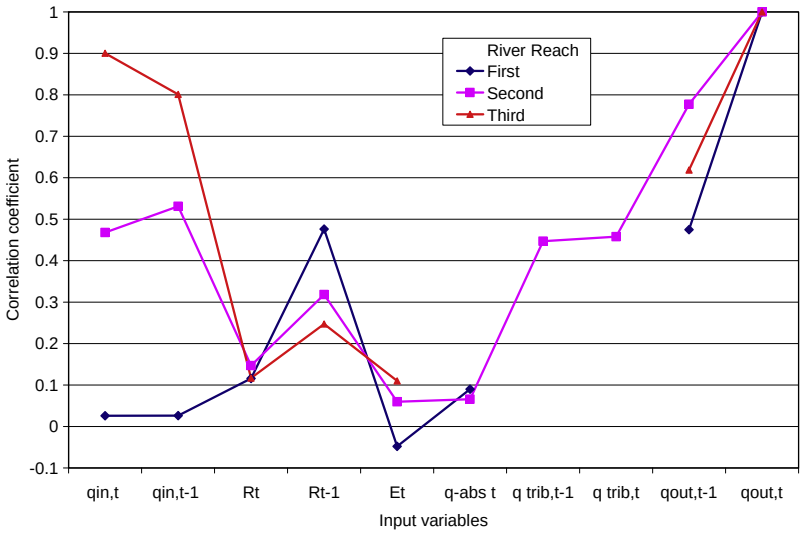


Fig. 3. Correlation coefficients for the input variables to the observed outflow.

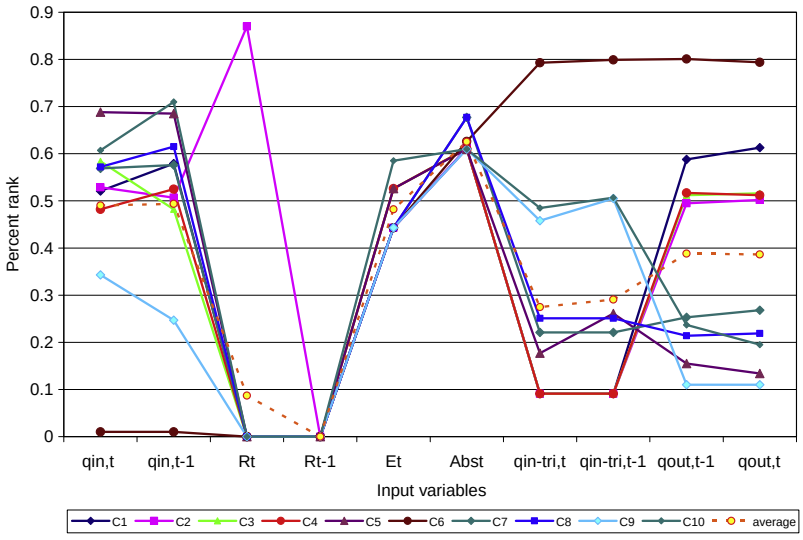


Fig. 4. Rank for the of the cluster center of each input variable for the second river reach.

Table 1
Variation of model coefficients with radius r_a for the second river reach.

r_a	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
1	0.004	1.321	0.323	0.622	−0.583	0.358	0.730	0.596
1.5	0.008	1.594	0.796	0.231	−0.507	0.387	0.630	0.547
2	0.008	1.131	0.910	0.608	−0.432	0.393	0.662	0.621
2.5	0.008	1.260	0.979	−0.169	−0.298	0.120	0.710	0.569
3	0.005	0.890	0.242	0.557	−0.180	0.447	0.536	0.598
3.5	0.003	1.703	0.506	0.387	−0.657	0.735	0.619	0.555
4	0.008	1.189	0.439	−0.117	−0.185	−0.195	0.697	0.609
4.5	0.010	1.110	0.685	0.470	−0.374	0.408	0.728	0.605
5	0.007	0.990	0.493	0.624	−0.377	0.430	0.807	0.589
5.5	0.005	1.100	0.821	0.722	−0.253	0.271	0.727	0.574
6	0.003	1.034	0.785	0.175	−0.157	0.125	0.518	0.618
6.5	0.008	1.570	0.527	0.906	−0.492	0.211	0.706	0.597
7	0.006	1.122	0.551	0.411	−0.188	0.318	0.681	0.557
7.5	0.000	1.401	0.936	0.733	−0.575	0.888	0.768	0.586
8	0.003	0.900	0.692	0.664	−0.204	0.379	0.641	0.597
8.5	0.006	0.920	0.160	0.323	−0.321	0.663	0.477	0.663
9	0.003	1.094	0.481	0.496	−0.441	0.794	0.704	0.605
9.5	0.002	1.502	0.377	0.744	−0.688	0.821	0.718	0.585
10	0.004	0.850	0.484	0.589	−0.265	0.354	0.720	0.576

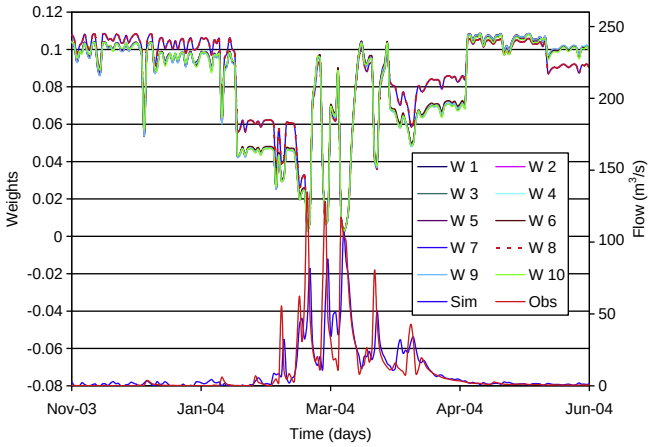


Fig. 5. Variation of the degree of membership for the second river reach.

abstraction estimates assuming a constant abstraction rate throughout the month. It is obvious that this does not reflect the reality but there are no daily measurements available.

The relative location of the ten cluster centers with respect to the data space was done by ranking the data for each variable and Fig. 4 shows a plot of the locations for the second river reach. The second river reach uses the largest number of model inputs and is therefore considered to be representative of the other river reaches. An average percent rank of 0.49 was observed for the inflow. In addition, three clusters were observed to have values less than the average value for the inflow into the second river reach for both with and without a day lag. For the outflow, half of the ten clusters where above and the other half are below the average percent rank of 0.39. With the exception of one cluster, the percent rank for rainfall inputs are all zero. The cluster locations for evaporation have a minimum rank of 0.44 and maximum of 0.59. The water abstraction estimates were found to have a minimum value of 0.61 and a maximum value of 0.68, which indicates that the abstraction values have a smaller range than the evaporation values. The contribution from the Letsitele River has a wider range of the rank, with minimum of 0.09 and maximum of 0.8 and an average of 0.27. The rank of the data cannot be strongly linked to the streamflow along the river since there are no observable trends.

6.2. Influence of radius of the hypersphere r_a and quash factor η to the model coefficients and the degree of membership

Each of the identified cluster centers have been ascribed to a linear function (Eq. (11) with indices set to 1), and the optimal coefficients of these functions were determined by the SCE-UA. Variation of the radius of the hypersphere and quash factor was done and there was no noticeable influence of variation of the radius of hypersphere to the model coefficients and the degree of membership. Although, the variations in the radius and quash factor was done for only the second river reach it was assumed that the characteristics of the second river reach are representative of both first and third river reach. Table 1 shows the average values of the coefficients for ten clusters for each value of radius of the hypersphere. The values in Table 1 show that variation in radius does not cause respective changes in the average values of the model coefficients. There is also no noticeable trend in coefficients

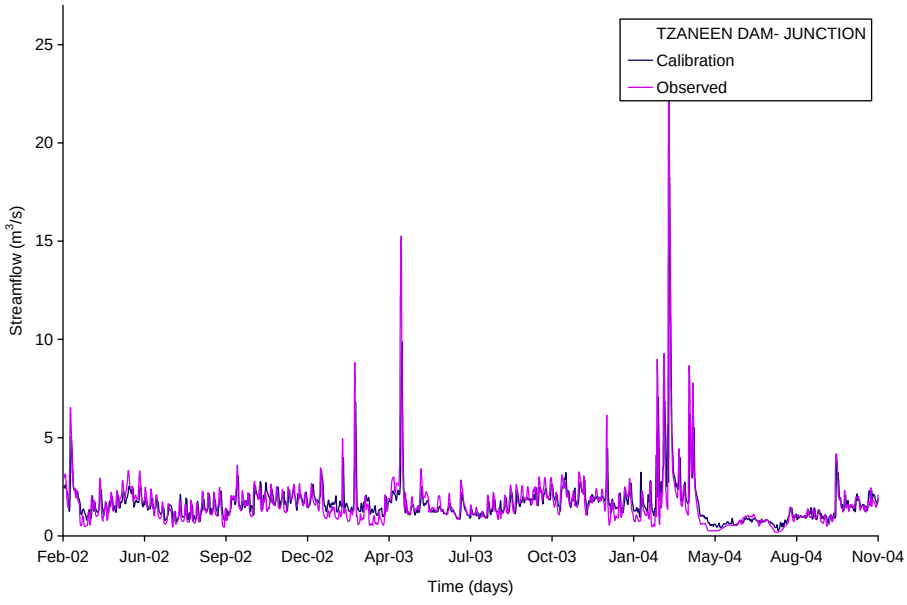


Fig. 6. Calibration and observed flow at the first river reach.

as a response to variation of the quash factor. A similar trend is observed in the variation of the weights derived from the degree of membership of each cluster as shown in Fig. 5. The degrees of membership do not vary significantly among the clusters suggesting that all the clusters work together in obtaining a particular simulated value for all the components of the streamflow hydrograph. The variation of the weights with the simulated flow is more pronounced in cases where there is a sudden change in flow with a sudden increase in flow corresponding to a uniform reduction in the degrees of membership for all the clusters.

The analysis in this Section and the previous Section indicates that there were no clusters that were specifically suited for simulating particular components of the flow hydrograph (e.g. high flows or low flows) and the model can be reasonably considered a pure black box.

6.3. Simulation results and model performance

The model was observed to perform better in simulating flows in the lower river reaches as compared to the upper reaches. The results from the calibration and verification of the model for all the three reaches are shown in Figs. 6–11. Simulations (calibration and verification) were also conducted with the two upstream river reaches combined and the results are presented in Figs. 12 and Fig. 13 while Figs. 14 and 15 show the results of the calibration and verification when all the three reaches are linked.

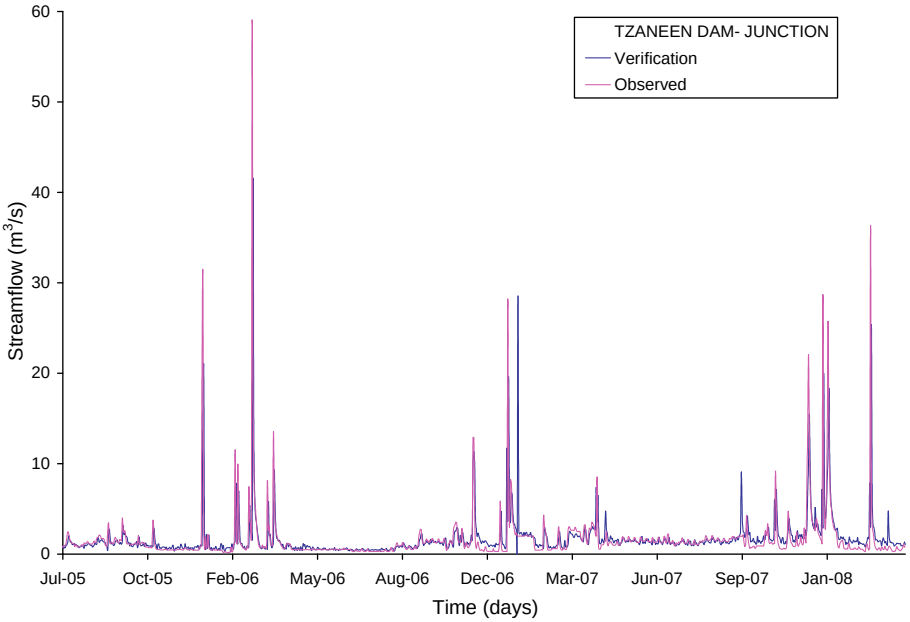


Fig. 7. Verification and observed flow at the first river reach.

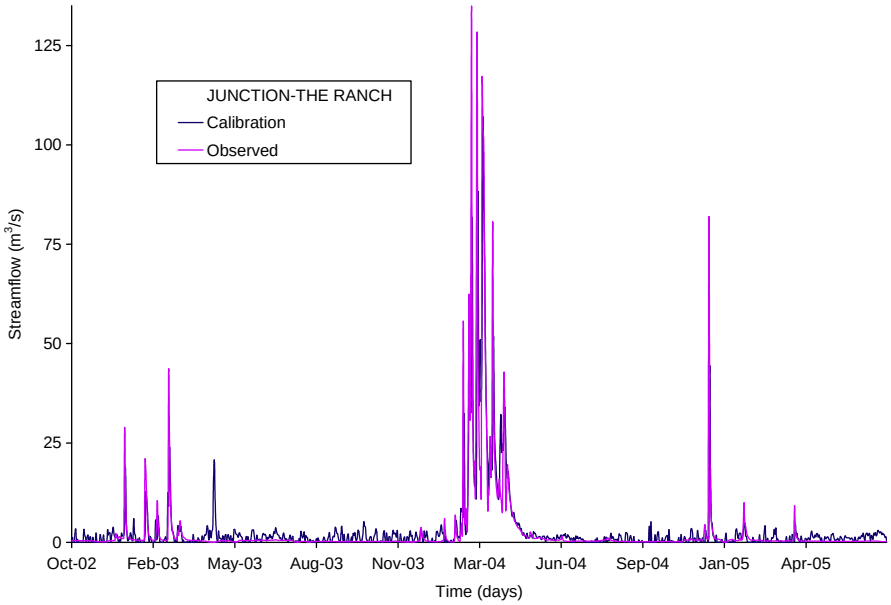


Fig. 8. Calibration and observed flows at the second river reach.

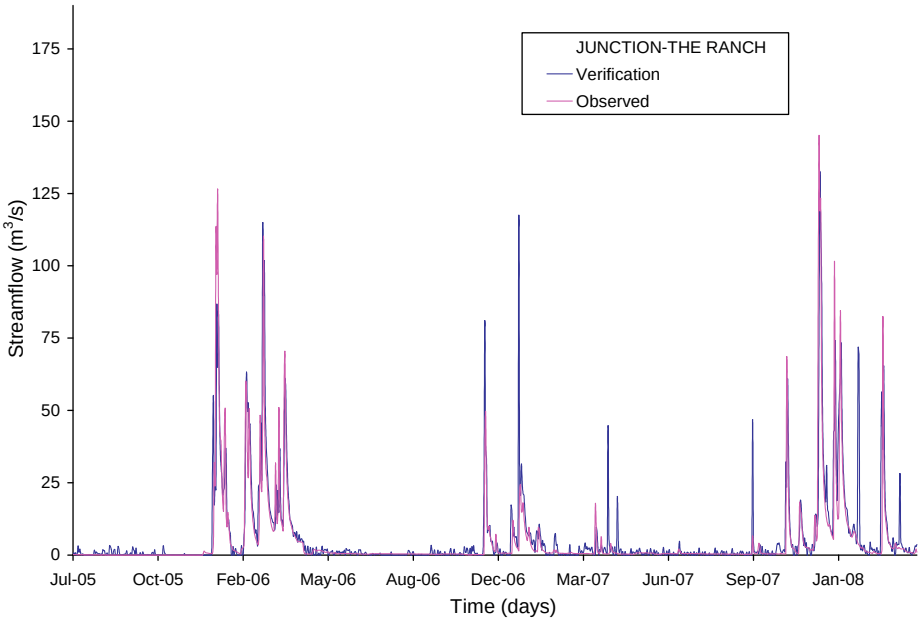


Fig. 9. Verification and observed flows at the second river reach.

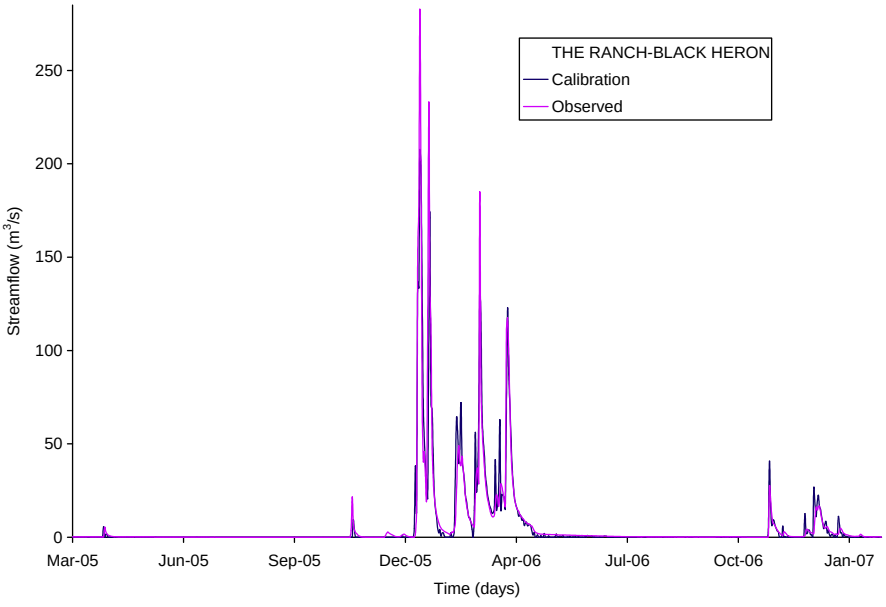


Fig. 10. Calibration and observed flows at the third river reach.

The values of *CCoef* during the calibration period ranged between 0.720 and 0.923 with the maximum value being obtained for the third river reach and the lowest value obtained for the first reach (Table 2). During verification, model performance for the first reach was unsatisfactory with a value of 0.470 for the *CCoef* while the second and third reach had values of 0.790 and 0.952, respectively. Model performance when the reaches were linked as earlier described yielded values of 0.740 and 0.813 for the combinations of the two upper and all three reaches, respectively. With the exception of the third river reach, the model indicates better values of *CCoef* during the calibration period than the verification period a factor attributed to the model's inability to attain a general suitable parameter set for both the hydrological and non-hydrological processes existing in the reaches. However, the model managed to obtain better values for the third river reach – a fact that indicates that the alluvial aquifer in this reach was modelled reasonably well.

The value of the *PBIAS* suggests that the model generally overestimates, although it was observed to underestimate flow at the Black Heron for both calibration and verification of the linked and individual reaches. The values of the *PBIAS* were observed to be between −13.95% and 8.40% for the second individual and

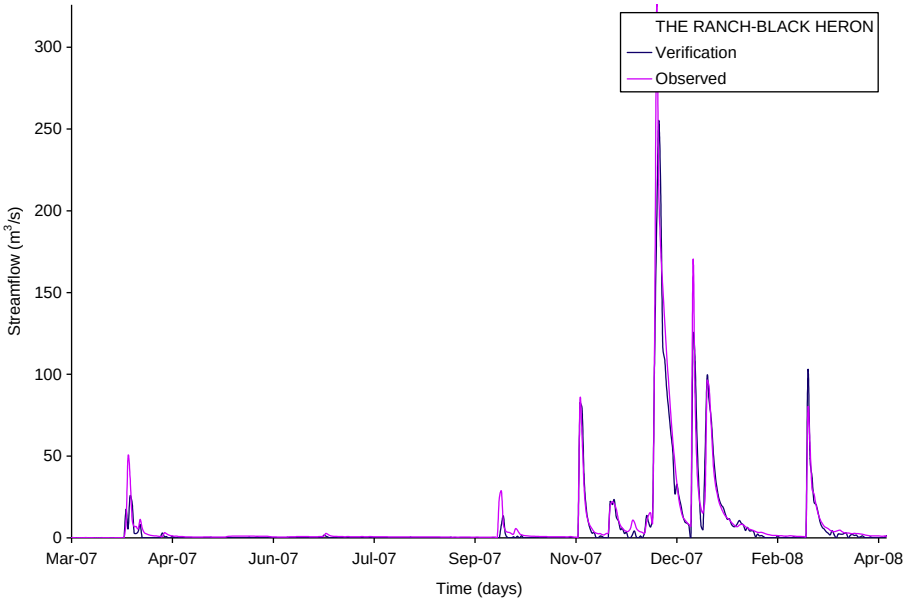


Fig. 11. Verification and observed flows at the third river reach.

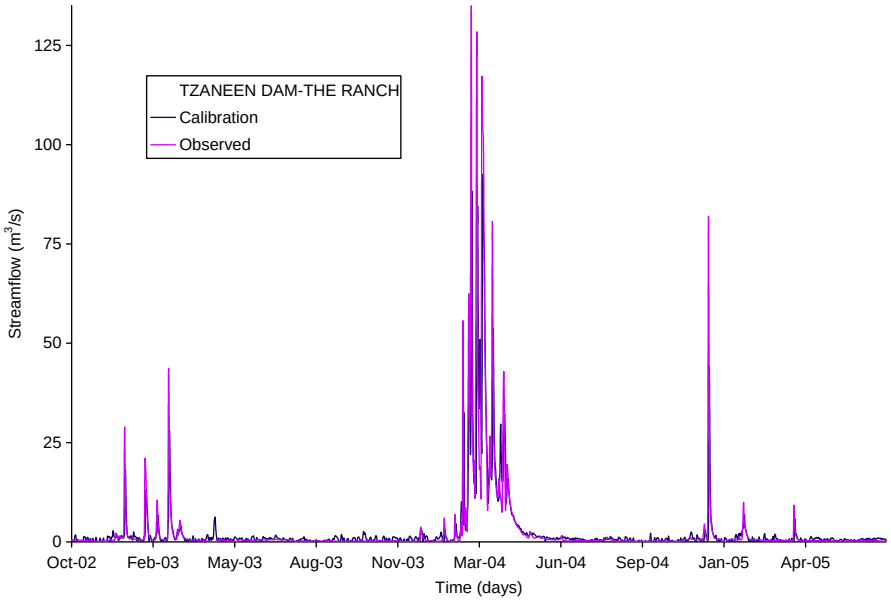


Fig. 12. Observed and calibration of the combined flows at the second river reach.

linked river reach, respectively. For the all the reaches, the best values of the *PBIAS* for both calibration and verification were −3.47 and 3.96 being for the first and third river reach. The *NSE* values obtained suggest that the general model performance is better for the lower reach that the upper reach. For the calibration processes the *NSE* values ranged from 0.507 (for the first reach) to 0.851 (for the third river reach). In verification, the model obtained a low value of 0.077 for the first river reach while all other values ranged between 0.484 and 0.903. In general, the performance based on the *NSE* indicates a satisfactory model performance and the poor values obtained for the first river reach may be attributed to the severe impact on flows by the hydraulic structures (e.g. storage weirs) and water abstraction. Considering all river reaches, the values of the root mean square error and observed standard deviation ratio ranged between 0.21 and 0.83 for the reaches, with the best (small) values obtained for the lower reaches. As expected, the performance of the linked models is influenced by the considerable human impacts in the first and second river reach and is consequently not as good as for the simulations of the individual reaches.

As Table 2 shows, with the exception of the first river reach, the correlation coefficients in calibration and validation match closely and are considered acceptable. Although only six years' data length have been used, the model performance is considered satisfactory for a river highly impacted with human activity. It is also quite possible that the approach can still perform satisfactorily with data length less than six years but additional studies would be needed to find this out. In addition, it is probable that the modelling could

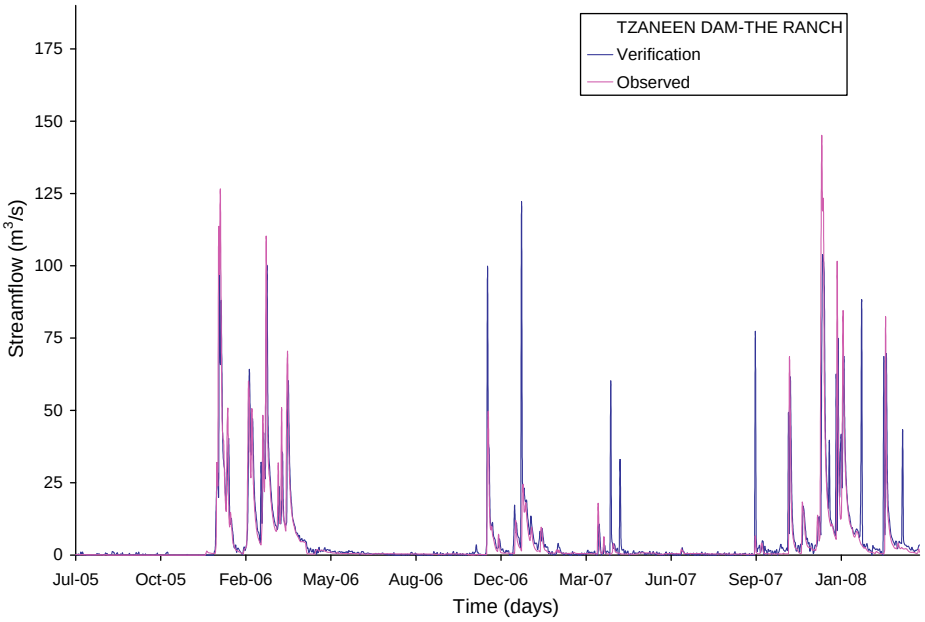


Fig. 13. Observed and verification of the combined flows at the second river reach.

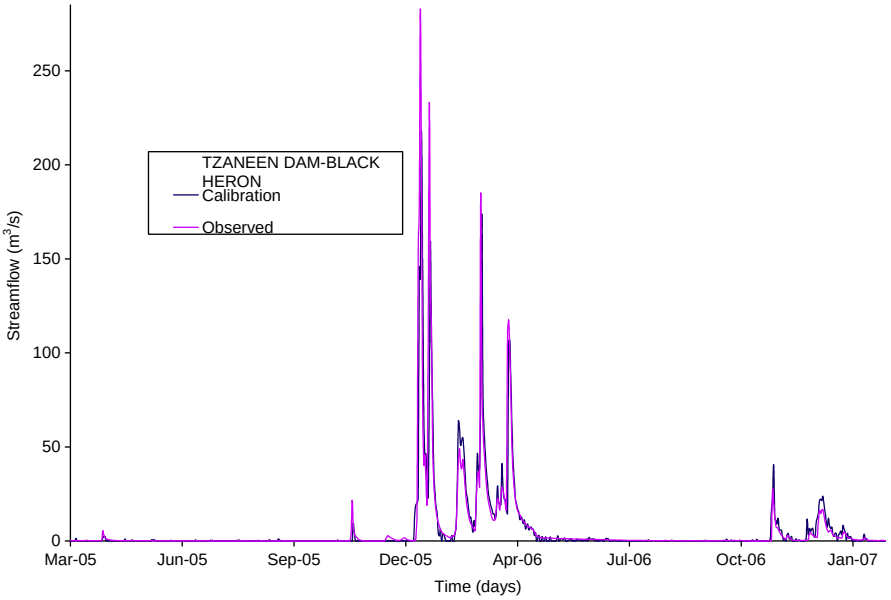


Fig. 14. Observed and calibration of the combined flows at the third river reach.

be further improved if a conceptual incorporation of the known hydrological processes and human activities is included in the model. This is the subject of further work.

7. Conclusions and recommendations

A Takagi–Sugeno fuzzy inference system model has been developed and applied in the simulation of daily streamflows of three reaches of the Letaba River in South Africa. The Takagi–Sugeno fuzzy inference system was selected owing to its ability in dealing with inadequate, imprecise and complex data that characterises Letaba River. The River is highly impacted by irregular 'rule of thumb' based dam and storage weir operation and intermittent water abstractions. Analysis of the various components of the modelling including (i) the location of cluster centers, (ii) the coefficients relating the modelled flow to the inputs, and (iii) degrees of membership, indicates that all the clusters and their associated equations work together to produce the simulations for all parts of the flow hydrograph. There being no evidence connecting any cluster to specific parts of the hydrograph more than any other, the modelling is considered a pure black box.

The model performance is considered satisfactory with average Nash–Sutcliffe efficiency values of 0.66 and 0.70 for calibration and verification, respectively. Lower values of Nash–Sutcliffe efficiency were obtained for the first river reach as a result of the more complex flow pattern resulting from the irregular operation of storage weirs and intermittent abstractions. The model estimates are generally lower than the observed values during wet seasons and are

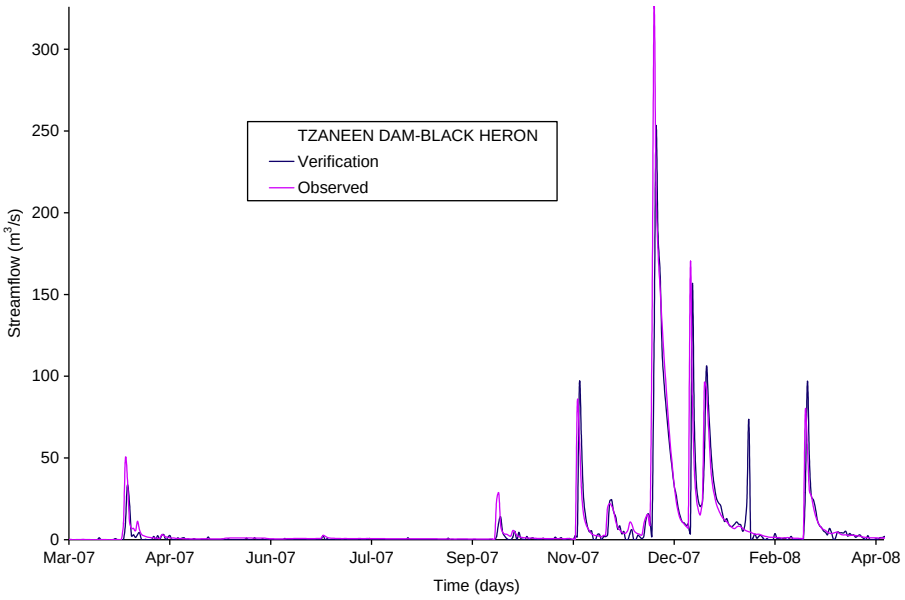


Fig. 15. Observed and verification of the combined flows at the third river reach.

Table 2
Model performance results.

	River reach	Calibration				Verification			
		CCoef	*NSE*	*PBIAS*	*RSR*	*CCoef*	*NSE*	*PBIAS*	*RSR*
River reach									
Tzaneen Dam-Junction	1st	0.720	0.507	−8.9	0.697	0.470	0.077	−3.47	0.83
Junction-Ranch	2nd	0.795	0.631	−9.43	0.595	0.790	0.557	−13.95	0.52
Ranch-Black Heron	3rd	0.923	0.851	3.96	0.382	0.952	0.903	6.95	0.21
Connected									
Tzaneen Dam-Ranch	1st to 2nd	0.764	0.583	8.40	0.633	0.740	0.484	−9.8	0.57
Tzaneen Dam-Black Heron	1st to 3rd	0.847	0.716	4.04	0.527	0.813	0.656	6.66	0.40

slightly higher during dry seasons. Low percent bias values are obtained in both calibration and verification indicating the model could have potential for modelling in situations where mass balance is the most important consideration. More of the bias values are positive but the overall average value is slightly negative.

Since the analysis here revealed the model as purely a black box; future work will focus on incorporating the known and perceived hydrological processes and human activities in the model.

Acknowledgements

The authors acknowledge and appreciate the invaluable contribution of the following organisations towards this study: The National Research Foundation (NRF) of South Africa, Mbeya Institute of Science and Technology (MIST), Department of Water Affairs and Forestry (DWAF) of South Africa, South Africa Weather Services (SAWS) and the Kruger National Park (KNP) for providing financial support and information. The comments of an earlier submission of the paper from reviewers are also acknowledged.

References

Altunkaynak, A., Sen, Z., 2007. Fuzzy logic model of lake water fluctuations in Lake Van, Turkey. Theoretical and Applied Climatology 90 (3–4), 227–233.
Chiu, S.L., 1994. Fuzzy model identification based on cluster estimation. Journal of Intelligent and Fuzzy Systems 2, 267–278.
Demirli, K., Cheng, S.X., Muthukumaran, P., 2003. Substrative clustering based modelling for job sequencing with parametric search. Fuzzy Sets and Systems 137, 235–270.
Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization of conceptual rainfall–runoff models. Water Resources Research 28 (4), 1015–1031.
Department of Water Affairs and Forestry, 2004. Internal Strategic Perspective: Luvuvhu/Letaba WMA. Report No. P WMA 02/000/00/0304.
Department of Water Affairs and Forestry, 2006. Letaba River System Annual Operating Analysis. Report No. WMA 02/000/00/0406.
Gőrgens, A.H.M., 1983. Conceptual modelling of the rainfall–runoff process in semi-arid catchments in South Africa. PhD Thesis, University of the Witwatersrand, Johannesburg, South Africa.
Hughes, D.A., 2004. Incorporating groundwater recharge and discharge functions into an existing monthly rainfall–runoff model. Hydrological Sciences—Journal 49 (2), 297–311.
Hundecha, Y., Bardossy, A., Theisen, H., 2001. Development of a fuzzy logic-based rainfall–runoff model. Hydrological Sciences—Journal 46 (3), 363–376.
Jacquin, A.P., Shamseldin, A.S., 2006. Development of rainfall–runoff models using Takagi–Seguno fuzzy inference system. Journal of Hydrology 329, 154–173.
Katambara, Z., Ndiritu, J., 2007. Developing a surface water–groundwater interaction model for Letaba River system in South Africa. In: Proceeding of the Eighth WATERNET Conference, Lusaka, Zambia.
Lohani, A.K., Goel, N.K., Bhatia, K.K.S., 2005. Development of a fuzzy logic based real time flood forecasting system for river Narmada in Central India. In: Preceding of an International Conference on Innovation Advances and Implementation of Flood Forecasting Technology, TromsØ, Norway.
Lohani, A.K., Goel, N.K., Bhatia, K.K.S., 2006. Takagi–Sugeno fuzzy inference system for modelling stage–discharge relationship. Journal of Hydrology 331, 146–160.
Mamdani, E.H., Assilian, S., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man–Machine Studies 7 (1), 1–13.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. American Society of Agricultural and Biological Engineers 50 (3), 885–900.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models: part 1. A discussion of principles. Journal of Hydrology 10 (3), 282–290.

Ndiritu, J.G., Daniell, T.M., 2001. An improved genetic algorithm for rainfall-runoff model calibration and function optimization. Mathematical and Computer Modelling 33 (6–7), 695–706.

Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its application to modelling and control. IEEE Transactions on System, Man and Cybernetics 15 (91), 116–132.

Tayfur, G., Ozdemir, S., Singh, V.P., 2003. Fuzzy logic algorithm for runoff-induced sediment transport from bare soils surface. Advances in Water Resources 26, 1249–1256.

Vernieuwea, H., Georgievab, O., De Baetsa, B., Pauwelsc, V.R.N., Verhoestc, N.E.C., De Trochc, F.P., 2005. Comparison of data-driven Takagi–Sugeno models of rainfall–discharge dynamics. Journal of Hydrology 302, 173–186.

Xiong, L., Shamseldin, A.Y., O'Connor, K.M., 2001. A non-linear combination of the forecast of rainfall–runoff models by the first-order Takagi–Seguno fuzzy system. Journal of Hydrology 245, 196–217.

Yager, R., Filev, D., 1994. Generation of fuzzy rules by mountain clustering. Journal of Intelligent and Fuzzy Systems 2 (3), 209–219.

Zimmermann, H.J., 2001. Fuzzy Set Theory and its Applications. Kluwer Academic Publisher, USA. ISBN 0-7923-7435-5.