HYDROLOGICAL PROCESSES
Hydrol. Process. **22**, 4301–4309 (2008)
Published online 7 May 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/hyp.7044

Model data selection using gamma test for daily solar radiation estimation

R. Remesan,* M. A. Shamim and D. Han
Water and Environmental Management Research Centre, Department of Civil Engineering, University of Bristol, Lunsford House, Cantocks Close, Clifton, Bristol, BS8 1UP, UK

Abstract:

Hydrological modelling is a complicated procedure and there are many tough questions facing all modellers: what input data should be used? how much data is required? and what model should be used? In this paper, the gamma test (GT) has been used for the first time in modelling one of the key hydrological components: solar radiation. The study aimed to resolve the questions about the relative importance of input variables and to determine the optimum number of data points required to construct a reliable smooth model. The proposed methodology has been studied through the estimation of daily solar radiation in the Brue Catchment, the UK. The relationship between input and output in the meteorological data sets was achieved through error variance estimation before the modelling using the GT. This work has demonstrated how the GT helps model development in nonlinear modelling techniques such as local linear regression (LLR) and artificial neural networks (ANN). It was found that the GT provided very useful information for input data selection and subsequent model development. The study has wider implications for various hydrological modelling practices and suggests further exploration of this technique for improving informed data and model selection, which has been a difficult field in hydrology in past decades. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS model data selection; gamma test; solar radiation

Received 17 May 2007; Accepted 28 February 2008

INTRODUCTION

Solar radiation is one of the key inputs for most hydrological models in estimating reference evapotranspiration (Tan *et al*., 2007). Daily solar radiation data is more popular than data at other time intervals for crop growth simulation models and hydrological and soil water balance models (Ball *et al*., 2004). In spite of the great importance of solar radiation, many published studies point out the major challenges associated with solar radiation data collection. Lack of solar radiation data is quite common, even in developed countries such as the USA (Richardson, 1985; Hook and McClendon, 1992) and Canada (De Jong and Stewart, 1993). Many researchers note the fact that solar radiation is an infrequently measured meteorological variable compared with temperature and rainfall (Liu and Scott, 2001; Weiss and Hays, 2004).

In past decades, many empirical and physical radiation models have been proposed (Sabbagh *et al*., 1977; Noia *et al*., 1993a,b; Tovar and Baldasano, 2001). The Angstrom equation, which was proposed by Angstrom (1924) and subsequently modified by Prescott (1940), is considered as the most popular and widely used method for the estimation of monthly averaged daily (global) irradiation value. Later several physical based empirical models were devised based on Chang (1968), who reported that there was a good relation between net radiation and global solar radiation, since the latter is the principal source of energy. Based on this argument Bristow and Campbell (1984), suggested an empirical relationship for daily global radiation, as a function of daily net radiation and the difference between maximum and minimum temperature. Later, Allen (1995) suggested the use of a self-calibrating model to estimate mean monthly global solar radiation based on the work of Hargreaves and Samani (1982). His research suggested that the mean daily global radiation can be estimated as a function of net radiation, and mean monthly maximum and minimum temperatures. The Bristow–Campbell model has been used in numerous hydrological related studies, and improvements have been developed over the years (Donatelli and Campbell, 1998). The Campbell–Donatelli method was implemented in many weather generators including MarkSim (Jones and Thornton, 2000) and ClimGen (Stöckle *et al*., 2001). Recently Donatelli *et al*, (2003, 2006) developed a windows based model named RadEst3·00 which estimates and evaluates daily global solar radiation values at given latitudes. Some other interesting work has been done in the area of solar radiation prediction using ARMA (autoregressive moving average) and Fourier analysis (Goh and Tan, 1977; Mustacchi *et al*., 1979). Furthermore, new approaches to predict solar radiation series have been developed using artificial neural networks(ANN), particularly in Turkey (Saylan *et al*., 2003; Ogulata and Ogulata, 2002; Tırıs *et al*., 1996; Togrul and

* Correspondence to: R. Remesan, Water and Environmental Management Research Centre, Department of Civil Engineering, University of Bristol, Lunsford House, Cantocks Close, Clifton, Bristol, BS8 1UP, UK.
E-mail: renji.remesan@bristol.ac.uk

Onat., 1999; Dinçer *et al*., 1996), but also in other places (Negnevitsky and Le, 1995; Alawi and Hinai, 1998; Mohandes *et al*., 1998; Kemmoku *et al*., 1999; Sfetsos and Coonick, 2001).

Despite an abundance of studies on prediction and modelling of solar radiation and many other variables using nonlinear techniques such as artificial neural networks (ANN), there are still many questions that need to be answered. For example, to what extent do the inputs determine the output from a smooth model? Given an input vector x how accurately can the output y be predicted? How many data points are required to make a prediction with best possible accuracy? Which inputs are relevant in making the prediction and which are irrelevant? So far, these questions have not been addressed adequately by the hydrological community (Han *et al*., 2007). However, owing to the advancement of modern computing technology and a new algorithm from the computing science community called the gamma test (GT) (Agalbjörn *et al*., 1997; Končar, 1997), it is possible that significant progresses could be made in tackling these problems. A formal proof for the Gamma Test can be found in Evans (2002) and Evans and Jones (2002). It is accomplished by estimating the variance of the noise var(r), computed from the raw data using efficient, scalable algorithms. This novel technique, the GT, enables one to quickly evaluate and estimate the best mean squared error that can be achieved by a smooth model on unseen data for a given selection of inputs, before model construction. This technique can be used to find the best embedding dimensions and time lags for time series analysis. This information would help to determine the best input combinations to achieve a particular target output. Overtraining is considered to be one of the serious weaknesses associated with almost all nonlinear modelling techniques including ANN, giving excellent results on the training data but very poor results on the unseen test data. The GT is designed to solve this problem efficiently by giving an estimate of how closely any smooth model could fit the unseen data. Thus we can avoid the guesswork associated with nonlinear curve fitting techniques.

The main objective of this study is to assess the performance of nonlinear techniques like local linear regression (LLR) and neural networks to estimate daily (global) irradiation values with the GT, based on different meteorological input data. This paper demonstrates the capability of the Gamma Test to identify the appropriate embedding nonlinear model dimensions through the estimate of variance of the noise associated with the data, before model construction and evaluation.

MATERIALS AND METHODS

Gamma test, V-ratio and M-test

The GT estimates the minimum mean square error (MSE) that can be achieved when modelling the unseen data using any continuous nonlinear models. The GT was first reported by Konča (1997) and Agalbjörn, *et al*. (1997), and later enhanced and discussed in detail by many researchers (Chuzhanova *et al*., 1998; De Oliveira, 1999; Tsui, 1999; Tsui *et al*., 2002; Durrant, 2001; Jones *et al*., 2002).

Only a brief introduction to the GT is given here and the interested readers should consult the aforementioned papers for further details. The basic idea is quite distinct from earlier attempts with nonlinear analysis. Suppose we have a set of data observations of the form

$$\{(\mathbf{x}_i, y_i), 1 \leq i \leq M\} \tag{1}$$

where the input vectors $\mathbf{x_i} \in R^m$ are vectors confined to some closed bounded set $C \in R^m$ and, without loss of generality, the corresponding outputs $y_i \in R$ are scalars. The vectors $\mathbf{x}$ contain predicatively useful factors influencing the output y. The only assumption made is that the underlying relationship of the system is of the following form

$$y = f(\mathbf{x}_1 \ldots \mathbf{x}_m) + r \tag{2}$$

where f is a smooth function and r is a random variable that represents noise. Without loss of generality it can be assumed that the mean of the distribution of r is zero (since any constant bias can be subsumed into the unknown function f) and that the variance of the noise Var(r) is bounded. The domain of a possible model is now restricted to the class of smooth functions which have bounded first partial derivatives. The Gamma statistic Γ is an estimate of the model's output variance that cannot be accounted for by a smooth data model.

The GT is based on $N[i, k]$, which are the kth ($1 \leq k \leq p$) nearest neighbours $\mathbf{x}_{N[i,k]}(1 \leq k \leq p)$ for each vector $\mathbf{x}_i(1 \leq i \leq M)$.Specifically, the GT is derived from the Delta function of the input vectors:

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} \left|\mathbf{x}_{N(i,k)} - \mathbf{x}_i\right|^2 \qquad (1 \leq k \leq p) \tag{3}$$

where $|\ldots|$ denotes Euclidean distance, and the corresponding Gamma function of the output values:

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} \left|y_{N(i,k)} - y_i\right|^2 \quad (1 \leq k \leq p) \tag{4}$$

where $y_{N(i,k)}$ is the corresponding y-value for the kth nearest neighbour of $\mathbf{x}_i$ in Equation (3). In order to compute Γ a least squares regression line is constructed for the p points $(\delta_M(k), \gamma_M(k))$

$$\gamma = A\delta + \Gamma \tag{5}$$

The intercept on the vertical axis ($\delta = 0$) is the Γ value, as can be shown

$$\gamma_M(k) \longrightarrow Var(r) \text{ in probability as } \delta_M(k) \longrightarrow 0 \tag{6}$$

Calculating the regression line gradient can also provide helpful information on the complexity of the system

DOI: 10.1002/hyp

under investigation. A formal mathematical justification of the method can be found in Evans and Jones (2002).

The graphical output of this regression line (Equation (5)) provides very useful information. First, it is remarkable that the vertical intercept Γ of the y (or Gamma) axis offers an estimate of the best MSE achievable utilizing a modelling technique for unknown smooth functions of continuous variables (Evans and Jones, 2002). Second, the gradient offers an indication of the model's complexity (a steeper gradient indicates a model of greater complexity).

The GT is a non-parametric method and the results apply regardless of the particular techniques used to subsequently build a model of f. The result can be standardized by considering another term V_{ratio}, which returns a scale invariant noise estimate between zero and one. The V_{ratio} is be defined as

$$V_{ratio} = \frac{\Gamma}{\sigma^2(y)} \tag{7}$$

where, $\sigma^2(y)$ is the variance of output y, which allows a judgement to be formed independent of the output range as to how well the output can be modelled by a smooth function. A V_{ratio} close to zero indicates that there is a high degree of predictability of the given output y.

The reliability of the Γ statistic can be determined by running a series of GT for increasing M, to establish the size of data set required to produce a stable asymptote. This is known as the M-test, and the result helps to avoid wasteful attempts at fitting the model beyond the stage where the MSE on the training data is smaller than Var(r), which may lead to 'overfitting'. The M-test also helps to decide how much data are required to build a model with a mean squared error which approximates the estimated noise variance. In practice, the GT can be achieved through winGamma™ software implementation (Durrant, 2001). Corcoran *et al*. (2003), applied the GT as a method for crime incident forecasting by focusing upon geographical areas of concern that transcend traditional policing boundaries. The authors believed this technique was very effective and could be potentially used for water management including flood prediction and other hydrological nonlinear modelling.

NONLINEAR MODELS

The GT helps to make decisions about input data selection and the actual modelling can then be carried out using one the nonlinear mathematical models. Nowadays, owing to the advancement of computer technology, there are a large number of nonlinear methods such as artificial neural networks, support vector machines, fuzzy logical systems, polynomial functions, local linear regressions, Bayesian belie networks, decision trees, etc. This study, because of constraints of time and resources, focused on only two popular model types: local linear regression (LLR) and artificial neural networks (ANN). Only brief introductions to them are given here and further details can be found in the references.

Local linear regression (LLR)

The LLF technique is a widely studied nonparametric regression method that has been widely used in many low-dimensional forecasting and smoothing problems. The advantage of the LLR technique is that reasonably reliable statistical modelling can be performed locally with a small amount of sample data. At the same time, LLR can produce very accurate predictions in regions of high data density in the input space. The LLR procedure requires only three data points to obtain an initial prediction and then uses all newly updated data as they becomes available to make further predictions. The only problem with LLR is to decide the size of $p_{\max}$, the number of near neighbours to be included for the local linear modelling. The method of choosing $p_{\max}$ for linear regression is called *influence statistics* and is explained below.

Given a neighbourhood of $p_{\max}$ points, we must solve a linear matrix equation

$$\mathbf{Xm} = \mathbf{y} \tag{8}$$

where $\mathbf{X}$ is a $p_{\max} \times d$ matrix of the $p_{\max}$ input points in d dimensions, $\mathbf{x}_i (1 \leq i \leq p_{\max})$ are the nearest neighbour points, $\mathbf{y}$ is a column vector of length $p_{\max}$ of the corresponding outputs, and $\mathbf{m}$ is a column vector of parameters that must be determined to provide the optimal mapping from $\mathbf{X}$ to $\mathbf{y}$, such that

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{x_{p\max}1} & x_{x_{p\max}2} & x_{x_{p\max}3} & \cdots & x_{x_{p\max}d} \end{pmatrix} \times \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ \vdots \\ m_d \end{pmatrix} = \begin{pmatrix} y_1 \\ y_1 \\ \vdots \\ y_{p\max} \end{pmatrix} \tag{9}$$

The rank $\boldsymbol{r}$ of the matrix X is the number of linearly independent rows, which will affect the existence or uniqueness of the solution for $\boldsymbol{m}$.

If the matrix $\mathbf{X}$ is square and non-singular then the unique solution to Equation (8) is $\mathbf{m} = \mathbf{X}^{-1}\mathbf{y}$. If $\mathbf{X}$ is not square or singular, we modify Equation (8) and attempt to find a vector $\mathbf{m}$ which minimizes

$$|\mathbf{Xm} - \mathbf{y}|^2 \tag{10}$$

which was proved by Penrose (1955) where the unique solution to this problem is provided by $\mathbf{m} = \mathbf{X}^{\#}\mathbf{y}$ where $\mathbf{X}^{\#}$ is a pseudo-inverse matrix (Penrose, 1955, 1956).

In this study, a *kd*-tree is used to organize the input training data, with a time-complexity in the order $O(M \log M)$. A *kd*-tree (short for k-dimensional tree) is a space-partitioning data structure for organizing points

DOI: 10.1002/hyp

in a k-dimensional space so that the LLR algorithms can be implemented using a minimum number of direct evaluations. More theoretical aspects of *kd*-tree can be found in Durrant (2001) and Jones (2004).

Artificial neural networks (ANN)

The theory of ANNs was first proposed in the early 1940s when McCulloch and Pitts developed the first computational representation of a neuron (McCulloch and Pitts, 1943). Later Rosenblatt proposed the idea of *perceptrons* (Rosenblatt, 1962) in which single layer feedforward networks of McCulloch–Pitts neurons could carry out various computational tasks with the help of weights and training algorithm. The applications of ANNs are based on their ability to mimic the human mental and neural structure to construct a good approximation of functional relationships between past and future values of a time series. The supervised ANN is the most commonly used ANN, in which the input is presented to the network along with the desired output, and the weights are adjusted so that the network attempts to produce the desired output. There are different learning algorithms, and a popular algorithm is the back propagation algorithm, which employs gradient descent and gradient descent with momentum; these are often too slow for practical problems because they require low learning rates for stable learning. Algorithms like conjugate gradient, quasi-Newton, Levenberg–Marquardt (LM), etc., are faster algorithms that all make use of standard numerical optimization techniques. Minsky and Papert (1969) highlighted the weaknesses of single layer perceptrons as their ability to solve linearly separable problems only. In practice nowadays, it is usually most effective to use two hidden layers (Jones, 2004). In this study, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) neural network training algorithm (Fletcher, 1987), and conjugate gradient training algorithms along with a two layer architecture embedded in WinGamma software were used. The BFGS algorithm is a quasi-Newton method performed iteratively using successively improved approximations to the inverse Hessian matrix, instead of the true inverse. The performance of the LLR technique and neural network based models were compared using three global statistics: correlation efficiency, root mean squared error (RMSE) and mean bias error (MBE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{N}} \tag{11}$$

The mean bias error can be estimated using the following equation

$$MBE = \frac{\sum_{i=1}^{N}(P_i - O_i)}{N} \tag{12}$$

In both equations, P denotes the predicted values of daily solar radiation (MJ m^{-2} day^{-1}), while O denotes the observed values of daily solar radiation (MJ m^{-2} day^{-1}) in the study area and N is the data point number.

STUDY AREA AND DATA SET

In this study, solar radiation data were collected from a meteorological station located in the Brue Catchment in south-west England. It was the site of the NERC (Natural Environment Research Council) funded HYREX project (Hydrological Radar Experiment) which ran from May 1993 to April 1997 (its data collection was extended to 2000). The Brue catchment area is located at 51·075 °N and 2·58 °W and drains an area of 135·2 sq km. It is predominantly a rural catchment of modest relief with spring-fed headwaters rising in the Mendip Hills and Salisbury Plain. The raingauge network at the Brue is quite dense and consists of 49 Casells 0·2 mm tipping bucket type raingauges, each having a tip time of 10 s. The average annual rainfall over the catchment is estimated as 867 mm. An automatic weather station (AWS) and automatic soil water station (ASWS) were located to record hourly (global) irradiation, net radiation and other physical parameters. The data sets contain hourly information of temperature, rainfall, atmospheric pressure, and wind velocity. The major issues associated with the raw data were lack of sunshine ratio, i.e. ratio of average daily sunshine hours S, and theoretical sunshine duration S_0, which is relevant in solar radiation modelling (Rietveld, 1978; Benson *et al.*, 1984; Gopinathan 1988; Akinoglu and Ecevit, 1990). In this study, the following input parameters were considered: horizontal extraterrestrial radiation (based on Allen *et al.*, 1998), mean air temperature (averaged over 24 h), maximum daily air temperature, minimum daily air temperature, wind speed (averaged over 24 h), and rainfall (summed over 24 h). The daily extraterrestrial radiation can be estimated from the solar constant, the solar declination and the time of the year (Allen *et al.*, 1998).

A major problem associated with HYREX during the study period was data discontinuity. In total, 1098 daily records from 1993–1996 were obtained after the missing data were taken out. All data were normalized before analysis by mapping the mean to zero and the standard deviation to 0·5. The process of normalization attempts to equalize the relative numerical significance between the input variables and to aid the analysis routines to perform efficiently, especially in the absence of any prior knowledge regarding input variable relevance. The asymptotic nature of the Gamma statistic remained valid for the normalized data. The training and testing data sets were selected by the randomization of the input data.

RESULTS AND DISCUSSION

Data analysis and model input selection based on the gamma test

The GT is able to provide the best mean square error that can possibly be achieved using any nonlinear

DOI: 10.1002/hyp

Table I. Gamma Test results on the daily sunshine data set

Parameters	Different combinations						
	ETR, T_{max}, *U* T_{mean}, T_{min}, *P*	T_{max}, *U* T_{mean}, T_{min}, *P*	*ETR*, *U* T_{mean}, T_{min}, *P*	*ETR*, T_{max} T_{mean}, T_{min}, *P*	*ETR*, T_{max}, *U* T_{min}, *P*	*ETR*, T_{max}, *U* T_{mean}, *P*	*ETR*, T_{max}, *U* T_{mean}, T_{min}
Gamma (Γ)	0·0354	0·0684	0·0434	0·0357	0·0361	0·0378	0·0401
Gradient (*A*)	0·1108	0·2254	0·1140	0·1731	0·1674	0·1265	0·1914
Standard error	0·0019	0·0046	0·0020	0·0019	0·0025	0·0020	0·0037
V-ratio	0·1438	0·2737	0·1736	0·1731	0·1430	0·1431	0·1605
Near neighbours	10	10	10	10	10	10	10
M	1098	1098	1098	1098	1098	1098	1098
Mask	111111	011111	101111	110111	111011	111101	111110

Note: Different combinations compared to study the input effects (inclusion and exclusion indicated by 1 or 0 in the mask).

smooth models. In this study, different combinations of input data were explored to assess their influence on the solar radiation modelling (Table I). There were $2^n - 1$ meaningful combinations of inputs; from which, the best one can be determined by observing the Gamma value, which indicates a measure of the best MSE attainable using any modelling method for unseen smooth functions of continuous variables. In Table I, some very interesting variations of the best MSE (Γ) are observed with different input combinations. The minimum value of Γ was observed when all available input data sets were used, i.e. extraterrestrial radiation (*ETR*), daily precipitation (*P*), daily mean temperature (T_{mean}), daily maximum temperature (T_{max}), daily minimum temperature (T_{min}) and daily mean wind velocity (*U*). The gradient (*A*) is considered to be an indicator of model complexity. A model with low MSE and low gradient is considered to be the best scenario for modelling. V-ratio is a measure of the degree of predictability of given outputs using available inputs. A smaller value of V-ratio was observed all inputs were considered.

The quantity of available input data to predict the desirable output was analysed using the M-test. The M-test results help to determine whether there were sufficient data to provide an asymptotic Gamma estimate and subsequently a reliable model. The M-test analysis results are shown in Figure 1. The test produced an asymptotic convergence of the Gamma statistic to a value of 0·0354 at around 770 data points (i.e. $M = 770$). The variation of the standard error (SE) corresponding to the data points is shown in Figure 2. In the figure it can be seen that the SE corresponding to $M = 770$ is very small at ∼0·0019, which shows the precision and accuracy of the Gamma statistic. M-tests were also performed in different dimensions, varying the number of inputs to the model (Table I), which clearly presents the response of the data model to different combinations of input data sets. From the table it can be deduced that the combination of precipitation, daily maximum temperature, daily minimum temperature and extraterrestrial radiation (*ETR*) gives a good model, comparable with the combination using all the inputs. The significance of the wind velocity and daily mean temperature data sets was relatively small when compared with other input sets since the elimination of these inputs made little difference to the Gamma statistic. The M-test analysis results for different scenarios are shown in Figure 3: scenarios are 'All', 'No ETR', 'No T_{max}', 'No T_{mean}', 'No T_{min}', 'No P' and 'No U'. ETR is observed to be the most significant input in solar radiation modelling; the 'No ETR' scenario resulted in a very high value of Gamma statistic.

The embedding 111111 model (a six input and one output set of I/O pairs) was identified as the best structure

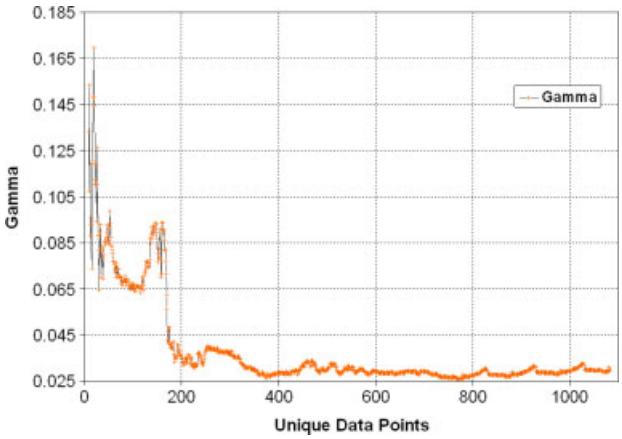

Figure 1. Gamma statistic (Γ) for the data set (M = 1098, p = 10)

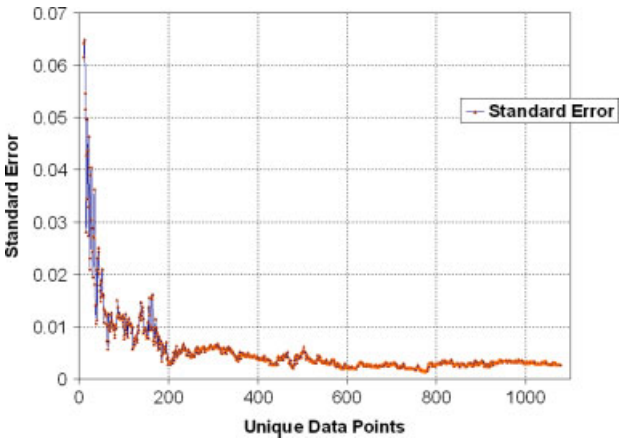

Figure 2. Variation of standard error for the data set (M = 1098, p = 10)

DOI: 10.1002/hyp

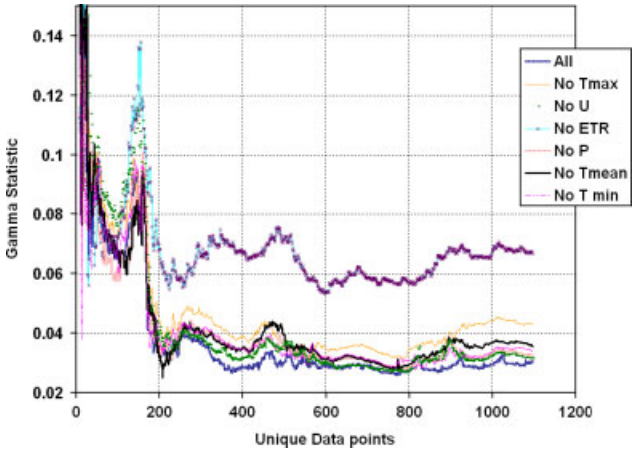

Figure 3. Variation of Gamma statistic (Γ) for the data corresponding to different combination of input data sets

because of its low noise level (Γ value), the rapid decline of the M-test SE graph (Figure 2), low V-ratio value (indicating the existence of a reasonably accurate smooth model), the regression line fit with slope A = 0·1108 (low enough for a simple nonlinear model with minimum complexity) and good fit with SE 0·0019. These values together give a clear indication that it is adequate to construct a nonlinear predictive model using around 770 data points with an expected MSE around 0·0354.

Nonlinear model construction and testing

In this study, two types of model were constructed and tested to predict daily solar radiation (LLR and ANN). The ANN were trained using the BFGS algorithm and the conjugate gradient algorithm. The nonparametric procedure based on LLR models does not require training in the same way as that of NN models. The optimal number of nearest neighbours for LLR (principally dependent on the noise level) was determined by a trial and error method and 16 nearest neighbours were implemented. The performance of the LLR technique was compared with that of the NN models using three global statistics (correlation efficiency, root mean squared error and mean bias error), as shown in Table II. Figures 4 and 5 show scatter plots of the computed (using LLR model with $p = 16$) and observed daily solar radiation during the training and validation periods. Figure 6 shows the observed and estimated solar radiation using the LLR model for 770 data points, which resulted in the minimum overall RMSE value of 1·79 MJ m^{-2} day^{-1}, which is 19·1% of the oserved daily solar radiation and the mean bias error (MBE) was observed to be −0·069 MJ m^{-2} day^{-1}.

In this study, various hidden layer neuron number combinations were tested for the ANN models. A feed forward 6-9-9-1 NN was constructed and trained using the BFGS algorithm and conjugate gradient algorithm and the performance was compared with that of the LLR model (shown in Table II). The size of training data was already determined as 770 data points through M-test analysis, and the target mean-squared error (MSError)

Table II. comparison of some basic performance indices of LLR and ANN models in daily solar radiation estimation

Models and algorithms	Training data (770 data points)				Validation data			
	RMSE* (MJ m^{-2} day^{-1} and %)	R^2	Slope	MBE (MJ m^{-2} day^{-1})	RMSE* (MJ m^{-2} day^{-1} and %)	R^2	Slope	MBE MJ m^{-2} day^{-1}
LLR	1·79 (19·1)	0·93	0·96	−0·069	2·50 (22·4)	0·90	0·97	0·017
ANN (conjugate gradient)	2·60 (28·5)	0·88	0·93	−0·179	3·06 (28·1)	0·83	0·93	−0·014
ANN (BFGS)	2·83 (30·2)	0·87	0·94	−0·178	3·39 (29·7)	0·80	0·93	−0·022

* Root mean square error is also shown as percentage of the mean value of observed daily solar radiation.

DOI: 10.1002/hyp

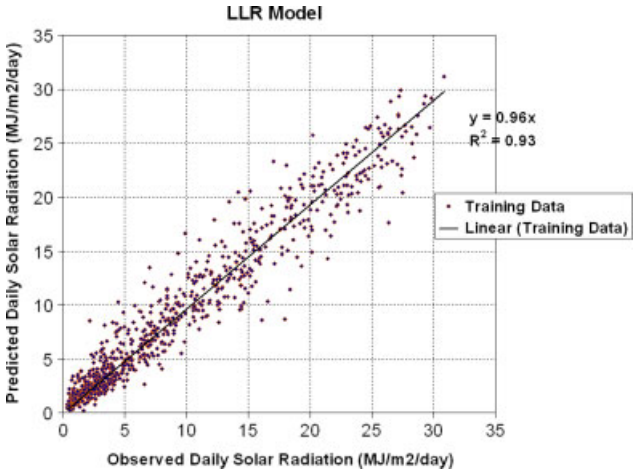

Figure 4. Observed versus LLR model of daily solar radiation in the training data set

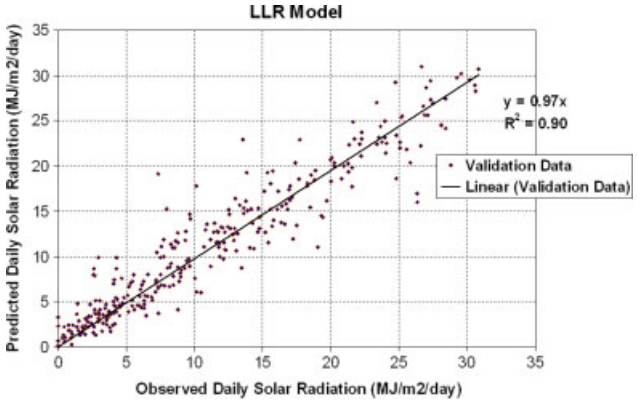

Figure 5. Observed versus LLR model of daily solar radiation in the validation data set

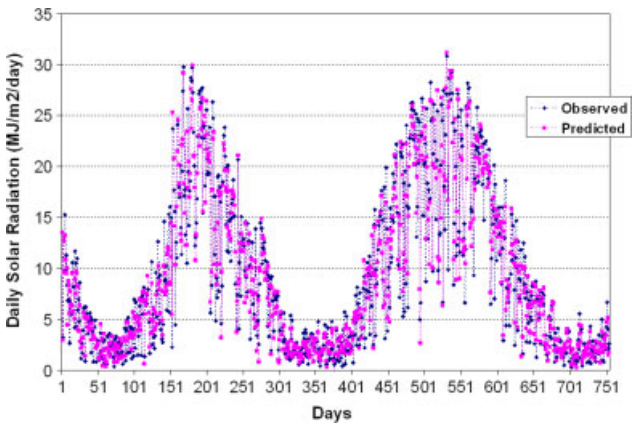

Figure 6. Solar radiation as observed (1993 to end 1995) and estimated using the LLR model for the training data set

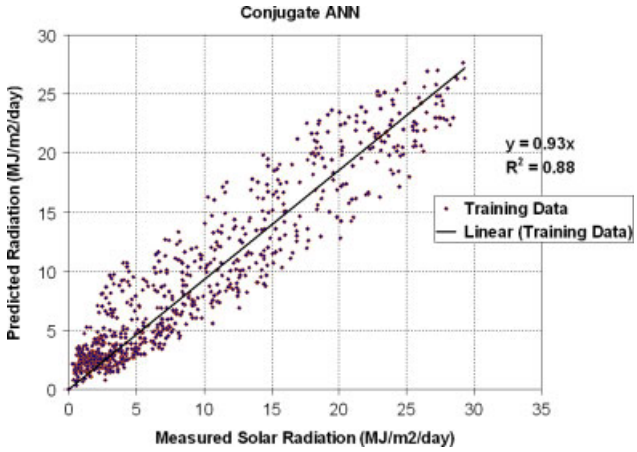

Figure 7. Observed versus ANN model of daily solar radiation for the training data set (conjugate gradient algorithm)

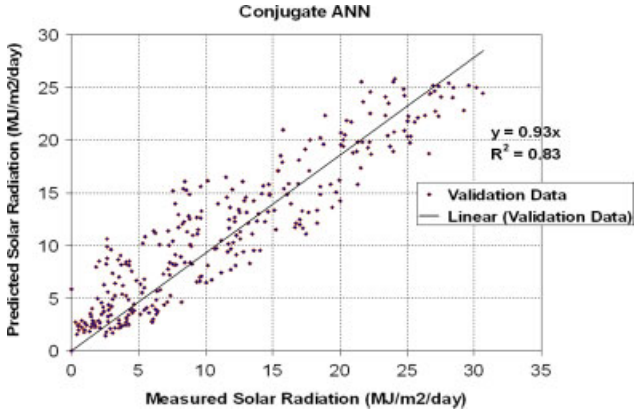

Figure 8. Observed versus ANN model of daily solar radiation for the validation data set (conjugate gradient algorithm)

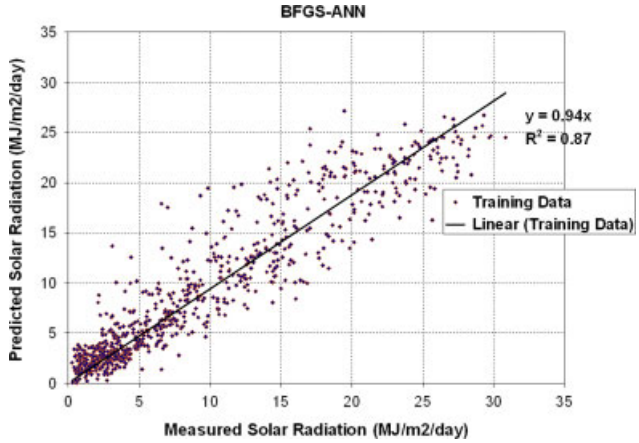

Figure 9. Observed versus ANN model of daily solar radiation for the training data set (BFGS algorithm)

was identified as 0·0354 (scaled) for $M = 770$. Scatter plots of training and validation results produced by the conjugate gradient algorithm based model are shown in Figures 7 and 8. The results predicted by the BFGS method based ANN model for training and validation data are shown in the form of scatter plots in Figures 9 and10. The conjugate gradient ANN model performed better on the validation data set than the BFGS algorithm based ANN model, with an RMSE value of 3·06 MJ m^{-2} day^{-1} (28·1% of mean observed solar radiation) and MBE value of $-0{\cdot}014$ MJ m^{-2} day^{-1}, whereas the latter produced 3·39 MJ m^{-2} day^{-1} (29·7% of mean observed solar radiation) and $-0{\cdot}022$ MJ m^{-2} day^{-1}, respectively. It is seen that the LLR model had superior performance to the BFGS and conjugate gradient ANN models. From Figures 9 and 10 one finds that both ANN

DOI: 10.1002/hyp

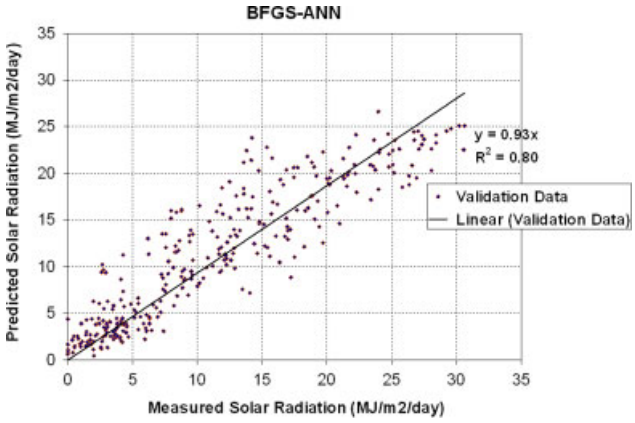

Figure 10. Observed versus ANN model of daily solar radiation in the validation data set (BFGS algorithm)

based models struggle to reproduce the highest values. At the same time, the LLR model is free from this handicap. Comparative analysis of these models using some basic statistics has been carried out and is shown in Table II, where the LLR model is shown to outperform both ANN models and provides the best performance, i.e. the lowest RMSE and highest R^2, for the training period and validation periods. The results of the study also indicate that the predictive capability of the BFGS algorithm is poor compared with those of conjugate gradient networks in daily solar radiation modelling.

CONCLUSIONS

This article describes a new approach to estimate daily solar radiation from meteorological data sets usuing the gamma test in combination with nonlinear modelling techniques. The study successfully demonstrated the informative capability of the GT in the selection of relevant variables in the construction of nonlinear models for daily (global) irradiation estimations. In this study, four relevant variables were used to estimate the daily solar radiation (extraterrestrial radiation, temperature, precipitation, and wind velocity). The quantity of data required to construct a reliable model was determined using the M-test, which identified $M = 770$ as a sufficient data scenario. The use of nonlinear modelling methods such as the LLR and ANNs with the BFGS NN training algorithm and conjugate gradient training algorithms has been demonstrated. Both the radial BFGS NN training algorithm and the conjugate gradient training algorithm performed reasonably well in modelling the validation data but both failed to reach the highest possible values. Among them, the conjugate gradient training algorithms was shown to be superior because of its better performance. The LLR technique was able to provide more reliable estimates than the ANN models. It would be interesting to explore this further in other catchments to confirm if similar results could be repeated.

The methodology described in the study might have significant implications for other types of hydrological modelling. In past decades, hydrologists have struggled to find an objective way of deciding the required data length for model calibrations. At the moment, the rule of thumb (e.g. six years) is still popular albeit such a method lacks consideration of the data characteristics (for example, how does one decide the required data length for a river like the Nile where there is only one peak flow per year or a typical river in England where many peak flows could be observed in the same period). Furthermore, if two similar floods are recorded, a hydrologist might consider one of them as a duplicate which provides little extra information for modelling. However, from the statistical point of view, a duplicated flood is still valuable and would help narrow down uncertainty bounds for model calibration. Also, there is no effective way for the hydrological community to check input data quality (rainfall, flow, temperature, wind, solar radiation, etc.) for hydrological models; this has hampered many model comparison activities. If the innate errors in the input data exceed the model's capability, it is very difficult for the model to perform, no matter how good the model itself is. In this regard, the GT presented in this study has the potential to help hydrologists to solve the uncertainty issues in the hydrological modelling process. It is hoped that this study will stimulate further exploration of this new technique in hydrology.

REFERENCES

Agalbjörn S, Končar N, Jones AJ. 1997. A note on the gamma test. *Neural Computing and Applications* **5**(3): 131–133. ISSN 0-941-0643.

Akinoglu BG, Ecevit A. 1990. A further comparison and discussion of sunshine based models to estimate global solar radiation. *Solar Energy* **15**: 865–872.

Alawi SM, Hinai HA. 1998. An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation. *Renewable Energy* **14**(1–4): 199–204. DOI:10·1016/S0960-1481(98)00068-8.

Allen R. 1995. Evaluation of procedures of estimating mean monthly solar radiation from air temperature. FAO: Rome.

Allen RG, Pereira LS, Raes D, Smith M. 1998. Crop evapotranspiration—guidelines for computing crop water requirements, irrigation and drainage. Paper 56, FAO: Rome.

Angstrom A. 1924. Solar and terrestrial radiation. *Quarterly Journal of the Royal Meteorological Society* **50**: 121–125.

Ball RA, Purcell LC, Carey SK. 2004. Evaluation of solar radiation prediction models in North America. *Agronomy Journal* **96**: 391–397.

Benson RB, Paris MV, Sherry JE, Justus CG. 1984. Estimation of daily and monthly direct, diffuse and global solar radiation from sunshine duration measurements. *Solar Energy* **32**: 523–535.

Bristow K, Campbell G. 1984. On the relationship between incoming solar radiation and daily maximum and minimum temperature. *Agricultural and Forest Meteorology* **31**: 159–166.

Chang JH. 1968. *Climate and Agriculture*. Aldine Publishing: Chicago.

Chuzhanova NA, Jones AJ, Margetts S. 1998. Feature selection for genetic sequence classification. *Bioinformatics* **14**(2): 139–143.

Corcoran J, Wilson I, Ware J. 2003. Predicting the geo-temporal variation of crime and disorder. *International Journal of Forecasting* **19**: 623–634. DOI:10·1016/S0169-2070(03)00095-5.

Donatelli M, Campbell GS. 1998. A simple model to estimate global solar radiation. In *Proceedings of the Fifth Congress of the European Society for Agronomy*. Nitra, Slovakia, II; 133–134.

Donatelli M, Bellocchi G, Fontana F. 2003. RadEst3·00: software to estimate daily radiation data from commonly available meteorological variables. *European Journal of Agronomy* **18**: 363–367. DOI:10·1016/S1161-0301(02)00130-2.

Donatelli M, Carlini L, Bellocchi G. 2006. A software component for estimating solar radiation. *Environmental Modelling & Software* **21**(3): 411–416. DOI:10·1016/j.envsoft.2005·04·002.

De Jong R, Stewart DW. 1993. Estimating global solar radiation from common meteorological observations in western Canada. *Canadian Journal of Plant Science* **73**: 509–518.

De Oliveira AG. 1999. *Synchronisation of chaos and applications to secure communications*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London.

Dinçer I, Dilmaç S, Türe IE, Edin M. 1996. Simple technique for estimating solar radiation parameters and its application for Gebze. *Energy Conversion and Management* **37**(2): 183–198. DOI:10·1016/0196-8904(95)00168-D.

Durrant PJ. 2001. *winGamma: A non-linear data analysis and modelling tool with applications to flood prediction*. PhD thesis, Department of Computer Science, Cardiff University, Wales, UK.

Evans D. 2002. *Data derived estimates of noise using near neighbour asymptotics*. PhD thesis, Department of Computer Science, Cardiff University, Wales, UK.

Evans D, Jones AJ. 2002. A proof of the gamma test. *Proceedings of Royal Society. Series A* **458**(2027): 2759–2799.

Fletcher R. 1987. *Practical Methods of Optimization*, 2nd edn. Wiley: New York.

Goh T, Tan K. 1977. Stochastic modelling and forecasting of solar radiation data. *Solar Energy* **19**(6): 755–757. DOI:10·1016/0038-092X(77)90041-X.

Gopinathan KK. 1988. A general formula for computing the coefficients of the correlations connecting global solar radiation to sunshine duration. *Solar Energy* **41**: 499–502.

Han D, Kwong T, Li S. 2007. Uncertainties in real-time flood forecasting with neural networks. *Hydrological Processes* **21**: 223–228. DOI: 10·1002/hyp.6184.

Hargreaves G, Samani Z. 1982. Estimating potential evapotranspiration. *Journal of Irrigation Drainage Engineering, ASCE* **108**: 225–230.

Hook JE, McClendon RW. 1992. Estimation of solar radiation data missing from long-term meteorological records. *Agronomy Journal* **84**: 739–742.

Jones AJ. 2004. New tools in non-linear modelling and prediction. *Computational Management Science* **1**: 109–149. DOI: 10·1007/s10287-003-0006-1.

Jones AJ, Tsui A, de Oliveira AG. 2002. Neural models of arbitrary chaotic systems: construction and the role of time delayed feedback in control and synchronization. *Complexity International* Vol 09, p. tsui01: 1–9 (with supplementary material in www.cs.cf.ac.uk/user/Antonia.J.Jones/GammaArchive/Complexity-InternationalPaper/SuppMat.html).

Jones PG, Thornton PK. 2000. MarkSim: software to generate daily weather data for Latin America and Africa. *Agronomics Journal* **92**: 445–453.

Kemmoku Y, Orita S, Nakagawa S, Skakibara T. 1999. Daily insolation forecasting using a multistage neural network. *Solar Energy* **66**(3): 193–199. DOI:10·1016/S0038-092X(99)00017-1.

Končar N. 1997. *Optimisation methodologies for direct inverse neurocontrol*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London.

Liu DL, Scott BJ. 2001. Estimation of solar radiation in Australia from rainfall and temperature observations. *Agricultural and Forest Meteorology* **106**(1): 41–59. DOI:10·1016/S0168-1923(00)00173-8.

McCulloch WS, Pitts W. 1943. A logical calculus of the ideas imminent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–133.

Minsky M, Papert S. 1969. *Perceptrons*. MIT Press: Cambridge, MA.

Mohandes M, Balchonaim A, Rehman KS, Halawani TO. 1998. Estimation of global solar radiation using artificial neural networks. *Renewable Energy* **14**(1–4): 179–184. DOI:10·1016/S0960-1481(98)00065-2.

Mustacchi C, Cena V, Rocchi M. 1979. Stochastic simulation of hourly global radiation sequences. *Solar Energy* **23**(1): 47–51. DOI:10·1016/0038-092X(79)90042-2.

Negnevitsky M, Le TL. 1995. Artificial neural networks application for current rating of overhead lines. *IEEE Transactions on Neural Networks* **1**: 418–422.

Noia M, Ratto CF, Festa R. 1993a. Solar irradiance estimation from geostationary satellite data. I. Statistical models *Solar Energy* **51**(6): 449–456. DOI:10·1016/0038-092X(93)90130-G.

Noia M, Ratto CF, Festa R. 1993b. Solar irradiance estimation from geostationary satellite data. II. Physical models. *Solar Energy* **51**(6): 457–465. DOI:10·1016/0038-092X(93)90131-7.

Ogulata RT, Ogulata SN. 2002. Solar radiation on Adana, Turkey. *Applied Energy* **71**(4): 351–358.

Penrose R. 1955. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society* **51**: 406–413.

Penrose R. 1956. On best approximate solution of linear matrix equations. *Proceedings of the Cambridge Philosophical Society* **52**: 17–19.

Prescott JA. 1940. Evaporation from a water surface in relation to solar radiation. *Transactions. Royal Society of South Australia* **64**: 148.

Rietveld MR. 1978. A new method for estimating the regression coefficients in the formula relating solar radiation to sunshine. *Agricultural Meteorology* **19**: 243–252.

Richardson CW. 1985. Weather simulation for crop management models. *Transactions of the ASAE* **28**: 1602–1606.

Rosenblatt F. 1962. *Priciples of Neurodynamics: Perceptrons and the Theory of Brain Mechanics*. Spartan.

Sabbagh JA, Sayigh AAM, El-Salam EMA. 1977. Estimation of total solar radiation from meteorological data. *Solar Energy* **19**: 307. DOI:10·1016/0038-092X(77)90075-5.

Saylan L, Sen O, Toros H, Arısoy A. 2003. Solar energy potential for heating cooling systems in big cities of Turkey. *Energy Conversion and Management* **43**(14): 1829–1837. DOI:10·1016/S0196-8904(01)00134-0.

Sfetsos A, Coonick H. 2001. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Solar Energy* **68**(2): 169–178. DOI:10·1016/S0038-092X(99)00064-X.

Stöckle CO, Nelson RL, Donatelli M, Castellvì F. 2001. ClimGen: a flexible weather generation pogram. In *Proceedings of the Second International Symposium Modelling Cropping Systems*, 16–18 July. Florence, Italy; 229–230.

Tan. SBK, Shuy EB, Chua LHC. 2007. Modelling hourly and daily open-water evaporation rates in areas with an equatorial climate. *Hydrological Processes* **21**: 486–499. DOI: 10·1002/hyp.6251.

Togrul IT, Onat E. 1999. A study for estimating solar radiation in Elazıg using geographical and meteorological data. *Energy Conversion and Management* **40**(14): 1577–1584. DOI:10·1016/S0196-8904(99)00035-7.

Tovar HF, Baldasano JM. 2001. Solar radiation mapping from NOAA AVHRR data in Catalonia, Spain. *Journal of Applied Meteorology* **40**: 1821–1834.

Tsui APM. 1999. *Smooth data modelling and stimulus-response via stabilisation of neural chaos*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London.

Tsui APM, Jones AJ, de Oliveira AG. 2002. The construction of smooth models using irregular embeddings determined by a gamma test analysis. *Neural Computing and Applications* **10**(4): 318–329. 10·1007/s005210200004.

Weiss A, Hays CJ. 2004. Simulation of daily solar irradiance. *Agricultural and Forest Meteorology* **123**(3–4): 187–199. DOI:10·1016/j.agrformet.2003·12·002.