Self-organising map rainfall-runoff multivariate modelling for runoff reconstruction in inadequately gauged basins

Adebayo J. Adeloye and Rabee Rustum

ABSTRACT

Water resources assessment activities in inadequately gauged basins are often significantly constrained due to the insufficiency or total lack of hydro-meteorological data, resulting in huge uncertainties and ineffectual performance of water management schemes. In this study, a new methodology of rainfall-runoff modelling using the powerful clustering capability of the self-organising map (SOM), unsupervised artificial neural networks, is proposed as a viable approach for harnessing the multivariate correlation between the typically long record rainfall and short record runoff in such basins. The methodology was applied to the inadequately gauged Osun basin in southwest Nigeria for the sole purpose of extending the available runoff records and, through that, reducing water resources planning uncertainty associated with the use of short runoff data records. The extended runoff records were then analysed to determine possible abstractions from the main river source at different exceedance probabilities. This study demonstrates the successful use of emerging tools to overcome practical problems in sparsely gauged basins.

Key words | hydrological data, Nigeria, rainfall-runoff modelling, self-organising map (SOM), water resources assessment

Adebayo J. Adeloye (corresponding author)
School of the Built-Environment,
Heriot-Watt University,
Riccarton,
Edinburgh, EH14 4AS,
UK
E-mail: *a.j.adeloye@hw.ac.uk*

Rabee Rustum
School of the Built Environment,
Heriot-Watt University,
Dubai International Academic City,
Dubai,
UAE

INTRODUCTION

Water resources planning activities are often hampered in poorly gauged basins such as those in developing countries and elsewhere, resulting in huge uncertainties in planning decisions (Adeloye 1990, 1996). This is typified by the situation of Osun river basin in southwest Nigeria where the runoff data are either unavailable or when available, they are too short and riddled with numerous gaps and erroneous readings, and are often difficult to retrieve because of the archaic storage and retrieval system in operation. Given the general paucity of data and other information for the Osun basin, the only realistic approach for modelling the rainfall-runoff transformation is by regression-correlation analysis.

Typically, regression analysis is usually univariate, in which there is only one dependent variable, predicted using one or more independent (or predictor) variables (Adeloye 2009). This implies that the predictor variables are totally independent of themselves, although they should exhibit strong correlations with the dependent variable. Where this is the case, the regression approach has been used in many hydrological analyses principally to predict runoff statistics such as the mean, median, low flow quantiles, flood quantiles, etc. (Wharton *et al.* 1989; IH 1999; Vogel & Sankarasubramanian 2000; Mohamoud & Parmar 2006; Laaha & Bloschl 2007; Mohamoud 2008). Often, however, correlations between the predictor variables cannot be avoided; indeed, advantage can be taken of such co-linearity to improve the predictability of the dependent variable. Such an exercise which explicitly accommodates co-linearity between predictor variables is known as a multivariate approach or regression. In other words, it is possible to take advantage of the different correlations between the available rainfall and runoff stations to improve the predictability of the runoff and achieve its record extension using the usually much longer rainfall records.

doi: 10.2166/nh.2012.017

The multivariate approach used in this study was based on the self-organising map (SOM), which is an unsupervised form of artificial neural networks (ANNs) developed by Kohonen (Kohonen *et al.* 1996). In general, ANNs are able to map any non-linear relationship, no matter how complex, without the need to specify explicitly the mathematical form of the model. Although ANNs have been widely applied in rainfall-runoff modelling (see e.g. Smith & Eli 1995; Minns & Hall 1996; Shamseldin 1997; Tokar & Johonson 1999; Chang *et al.* 2004), these have been limited to supervised, feed-forward back propagation types. However, a problem with feed-forward, back-propagation ANNs is that they can be affected by missing values and outliers. Indeed where such noise is available in the data, feed forward ANNs have been known to give unrealistic results (Parasuraman *et al.* 2006; Rustum *et al.* 2007). On the contrary, unsupervised ANNs, typified by the SOM, while they have distinct input and output layers, have no specific input or output variables, since all the variables in the input vector are also contained in each node of the output layer.

The SOM approach possesses a powerful clustering capability which can reduce a multi-dimensional input data array into a two-dimensional array of features or best matching units (BMUs). Because it clusters, its implementation is unhindered by missing values. The prediction of elements of the multi-dimension input array is then determined by identifying its BMU. This approach was used to reconstruct and extend monthly rainfall and runoff data for stations within the lower Osun basin in southwest Nigeria.

Despite the paucity of data, the Osun basin has many water development projects and the latest development being planned that instigated the current study was a pumped storage water supply reservoir at Igbonla close to its mouth. However, there were no runoff data whatsoever at the proposed abstraction site with which to determine abstracted flows and their probabilities. Rather, there were two upstream stations with short, incomplete runoff records and a number of rainfall stations with much longer records, albeit with numerous missing values. The SOM clustering was thus used to extend the short runoff records at the two upstream stations, which were then used to derive the needed runoff record at the proposed abstraction site downstream.

Thus, the aim of the work is to demonstrate the use of the SOM as a viable rainfall-runoff modelling approach for runoff extension in data sparse situations. The objectives are to:

- Train and validate SOM models using existing rainfall-runoff data for Osun basin, Nigeria;
- use the validated models to fill-in the missing data in all the records;
- use the SOM model to extend the available Osun runoff records;
- analyse the extended runoff data to determine abstraction quantities and their associated probabilities and make recommendations.

In the next section, further details about the SOM modelling are given. This is then followed by the methodology applied for the Osun catchment study. Finally, the results are presented and discussed.

SOM MODELLING

Basics of the self-organising map

The SOM (also called feature map or Kohonen map) is one of the most widely used ANNs algorithms (Kohonen *et al.* 1996). It is usually presented as a dimensional grid or map whose units (nodes or neurons) become tuned to different input data patterns. Its algorithms are based on unsupervised competitive learning, which means that training is entirely data driven and the neurons or nodes on the map compete with each other (Alhoniemi *et al.* 1999; Rivera *et al.* 2011).

The principal goal of the SOM is to transform an incoming signal pattern of arbitrary dimension into a two-dimensional discrete map. It involves clustering the input patterns in such a way that similar patterns are represented by the same output neurons, or by one of its neighbours (Back *et al.* 1998). In this way, the SOM can be viewed as a tool for reducing the amount of data by clustering, thus converting complex, nonlinear statistical relationship between high dimensional data into a simple relationship on a low (usually two) dimensional display (Kohonen *et al.* 1996). This mapping preserves the most important topological and metric relationship of the original data elements, implying that not much information is lost during the mapping

making the technique a very good tool for prediction. Problems do arise though if the tool is used for extrapolation, i.e. for predicting values outside the range used for the clustering, because as with most data-driven techniques, the SOM is a very poor extrapolator (Adeloye *et al.* 2011).

The SOM consists of two layers: the multi-dimensional input layer and the competitive or output layer; both of these layers are fully interconnected as illustrated in Figure 1. The output layer consists of M neurons arranged in a two-dimensional grid of nodes. Each node or neuron i ($i = 1, 2, \ldots, M$) is represented by an n-dimensional weight or reference vector $\mathbf{W_i} = [w_{i1},\ldots, w_{in}]$, where n is the dimension of each input vector, i.e. the maximum number of variables in the input vector. Garcia & Gonzalez (2004) offer guidance on determining the optimum number of neurons, which is:

$$M = 5\sqrt{N} \tag{1}$$

where N is the total number of data samples. Once M is known, the number of rows and columns in the KSOM can be determined. A guideline by Garcia & Gonzalez (2004) is that:

$$\frac{l_1}{l_2} = \sqrt{\frac{e_1}{e_2}} \tag{2}$$

where l_1 and l_2 are the number of rows and columns, respectively, e_1 is the biggest eigenvalue of the training data set and e_2 is the second biggest eigenvalue.

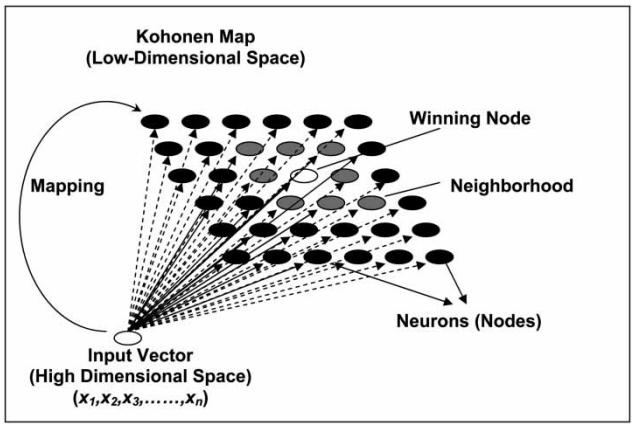

Figure 1 | Illustration of the winning node and its neighbourhood in the Kohonen self-organising map.

Training the SOM

The multi-dimensional input data are first standardised by deducting the mean and then dividing the result by the standard deviation. To start the training, the neurons in the output layer are seeded with randomly generated values. A standardised input vector is chosen at random and presented to each of the individual neurons of the SOM for comparison with their code vectors in order to identify the code vector most similar to the presented input vector. The identification uses the Euclidian distance, which is defined as:

$$D_i = \sqrt{\sum_{j=1}^{n} m_j (x_j - w_{ij})^2};\ i = 1, 2, \ldots, M \tag{3}$$

where D_i is the Euclidian distance between the input vector and the code vector i; x_j is the j-th element of the current input vector; w_{ij} is the j-th element of the code vector i; n is the dimension of the input vector; and m_j is the so-called 'mask' which is used to include in ($m_j = 1$) or exclude from ($m_j = 0$) the calculation of the Euclidian distance, the contribution of a given element x_j of the input vector. This is very useful where the input vector contains missing elements because all that needs to be done is to set the mask (m_j) to zero for such elements. In this way, the SOM is able to handle missing values in the input vector without any problem.

The neuron whose vector most closely matches the input data vector (i.e. for which the D_i is minimum) is chosen as a winning node or the BMU as indicated in Figure 1. The code vectors of this winning node and those of its adjacent neurons are then adjusted to match the input data using Equation (4), thus bringing the code vectors further into agreement with the input vector (Vesanto *et al.* 2000).

$$w_i(t+1) = w_i(t) + \alpha(t) h_{ci}(t)[x(t) - w_i(t)] \tag{4}$$

where t denotes time, $\alpha(t)$ is the learning rate at t, $h_{ci}(t)$ is the neighbourhood function centred in the winner unit c at time t and all the other variables are as defined previously. In this manner each node in the map internally develops the ability to recognise input vectors similar to itself. This characteristic is referred to as self-organising, because no external information is supplied to lead to a classification (Penn 2005).

The process of comparison and adjustment continues until the optimal number of iterations is reached or the

specified error criteria are attained. Both the learning rate and the neighbourhood function (see Equation (4)) affect the learning effectiveness of the KSOM and must be chosen carefully. In particular, the learning rate decreases monotonically with increased number of iterations as in Equation (5):

$$\alpha(t) = \alpha_0(0.005/\alpha_0)^{t/T} \tag{5}$$

where α_0 is the initial learning rate and T is the training length (Vesanto *et al.* 2000), thus forcing the weight vector to converge. The neighbourhood function is normally chosen to be Gaussian centred in the winner unit c, such that:

$$h_{ci}(t) = \exp^{\left(-\|r_c - r_i\|^2/(2\sigma^2(t))\right)} \tag{6}$$

where r_c and r_i are the positions of nodes c and i on the SOM grid and $\sigma(t)$ is the neighbourhood radius. Like the learning rate $\alpha(t)$, $\sigma(t)$ also decreases monotonically as the number of iterations increases.

The quality of the trained SOM is measured by the total average quantisation error and total topological error. The quantisation error is:

$$q_e = \frac{1}{N}\sum_{i=1}^{N} \|X_i - W_c\| \tag{7}$$

where q_e is the quantisation error, $\boldsymbol{X_i}$ is the i-th data sample or vector, $\boldsymbol{W_c}$ is the prototype vector of the BMU for $\boldsymbol{X_i}$ and $\|.\|$ denotes the Euclidian distance (Equation (3)). The topological error is:

$$t_e = \frac{1}{N}\sum_{i=1}^{N} u(X_i) \tag{8}$$

where u_i is a binary integer such that it is equal to 1 if the first and second BMUs for X_i are not adjacent units; otherwise it is zero.

The SOM can be used for many practical tasks, such as the reduction of the amount of training data for model identification, nonlinear interpolation and extrapolation (i.e. prediction), generalisation and compression of information for easy transmission (Kangas & Simula 1995; Kohonen *et al.* 1996; Parasuraman *et al.* 2006; Tananaki *et al.* 2007; Chang *et al.* 2010; Adeloye *et al.* 2011; Rivera *et al.* 2011). The application of the SOM for prediction purposes is illustrated in Figure 2 (see also Rustum & Adeloye 2007; Rustum *et al.* 2008). First, the model is trained using the available data set. Then the depleted vector, i.e. with the predictand either missing or deliberately removed, is presented to the SOM to identify its BMU using the computed Di (Equation (3)). The values for the missing variables are then obtained as their corresponding values in the BMU.

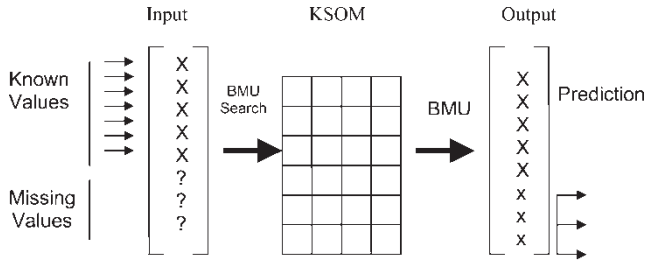

Figure 2 | Schematic illustration of prediction of missing components of an input data vector using the KSOM. BMU is the best matching unit (adapted from Rustum *et al.* 2008).

METHODOLOGY

Case study and data

The Osun basin in south west Nigeria is highly developed with many water abstractions and impounding schemes as shown in Figure 3. Together with the neighbouring Ogun river basin, the Osun basin is managed by the Ogun-Osun River Basin Authority, one of the 11 statutory river basin organisations in Nigeria. Of the existing impoundments, the Asejire dam is the most prominent supplying water to over 5 million people in Ibadan city (Tokun & Adeloye 2005). Most of the other impounding schemes within the basin are still in the planning stage (see Figure 3(a)) and may never be realised. However, the latest proposed development on the river being given active consideration and which is the subject of this study is an off-line, pumped storage scheme at Igbonla, just before the river discharges into the Lagos lagoon. To achieve this, probabilities associated with different pumped abstraction scenarios as shown in Table 1 for filling the reservoir will need to be determined. Without runoff data at Igbonla, however, this proved an impossible task.

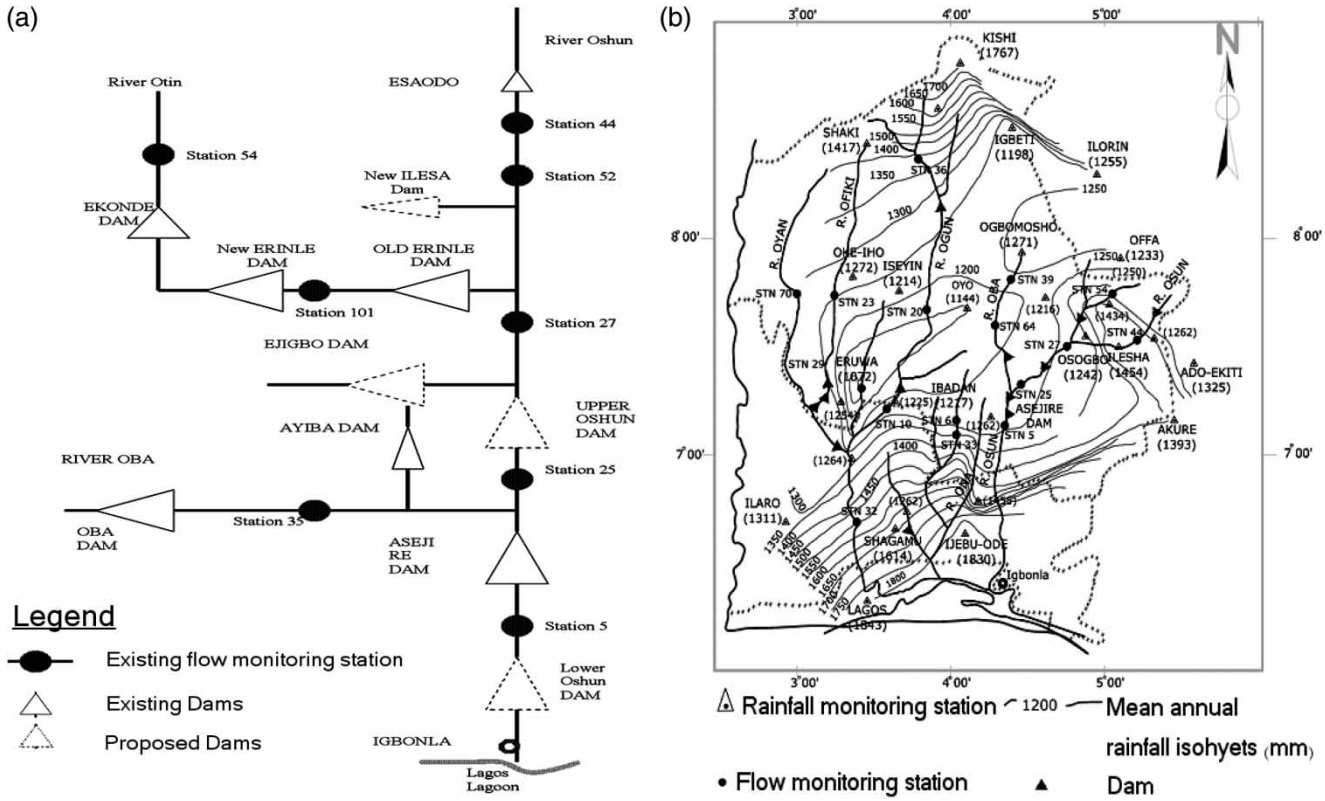

Figure 3 | The Osun basin in southwest Nigeria showing (a) simplified schematic of rivers, dams and gauging stations and (b) map of isohyets, and location of main towns, dams, runoff and rainfall stations.

Table 1 | Proposed abstraction scenarios from River Osun at Igbonla

Target year	Desired daily pumping rate (MCM day^{-1})
2011	0.63
2015	1.20
2020	1.82

Rainfall data are widely available within the basin, including: monthly data for Osogbo (1970–1983); Lagos Marina, Ijebu-Ode and Abeokuta (1987–2006, albeit with up to 40% of the monthly values missing at Ijebu-Ode and Abeokuta); and Lagos Island, Ikeja and Ibadan (1941–2006). Thus potentially, it could be assumed that rainfall data are available for 1941–2006, i.e. 66 years which can be used as the basis for extending/reconstructing monthly runoff records at desired locations in the basin.

Stations with runoff records in the basin include Station 44 and Station 25 as shown in Figure 3; all the other gauging sites in Figure 3 have no time series runoff records. The available monthly record at Station 44 is only 10 years (1970–1979) while that at station Station 25 is 11 years (1973–1983). Apart from the fact that these runoff monitoring stations are far upstream of the Igbonla site, the available record lengths are far too short to carry out any meaningful statistical analysis for the purpose of reservoir planning without incurring huge uncertainties (Adeloye 1990, 1996). To reduce these uncertainties, the runoff records will need to be extended and the way this was achieved in the current situation was by harnessing the multivariate correlations between the rainfall and runoff using the SOM. Summary statistics of the runoff at some of the gauging stations are shown in Table 2. Apart from Stations 44 and 25 which have time series data, the basis of the min, max and average values in Table 2 is unknown.

SOM modelling

The SOM analysis used monthly rainfall at four stations, namely Ibadan (No. 65208), Ikeja (No. 65201), Lagos

Table 2 | Characteristics of some of the runoff stations in Osun basin

Station number	Min MCM year^{-1}	Max MCM year^{-1}	Average MCM year^{-1}	Catchment area (km^2)
44	90.52	406.44	254.71	295.50
25	366.20	2,573.39	1,179.52	4,325.30
52	61.86	481.57	293.92	1,554.00
54	21.79	255.30	153.38	194.35
35	5.58	274.01	172.90	1,320.95
5	197.40	2,032.50	864.40	7,174.30

Island (No. 65203), and Osogbo (No. 65215) and the two runoff stations (25 and 44) for the multivariate, KSOM modelling. This led to the reconstruction of runoff at Stations 25 and 44 for the period 1941–2006. The SOM toolbox for Matlab 5 was used for this case study. The toolbox was developed by the SOM team at the Laboratory of Computer and Information Science, Helsinki University of Technology, Finland (http://www.cis.hut.fi).

Runoff reconstruction at Igbonla

Simple area-ratio scaling (Loucks *et al.* 1981) was used for transposing runoff data to Igbonla. The use of area-ratio scaling for transferring flow from a gauged site to an ungauged one is widely practised in the literature (see e.g. McCuen & Levy 2000; Vogel & Sankarasubramanian 2000; Laaha & Bloschl 2005; Mohamoud & Parmar 2006; Mohamoud 2008). Mohamoud (2008) recommends using a simple, linear scaling factor given by $F = A_u/A_g$, where A_u and A_g are the catchment areas at the ungauged (or destination) and gauged (or source) sites, respectively, if $0.5 \leq F^{-1} \leq 1.5$. According to Loucks *et al.* (1981), a simple scaling factor will give reliable results if the two sites are located on a catchment exhibiting increasing runoff in its downstream direction. Multiple donor (or source) sites could also be used for linear scaling as described by Loucks *et al.* (1981) but this will require information on the relative weights for the donor sites flow in determining the flow at the ungauged site. Where the conditions for linear scaling are not met or where there is evidence that the method does not give reliable results, non-linear scaling approaches are recommended. Examples of non-linear approaches include using the β-th power of F, i.e. F^β where β is an empirical constant (McCuen & Levy 2000; Vogel & Sankarasubramanian 2000), or using $\tan(F)$ or $\arctan(F)$ (Mohamoud 2008). To improve the performance of non-linear scaling, a two-stage approach is recommended by Mohamoud (2008) involving first transferring the gauged flow to an intermediate station and then transferring the flow at the intermediate site to the target site.

Since Station 25 is closer to Igbonla, its reconstructed (1941–2006) data formed the basis of the transposition using the respective catchment areas, i.e. 8,174 km^2 at Igbonla and 4,325 km^2 at Station 25. The corresponding value of F^{-1} for these catchment areas is 0.53, which falls within the recommended range for use of the simple linear scaling method. Additionally, there is evidence of increasing wetness from upstream to downstream in the Osun basin. For example, in a previous study (Adeloye & Rustum 2011), it was established that the rainfall at Ibadan was generally lower than Ikeja rainfall downstream which was in turn lower than rainfall at the maritime Lagos Island further downstream. The isohyets in Figure 3(b) also confirm the increasing wetness in the downstream direction within the basin. Consequently, the Osun catchment meets this criterion of progressively increasing wetness downstream to make the use of the linear area-ratio scaling approach justified. It might be argued that since Station 44 also had runoff, both Stations 25 and 44 could have been used as donor sites. However, to achieve this, the weights corresponding to each station must be objectively determined which is problematic.

Other catchment areas that went into the calculations are detailed in Table 2. In scaling, allowance was also made for the Asejire dam (catchment area = 5,646 km^2) and its downstream compensation releases, which was taken to be 10% of combined Station 25 and Station 35 mean runoff (see Table 2). The combined Stations 25 and 35 flow approximately represents the inflow into the Asejire dam (see Figure 3(a)) and the 10% allowance is in accordance with recommended practice (see Twort *et al.* 2000). However, because there is no statutory requirement to make dam compensation releases in Nigeria, an alternative situation in which Asejire makes no downstream compensation releases was also investigated. Full details of the area-ratio scaling calculations are shown in the Appendix (available online at http://www.iwaponline.com/nh/043/017.pdf).

Statistical analysis to determine runoff probabilities at Igbonla

Minimum 1-month annual runoff series were extracted from the reconstructed Igbonla runoff. In extracting the low flows, two situations were considered: (a) the whole calendar year; and (b) the 'wet season', taken to be the period April–October during which it is actually planned to carry out the abstractions. The extracted, low runoff series were then fitted to the 3-parameter log-normal distribution as recommended by Twort *et al.* (2000) as follows.

Let the lower bound of the 3-parameter lognormal distribution be denoted by ϑ, then if the variable Q is 3-parameter log-normally distributed, the variable Y given by Equation (9) is normally distributed.

$$Y = \log_e(Q - \vartheta) \tag{9}$$

The probability density function (pdf) of Q is given by Equation (10) (IH 1975):

$$f_Q(q) = \frac{1}{(q-\vartheta)\sigma_y\sqrt{2\pi}}\exp\left(-0.5\left[\left(\ln(q-\vartheta)-\mu_y\right)\Big/\sigma_y\right]^2\right) \tag{10}$$

where μ_y and σ_y are the mean and standard deviation of Y, respectively, and ϑ is the lower bound of the distribution of Q as stated before; all three constitute the parameters of the 3-parameter log-normal distribution. If these parameters are known, then quantiles for q, i.e. the value of q corresponding to a given cumulative probability value, can be obtained as:

$$q_p = \vartheta + \exp(\mu_y + z_p \times \sigma_y) \tag{11}$$

where z_p is the standard normal variate at p% non-exceedance level, which can be obtained if p is known using (Stedinger *et al.* 1993):

$$z_p = \frac{(0.01p)^{0.135} - (1-0.01p)^{0.135}}{0.1975} \tag{12}$$

Thus, to use the 3-parameter log-normal distribution, all the three distribution parameters must first be estimated. If the lower bound parameter ϑ is known, then the method of maximum likelihood is recommended as the most efficient method for estimating the mean (μ_y) and standard deviation (σ_y) as follows (Stedinger 1980):

$$\mu_y = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{13}$$

$$\sigma_y = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}\left(y_i - \mu_y\right)^2} \tag{14}$$

where y_i is as given by Equation (9) and k is the sample size, i.e. the total number of the y_i's. A simple method for estimating ϑ was also given by Stedinger (1980) as:

$$\vartheta = \frac{\breve{q}\times\hat{q} - \tilde{q}^2}{\breve{q} + \hat{q} - 2\tilde{q}} \tag{15}$$

where $\hat{q}$, $\breve{q}$, and $\tilde{q}$ are the maximum, minimum and median of q, respectively.

Often, as is the case in this study, the objective will be to estimate the probability p for a given q using the 3-parameter log-normal distribution. This is also possible once parameters ϑ, μ_y and σ_y are known. To do this, the value of z_p corresponding to the given q will be determined using the known distribution parameters by re-arranging Equation (11), i.e.

$$z_p = \frac{\ln(q-\vartheta) - \mu_y}{\sigma_y} \tag{16}$$

The corresponding probability can be approximated using (Stedinger *et al.* 1993; McMahon & Adeloye 2005):

$$p(\%) = 100\times\left\{1 - 0.5\times\exp\left[\frac{-(83z_p + 351)\times z_p + 562}{703/z_p + 165}\right]\right\};\quad z_p > 0.0 \tag{17}$$

Due to the symmetry of the normal distribution, the probability for negative values of z_p is the complement of the corresponding positive value, i.e. 100–p%. Also, for $z_p = 0$, p is 50%.

Equation (11) assumes that the 3-parameter log-normal density function is appropriate for describing the low flow

sequences and that the parameters ϑ, μ_y and σ_y of the distribution have been estimated accurately. As will become clear in the Results and Discussions section, the 3-parameter lognormal is appropriate and although the errors in the estimated parameters will not be evaluated specifically because this is not within the purview of the study, the long data records to be used should ensure that any parameter estimation uncertainty will be minimal.

RESULTS AND DISCUSSIONS

Rainfall-runoff analysis using the KSOM

The component planes, a major feature of SOM analysis, are shown in Figure 4 and help to visually illustrate areas in which a variable is high, low or average and thus the relationship or correlation between the variables. The component planes show the values of the variables in each map unit that can be used to estimate the data variable of the input spaces (Vesanto *et al.* 2000). The planes are filled using grey or coloured shades and the way the gradients of these shades relate is an indication of the correlation. Variables exhibiting parallel shades gradients will have high positive correlations; anti-parallel gradients are indicative of negative correlations.

As Figure 4 visually illustrates, the gradients of the plane for Ikeja and Lagos rainfall are parallel, indicating high correlation between these two variables. Indeed for these two stations, there is a nearly perfect correlation in the occurrence of low rainfall (see the dark shade in the top half of the respective component planes) and high rainfall (see the bottom, left hand corner of the component planes). Both of these observations are to be expected, given the close proximity of the two stations. A further feature revealed by the component planes is the generally higher monthly rainfall at Lagos when compared with Ikeja, a situation also not unexpected given the maritime climate that prevails in Lagos Island due to its close proximity to the sea.

The Ibadan and Osogbo rainfall component planes gradients are also parallel, implying good correlation between them. However, although low rainfall at these two stations tends to be clustered in the top part of the plane as observed for Ikeja and Lagos, the overall relationship between Ibadan/Osogbo and Ikeja/Lagos is not as strong as that between either Ibadan and Osogbo or between Ikeja and Lagos. Additionally, rainfall at Ibadan and Osogbo tends

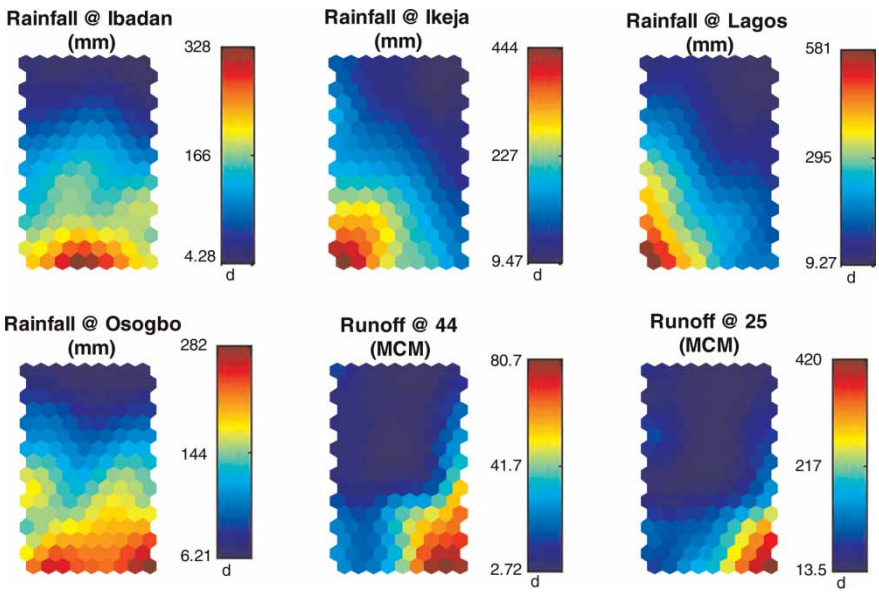

Figure 4 | Component planes for the KSOM model of rainfall and runoff data for the Osun Basin, Nigeria.

to be generally much lower than those observed at both Ikeja and Lagos, further confirming the increasing wetness as one travels downstream within the Osun catchment.

The runoff component planes at both Stations 44 and 25 are broadly similar, exhibiting parallel gradients and co-incidental occurrences of high and low runoff magnitudes. Both of these imply high correlation between the runoff at the two stations, which means that they could be good predictors of each other for the purpose of record extension if one of the stations had a substantially longer record than the other but unfortunately, this was not the case. In relation to the rainfall stations, the component planes of the runoff at both Stations 44 and 25 appear to correlate more with the Osogbo rainfall (and to some extent the Ibadan rainfall) than with either the Ikeja and Lagos rainfall. The Osogbo rainfall station is close to runoff Station 44 whilst the Ibadan rainfall site is close to runoff Station 25; however, information revealed by the component planes would tend to suggest that Osogbo rainfall would be a much better predictor for the runoff at both Stations 44 and 25. Information such as this, if available *a priori*, would have made the choice of the Osogbo rainfall to be sole predictor variable for runoff at the sites if a univariate regression approach were to be adopted. The multivariate approach utilising the SOM that was adopted for this study will take advantage of this high correlation exhibited between the Osogbo rainfall and the runoff sites whilst also considering the correlations, albeit smaller, with the other rainfall data at Ibadan, Ikeja and Lagos. As a consequence, the resulting predictions of the runoff at both sites 25 and 44 would have improved significantly over the use of a univariate approach. The computed correlation coefficients between the variables are shown in Table 3 from where it can be seen that they broadly agree with the indications provided by the visual component planes.

The estimation of the missing values relies on the component planes. The vector that contains the missing element is processed through SOM to search for its BMU. This BMU is a vector like the input vector; the difference is that the BMU contains the full complement of elements. Thus, elements in the BMU that correspond to the missing elements in the input vector are used as the best estimates of these missing values. This procedure was illustrated previously in Figure 2.

Table 3 | Nash-Sutcliffe efficiencies for the Osun Basin data and SOM modelling

		Rainfall (Ibadan)	Rainfall (Ikeja)	Rainfall (Lagos)	Rainfall (Osogbo)	Runoff @ 44	Runoff @ 25	SOM Rainfall (Ibadan)	SOM Rainfall (Ikeja)	SOM Rainfall (Lagos)	SOM Rainfall (Osogbo)	SOM Runoff @ 44	SOM Runoff @ 25
	Rainfall (Ibadan)	1.00											
	Rainfall (Ikeja)	0.62	1.00										
	Rainfall (Lagos)	0.60	0.82	1.00									
	Rainfall (Osogbo)	0.71	0.61	0.57	1.00								
	Runoff @ Station 44	0.38	0.22	0.22	0.49	1.00							
	Runoff @ Station 25	0.40	0.22	0.21	0.49	0.89	1.00						
SOM	Rainfall (Ibadan)	0.94	0.61	0.59	0.75	0.61	0.57	1.00					
SOM	Rainfall (Ikeja)	0.59	0.94	0.73	0.60	0.26	0.26	0.78	1.00				
SOM	Rainfall (Lagos)	0.54	0.71	0.96	0.57	0.20	0.20	0.72	0.96	1.00			
SOM	Rainfall (Osogbo)	0.68	0.58	0.59	0.97	0.70	0.67	0.94	0.75	0.71	1.00		
SOM	Runoff @ Station 44	0.39	0.22	0.22	0.50	0.99	0.89	0.61	0.27	0.21	0.71	1.00	
SOM	Runoff @ Station 25	0.37	0.22	0.21	0.47	0.93	0.94	0.59	0.27	0.21	0.70	0.94	1.00

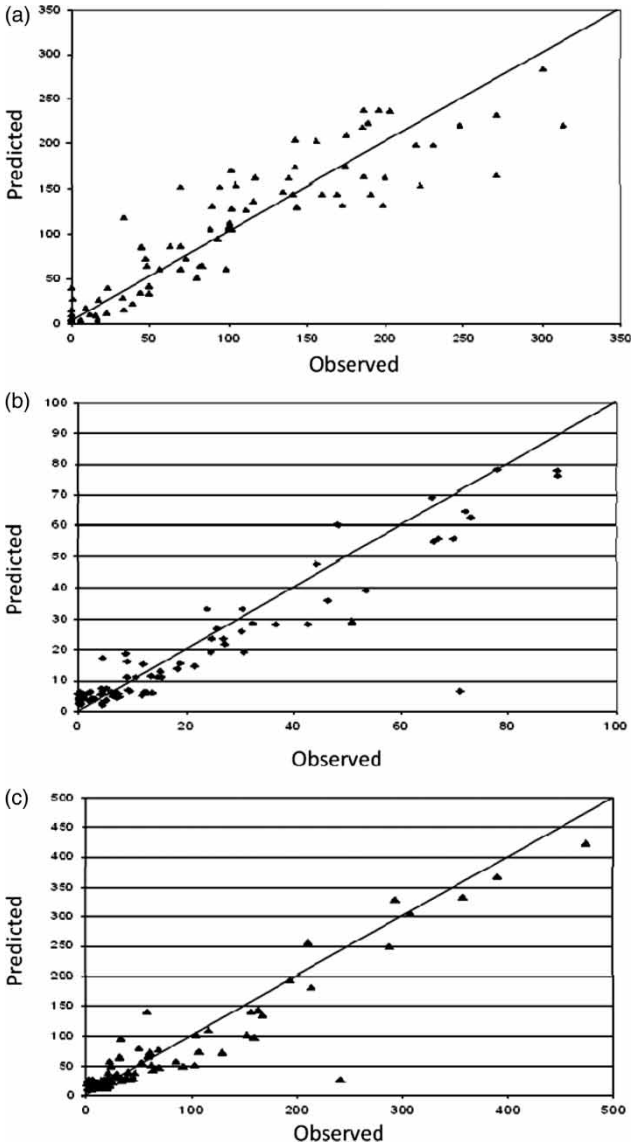

Figure 5 | The performance of the KSOM model during calibration. (a) Osogbo monthly rainfall (mm). (b) Station 44 monthly runoff (MCM). (c) Station 25 monthly runoff (MCM).

The performance of the SOM during calibration is demonstrated in Figure 5 where the observed and predicted quantities are compared, using Osogbo rainfall and the runoff at Stations 44 and 25. Apart from two observations, one at Station 44 and another at Station 25, which appear discordant or like outliers, the performance is generally good. Although not tested formally, the spread of the data points around the line of equality is an indication that the model errors are random, a necessary condition for the validity of the model. This should ensure confidence in using the model for the purpose of prediction. The time series plots of the monthly data are shown in Figure 6 both for the observed and those estimated by the SOM and further confirm the good performance of the SOM. Table 3 shows the Nash-Sutcliffe efficiency indices for the SOM models. These are generally above 94%, further supporting the excellent performance of the SOM in modelling the rainfall and runoff data.

Low flow quantiles

Table 4 contains the summary statistics for the minimum 1-month runoff series at Igbonla based on the entire calendar year and also on the wet season (April–October) of the year. Q_{igbonla} represents the reconstructed runoff at Igbonla assuming there were no compensation releases by the Asejire dam, while $Q_{\text{igbonla-adj}}$ includes Asejire dam compensation releases. Also included in the Table are estimates of the parameters ϑ, μ_y and σ_y for the 3-parameter log-normal density function. As seen in the Table, the minimum series based on the wet season period exhibited positive skewness, implying that they could be modelled using the 3-parameter distribution function. Furthermore, the lower bound parameter, ϑ, for the two cases (Q_{igbonla} and $Q_{\text{igbonla-adj}}$) in this category was below the least runoff value in the minimum 1-month runoff series. On the contrary, the minimum 1-month runoff series based on the complete calendar year exhibited negative skewness which cannot be modelled using the 3-parameter log-normal distribution. Additionally, the estimate for parameter ϑ in this case far exceeds the least value in the series, a further condition for precluding the use of the 3-parameter log-normal distribution (see Stedinger 1980). Finally, given the proposed abstraction rates at Igbonla both for the present and future time horizons (Table 1), there is no way that they will be met by the low flows that ensue from the calendar year minimum flow analysis. Consequently, further consideration of the calendar year series was stopped and the following discussions will be limited to the wet season series.

The fitted 3-parameter log-normal distribution is shown in Figure 7 for the wet season 1-month minimum runoff $Q_{\text{igbonla-adj}}$ series. As Figure 7 reveals, the 3-parameter log–normal distribution fits the data well ($R^2 = 0.97$). This is

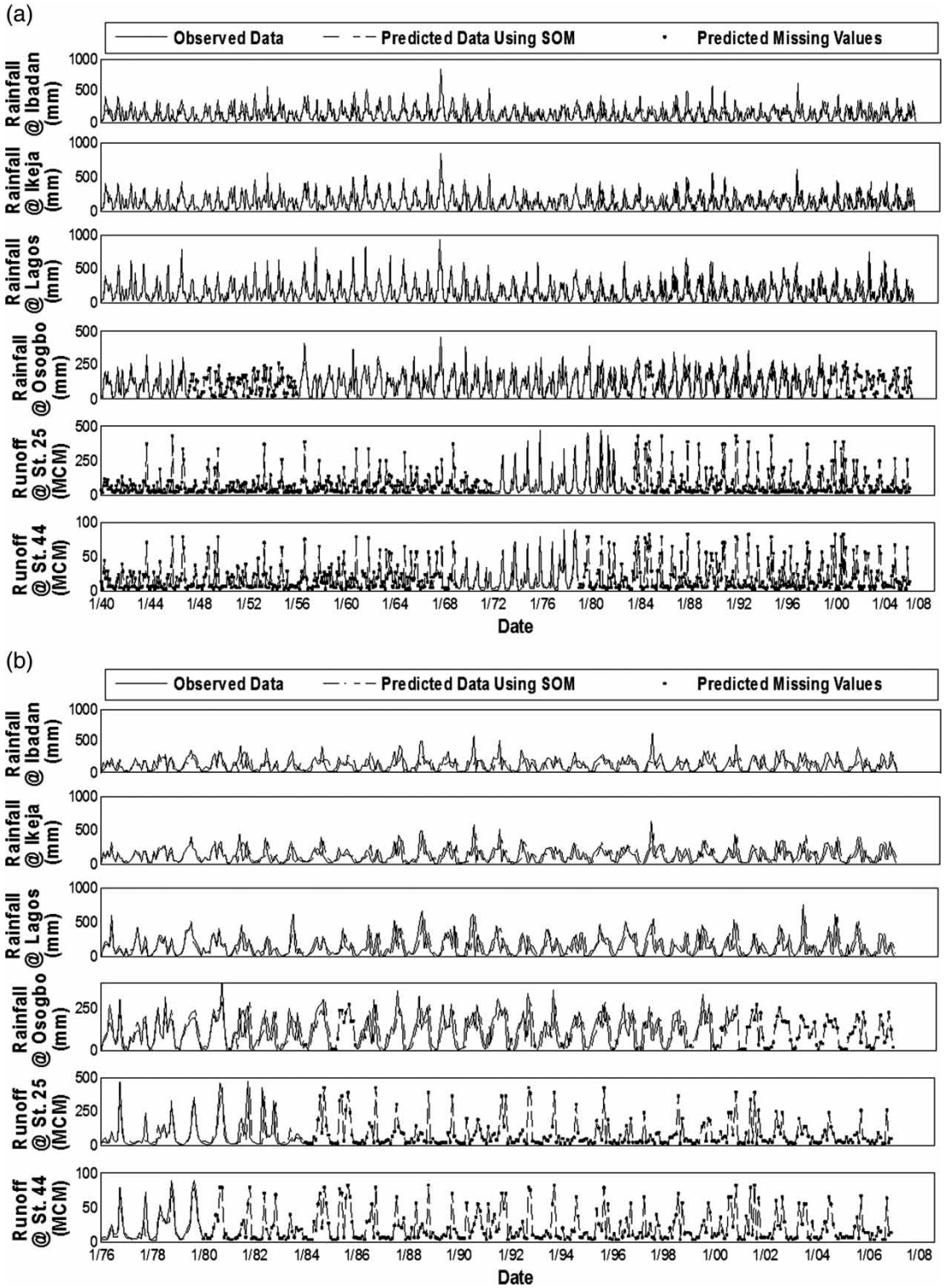

Figure 6 | Time series plot of observed and KSOM-predicted monthly rainfall and runoff within Osun Basin, Nigeria, (a) complete time series; (b) recent time slice.

Table 4 | Summary statistics for minimum 1-month low flow series at Igbonla and parameters of the fitted 3-parameter log-normal distribution function

	Calendar year		Wet season	
	$Q_{igbonla}$	$Q_{igbonla-adj}$	$Q_{igbonla}$	$Q_{igbonla-adj}$
Mean (MCM)	7.02	18.29	10.33	21.60
Std (MCM)	2.41	2.41	6.21	6.21
Skew	−0.42	−0.42	1.86	1.86
CV	0.34	0.13	0.60	0.29
Min (MCM)	0.51	11.78	2.02	13.29
Max (MCM)	13.30	24.57	34.37	45.64
Median (MCM)	6.997	18.27	8.43	19.70
ϑ (MCM)	220.16	231.43	−0.09	11.18
μ_y	–	–	2.196	2.196
σ_y	–	–	0.547	0.547

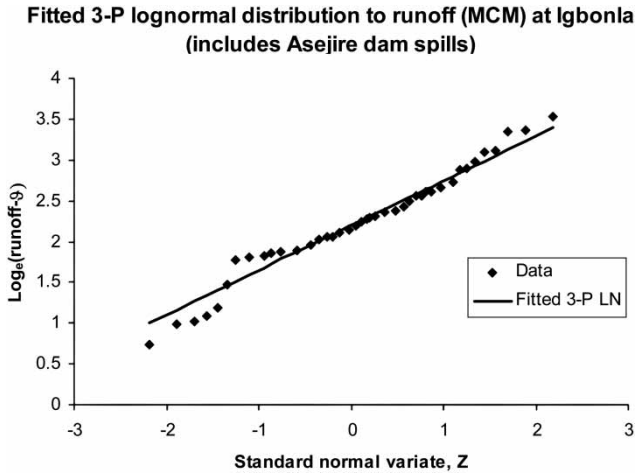

Figure 7 | Fitted 3-parameter lognormal distribution to wet season 1-month low runoff series at Igbonla ($Q_{igbonla-adj}$ runoff scenario).

especially so for the right hand tail region of the distribution, i.e. at high positive z values corresponding to the investigated abstraction rates at Igbonla (see Table 5). The extreme low flow region of Figure 7 exhibits consistent upward bias in the 3-parameter log-normal predictions, which may have been caused by the fewer number of observations in this region in comparison to the other parts of the distribution function. The probabilities for the various abstraction rates as estimated from the fitted 3-parameter lognormal distribution are summarised in Table 5 for both the $Q_{igbonla-adj}$ and $Q_{igbonla}$. Although daily abstraction rates were proposed for the project, these had to be converted into their monthly equivalents because the low flow frequency analysis was carried out using monthly data. For this purpose, each of the months was assumed to have 30.42 days.

As seen in Table 5 for the $Q_{igbonla-adj}$, except for the 2011 abstraction where the probability of achieving the required abstraction or higher was 59%, all the other abstraction scenarios had extremely low probabilities. The probability of exceedance, i.e. the probability of achieving an abstraction rate that is equal to or higher than the desired, is a measure of the reliability of the direct abstraction system. In the UK and other developed economies, it is customary for water supply infrastructures to be designed for 98% reliability (Twort *et al.* 2000), i.e. probability of exceedance of 98%. If the facility is for irrigation or other (less strategic) purposes for which failure to meet demand does not produce catastrophic consequences, reliabilities as low as 80% may suffice. Thus, Table 6 has been produced which shows what the abstraction rates at Igbonla would be for a range of acceptable probabilities. Due to the upward bias in the prediction of low flows by the 3-parameter log-normal distribution as noted earlier, the estimated flows in Table 6 are probably higher than what they should be but despite this, it is clear

Table 5 | Modelled probabilities of projected abstractions at Igbonla for both $Q_{igbonla-adj}$ and ($Q_{igbonla}$)

Year	Abstraction rate, q (MCM month^{-1})	$q-\vartheta$	$\ln(q-\vartheta)$	z	Cumulative probability	Exceedance probability (%)
2011	19.16	7.98	2.08	−0.22	0.41	58.67
		(19.25)	(2.96)	(1.39)	(0.92)	(8.21)
2015	36.50	25.32	3.23	1.89	0.971	2.92
		(36.59)	(3.6)	(2.57)	(0.995)	(0.52)
2020	55.36	44.18	3.79	2.91	0.998	0.1808
		(55.44)	(4.02)	(3.32)	(0.9996)	(0.0442)

Table 6 | Modelled shortfall at Igbonla (based on $Q_{igbonla-adj}$ runoff scenario) at future time horizons for acceptable exceedance probability levels

Exceedance probability (%)	z	Possible abstraction rate, q (MCM month^{-1})	Shortfall (%) Year 2011	Year 2015	Year 2020
98	−2.06362	14.09099	26.5	61.4	74.5
95	−1.64931	14.8307	22.6	59.4	73.2
90	−1.28128	15.64421	18.3	57.1	71.7
80	−0.83856	16.86673	12.0	53.8	69.5

that even at 80% reliability, the possible abstraction profile is much below the 2011 projection, not to mention the higher 2015 and 2020 abstraction profile requirements.

What the results in Tables 5 and 6 suggest is that it will be extremely difficult to meet the projected abstraction rate at Igbonla based on the $Q_{igbonla-adj}$ runoff series at the site; some other means of meeting this shortfall, such as through groundwater resources would have to be sought. The situation is even worse for the $Q_{igbonla}$ scenario as shown in Table 5. As a reminder, $Q_{igbonla}$ assumes that Asejire makes no compensation releases whatsoever.

As implied earlier, assuming that Asejire makes no compensation release is rather draconian even by Nigerian hydrological practice standards and thus little credence should be placed on the results based on this assumption. The true situation would be that Asejire does make some compensation releases, either managed or during spills when the reservoir is full. The assumption here that a fixed amount equal to 10% of the average annual runoff is spilled each month is a gross over-simplification of a complex reservoir operating policy decision making. However, without the information, not much can be done to arrive at the precise amount and scheduling of the downstream releases. If the releases are made up of the involuntary spills during the wet season when the reservoir is likely to be full, then the analysis reported here would have underpredicted the runoff during the wet season and consequently the associated probabilities would have been underestimated as well. This is not a bad thing as it would mean that the current analysis has been conservative.

Thus, although the use of the SOM has enabled specific answers regarding the abstraction decisions at the proposed Igbonla to be arrived at, the wider analysis did rely on a number of assumptions. These assumptions were based on sound scientific judgement; however, they nonetheless bestow some limitations on the analysis results. It is therefore very important that these caveats are well recognised when using the results presented in this work. For example, it is doubtless that more certainty in results would have been achieved if: measured runoff had been available at Igbonla and used instead of the modelled runoff; information about the operational practices at Asejire had been available which would have improved the reconstruction of the runoff downstream of the Asejire dam; analysis had been based on daily runoff data rather than the lumped monthly approach, and finally account had been taken of the impact of climate change on future levels of runoff in the catchment.

CONCLUSIONS

A new methodology based on the SOM, unsupervised ANNs has been presented as a viable rainfall-runoff modelling paradigm to achieve robust runoff data record extension in data sparse areas. The attractiveness of the SOM is that because of its powerful clustering ability, rainfall-runoff modelling is unhindered by missing values; indeed, as seen in this work, one of the outcomes of the SOM modelling is the provision of reliable estimates for the missing values. Consequently, SOM rainfall-runoff modelling offers a huge potential for redressing the difficulties associated with water resources planning in regions where available runoff data records are short and riddled with gaps and missing values, a common problem (Khan *et al*. 2010).

On the basis of the results obtained here, the runoff at Igbonla will only provide the projected abstraction rates with a low level of reliability. Given the rather patchy nature of the water supply infrastructure in most parts of Nigeria, a reliability of 59%, the best of the scenarios analysed, is probably adequate but would be judged unacceptable in more advanced communities of the world where the provision of uninterrupted water supply is normally taken for granted. For example, in the UK, most water undertakers are required by statute to supply with at least 98% reliability. Achieving such a high level of reliability at Igbonla would require a significant reduction in the abstraction rates as shown in Table 6.

Alternatively, the water available from Igbonla abstractions could be supplemented with groundwater development. Indeed, conjunctive use of ground and surface water resources is very popular in different parts of the world and represents a viable way of maximising water yield while at the same time benefiting either water resource compartment. For example, more of surface water resources could be used when water in rivers and reservoirs is plentiful, leaving the groundwater reservoirs to recharge and to be used later when surface water are significantly depleted. Thus, the fear that using groundwater will accelerate saltwater intrusion in Lagos can be allayed if groundwater abstraction is well managed, which will be the case if its abstraction is only used to supplement that derived from development of surface water resources. Another option for the Lagos project which is reportedly being actively considered by the authorities is the use of desalination. Desalination is feasible given the proximity of Lagos to the sea but potentially expensive and energy intensive and may thus be problematic in a city with unreliable, intermittent power supply infrastructure.

REFERENCES

Adeloye, A. J. 1990 Streamflow data and surface water resource assessment. *Journal Water Supply, Research and Technology-AQUA* **39**, 225–236.

Adeloye, A. J. 1996 An opportunity loss model for estimating value of streamflow data for reservoir planning. *Water Resources Management* **10**, 45–79.

Adeloye, A. J. 2009 The relative utility of multiple regression and ANN models for rapidly predicting the capacity of water supply reservoirs. *Environmental Modelling and Software* **24**, 1233–1240.

Adeloye, A. J. & Rustum, R. 2011 Lagos (Nigeria) flooding and influence of urban planning. *Proceedings of ICE Urban Design and Planning* **164** (DP3), 175–187.

Adeloye, A. J., Rustum, R. & Ibrahim, I. D. 2011 Kohonen self-organising map estimator for the reference crop evapotranspiration. *Water Resource Research* **47**, W08523.

Alhoniemi, E., Hollmén, J., Simula, O. & Vesanto, J. 1999 Process monitoring and modeling using the self-organizing map. *Integrated Computer-Aided Engineering* **6**, 3–14.

Back, B., Sere, K. & Hanna, V. 1998 Managing complexity in large database using self organising map. *Accounting Management and Information Technologies* **8**, 191–210.

Chang, Y. M., Chang, L. C. & Chang, F. J. 2004 Comparison of static-feedforward and dynamic-feedforward neural networks for rainfall-runoff modelling. *Journal of Hydrology* **290**, 297–311.

Chang, F. J., Chang, L. C., Kao, H. S. & Wu, G. W. 2010 Assessing the effort of meteorological variables for evaporation estimation by self-organizing map neural network. *Journal of Hydrology* **384**, 118–129.

Garcia, H. & Gonzalez, L. 2004 Self-organizing map and clustering for wastewater treatment monitoring. *Engineering Applications of Artificial Intelligence* **17**, 215–225.

IH 1975 Flood Studies Report. Vol. I- Hydrological Studies. Institute of Hydrology, Wallingford, England. 550 pp.

IH 1999 *Flood Estimation Handbook (5 Volumes)*. Institute of Hydrology, Wallingford, England.

Kangas, J. & Simula, O. 1995 Process monitoring and visualization using self organizing map. In: *Neural Networks for Chemical Engineers, Ch. 14* (A. B. Bulsari, ed.). Elsevier Science, Dordrecht, The Netherlands.

Khan, S., Savenije, H. H. G., Demuth, S. & Hubert, P. (eds) 2010 *Hydrocomplexity: New Tools for Solving Wicked Water Problems*. IAHS Pub. No. 338. IAHS Press, Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK.

Kohonen, T., Simula, O. & Visa, A. 1996 Engineering applications of the self organising map. *IEEE* **84**, 1358–1384.

Laaha, G. & Bloschl, G. 2005 Low flow estimates from short stream flow records-a comparison of methods. *Journal of Hydrology* **306**, 264–286.

Laaha, G. & Bloschl, G. 2007 A national low flow estimation procedure for Austria. *Hydrological Sciences Journal* **52**, 625–644.

Loucks, D. P., Stedinger, J. R. & Haith, D. A. 1981 *Water Resources Systems Planning and Analysis*. Prentice-Hall, Inc., Englewood Cliffs, NJ, USA. 559 pp.

McCuen, R. H. & Levy, B. S. 2000 Evaluation of peak discharge transposition. *Journal of Hydrologic Engineering ASCE* **5**, 278–289.

McMahon, T. A. & Adeloye, A. J. 2005 *Water Resources Yield*. Water Resources Publications LLC, Colorado, USA. Available from: www.wrpllc.com.

Minns, A. W. & Hall, M. J. 1996 Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal* **41**, 399–417.

Mohamoud, Y. M. 2008 Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. *Hydrological Sciences Journal* **53**, 706–724.

Mohamoud, Y. M. & Parmar, R. S. 2006 Estimating streamflow and associated hydraulic geometry, the mid-atlantic region, USA. *Journal of the American Water Resources Association (JAWRA)* **42**, 755–768.

Parasuraman, K. A., Elshorbagy, A. & Carey, S. K. 2006 Spiking modular neural networks: a neural network modelling approach for hydrological processes. *Water Resource Research* **42**, W05412.

Penn, B. S. 2005 Using self-organising maps to visualize high dimensional data. *Computers and Goesciences* **31**, 531–544.

Rivera, D., Lillo, M., Uvo, C. B., Billib, M. & Arumi, J. L. 2011 Forecasting monthly precipitation in central Chile: a

self-organising map approach using filtered sea surface temperature. *Theoretical and Applied Climatology* **107** (1–2), 1–13.

Rustum, R. & Adeloye, A. J. 2007 Replacing outliers and missing values from activated sludge data using Kohonen self organising map. *Journal of Environmental Engineering, ASCE* **133**, 909–916.

Rustum, R., Adeloye, A. J. & Simala, A. 2007 *Kohonen Self-Organising Map (KSOM) Extracted Features for Enhancing MLP-ANN Prediction Models for BOD_5*. IAHS Publ. No. 314, 181–187. IAHS Press, Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK.

Rustum, R., Adeloye, A. J. & Scholz, M. 2008 Applying Kohonen self-organising map as a software sensor to predict the biochemical oxygen demand. *Water Environment Research* **80**, 32–40,43.

Shamseldin, A. Y. 1997 Application of neural network technique to rainfall-runoff modelling. *Journal of Hydrology* **199**, 272–294.

Smith, J. & Eli, R. N. 1995 Neural-networks models of rainfall-runoff process. *Journal of Water Resources Planning and Management – ASCE* **121**, 499–508.

Stedinger, J. R. 1980 Fitting log-normal distribution to hydrologic data. *Water Resource Research* **16**, 481–490.

Stedinger, J. R., Vogel, R. M. & Foufoula-Georgiou, E. 1993 Frequency analysis of extreme events. In: *Handbook of Applied Hydrology, Chapter 19* (D. R. Maidment, ed.). McGraw-Hill, New York, USA.

Tananaki, C., Thrasyvoulou, A., Giraudel, J. L. & Montury, M. 2007 Determination of volatile characteristics of Greek and Turkish pine honey samples and their classification by using Kohonen self organizing maps. *Food Chemistry* **101**, 1687–1693.

Tokar, A. S. & Johonson, P. A. 1999 Rainfall-runoff modelling using artificial neural networks. *Journal of Hydrologic Engineering – ASCE* **4**, 232–239.

Tokun, A. & Adeloye, A. J. 2005 *Sustainable Water Management Solution for Ibadan City, Nigeria*. IAHS Publ. No. 293, 41–48. IAHS Press, Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK, 41–48.

Twort, A. C., Ratnayaka, D. D. & Brandt, M. J. 2000 *Water Supply (5th Edition)*. Arnold Publishing, London, England, 712 pp.

Vesanto, J., Himberge, J., Alhoniemi, E. & Parhankangas, J. 2000 Self-organizing Map (SOM) Toolbox for Matlab 5, Report No. A57. Helsinki University of Technology, Laboratory of Computer and Information Science, Helsinki, Finland.

Vogel, R. M. & Sankarasubramanian, A. 2000 Spatial scaling properties of annual streamflow in the United States. *Hydrological Sciences Journal* **45**, 465–476.

Wharton, G., Arnell, N. W., Gregory, K. J. & Gurnell, A. M. 1989 River discharge estimated from channel dimensions. *Journal of Hydrology* **106**, 365–376.

First received 17 January 2011; accepted in revised form 26 September 2011. Available online 8 May 2012