Performance Evaluation of LLR, SVM, CGNN and BFGSNN Models to Evaporation Estimation

1ALIREZA MOGHADDAMNIA, 2MOHSEN GHAFARI GOSHEH, 3MEHRDAD NURAIE, 4MOHAMMAD ALIZADEH MANSURI and 5DAWEI HAN
1Assistant Professor of Hydrology, Department of Range and Watershed Management, University of Zabol, Zabol, IRAN (ali.moghaddamnia@gmail.com)
2MSc Graduate in Combating Desertification, Department of Range and Watershed Management, University of Zabol, Zabol, IRAN
3MSc Student in Water Engineering and Hydraulic Structures, Department of Civil Engineering, Islamic Azad University, Central Tehran Branch, Tehran, IRAN (mehrdad.nuraie@gmail.com)
4MSc Graduate in Civil Engineering (Water), Department of Civil Engineering, Islamic Azad University of Mahabad, Mahabad, IRAN
5Reader in Civil and Environmental Engineering, Faculty of Engineering, University of Bristol, UK

Abstract: This study assessed the ability of three models of Local Linear Regression (LLR), Support Vector Machine (SVM) and Artificial Neural Network (ANN) to estimate evaporation from reservoirs as one of the critical components of hydrological cycle in arid and semi-arid regions. A case study has been carried out in the Chahnimeh water reservoirs of Zabol located in the Sistan plain of Iran. Among the models used, in terms of the evaluation criteria of root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination ($R^2$), it is demonstrated that use of the nonlinear models of support vector machine (SVM), Broyden-Fletcher-Goldfarb-Shanno neural network (BFGSNN) and Conjugate Gradient neural network (CGNN) performed reasonably well in modeling the validation data compared to Local Linear Regression (LLR) model but both BFGSNN and CGNN failed to reach the highest possible values. In the meantime, the SVM model was able to provide more reliable estimations compared to others.

Key-Words: Evaporation, SVM, ANN, LLR, modeling, Chahnimeh reservoirs of Zabol

1 Introduction
Evaporation as one of the most important hydrological losses should be considered in the design of surface water reservoirs, in particular, in arid and semi-arid regions. Thus, accurately estimating water losses by evaporation from water reservoirs is of vital importance to water resources optimum allocation. The key factors affecting evaporation include solar radiation, temperature, relative humidity, saturation vapor pressure deficit, atmospheric pressure, and the wind speed. There are a massive number of empirical formulas for evaporation estimation. A comprehensive comparison among existing evaporation formulas was made by Ryan and Harleman (1973) [36]. They concluded that there is no significant discrepancy among these formulas. A number of direct and indirect methods are available for estimating potential evaporation from free water surfaces. The sole direct method is the U.S. Weather Bureau Class A pan measurement, which is 4 ft in diameter and 10 in. deep and is mounted on a timber grill about 6 in. above the soil surface. The indirect methods, in increasing order of complexity and data requirements, consist mainly of temperature-based formulas [43]; radiation-based approximations [45]; humidity-based formulas [34]; combination formulas, including allowance for humidity and wind speed [31]; or even more intensive evaluations of an energy balance at the evaporation surface [28]. These formulas and similar techniques have been comparatively used for evaporation estimation by many researchers [1,16,23,48,28,5]. Although these approaches are based directly on the Penman method, they are rather restrictive and sensitive to site-specific evaporation estimations, which can vary widely from one place to another. It is not possible to consider simultaneously all the factors affecting evaporation by any of the aforementioned methods over a period of time, due to existence of a set of restructured phenomenological (constant temperature, pressure, wind velocity, uniform environmental conditions, etc.) and procedural (linearity, homogeneity, isotropy, etc.) restrictive assumptions, which limit the applicability of any methodology except under specific environmental circumstances and meteorological conditions. Amongst the components of the hydrological cycle, evaporation is perhaps the most difficult to estimate owing to complex interactions between the components of the land-plant-atmosphere system [37]. In spite of the fact that there has been the large...
amount of literature published, most techniques reported are too demanding for observed weather data and prone to errors if local parameters are not available. Moreover, evaporation is a nonlinear, complex, and unsteady process so it is difficult to derive an accurate formula to represent all the physical processes involved. As a result, there is a new trend in using data mining techniques such as Fuzzy logic, Artificial Neural Networks and ANFIS to estimate evaporation. There are a large number of studies in which some hydrological processes are simulated by nonlinear models based on Artificial Neural Networks, Support Vector Machines, Fuzzy Logical system, Local Linear Regression, Bayesian Networks and so on. Keskin et al. (2004) examined the potential of the fuzzy logic approach in estimation of daily pan evaporation [21]. Kisi (2005) discusses the important viewpoints of the aforementioned paper. He points out the authors’ state in the second paragraph of the introduction section of the Keskin et al. paper: “The probabilistic, statistical, and stochastic approaches require large amounts of data for the modeling purposes and therefore are not practical in local evaporation studies. Thus it is necessary to adopt a better approach for evaporation modeling, such as the fuzzy logic principle used herein.”[22]. Keskin et al. (2004) describe that in fuzzy logic approaches, restrictions of data amounts are not a problem in modeling, even with incomplete and vague data [21]. Some typical studies reported thus far include ANNs in modeling daily soil evaporation [15], daily evapotranspiration [25], daily pan evaporation [39,41,21,42] and hourly pan evaporation [40]. From those reports, it is clear that ANN models are superior to the conventional regression models, since ANNs do not require any predetermination of regression forms. This advantage becomes more promising when an engineering problem is too complex to be represented by regression equations [40]. In comparison to a wider application of ANNs and SVMs in other fields of water sciences like flood forecasting[17,9,44,8,14,4,12,13,26,3], the experience of ANNs and SVMs application in evaporation estimation is still quite limited and there is a growing need for studying and trialing these techniques by researchers. There is still a huge gap in knowledge applying ANNs and SVMs in evaporation estimation and some conflicting results obtained for different climate regions from different researchers are a good indication that this technique is still in its infancy and more exploratory work is necessary to improve our understanding of these potentially robust tools from the Artificial Intelligence community. In the initial stages, modeling of evaporation variables was done using the traditional statistical models. In recent years, novel techniques have been proposed as robust modeling tools in hydrology and water management. The main objective of this study is the postulation and performance evaluating of LLR, SVM, BFGSNN and CGNN.

2 Material and Methods
2.1 Local Linear Regression (LLR)
The LLR technique is a widely studied nonparametric regression method which has been widely used in many low dimensional forecasting and smoothing problems. The advantage of LLR technique is that a reasonably reliable statistical modeling can be performed locally with a small amount of sample data. In 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 $p_{\text{max}}$, the number of near neighbors to be included for the local linear modeling. The method of choosing $p_{\text{max}}$ for linear regression is called influence statistics and is explained below. Given a neighborhood of $p_{\text{max}}$ points, we must solve a linear matrix equation of $Xm = y$, where $X$ is a matrix of the $p_{\text{max}}$ input points in d-dimensions, are the nearest neighbor points, $y$ is a column vector of length $p_{\text{max}}$ of the corresponding outputs, and $m$ is a column vector of parameters that must be determined to provide the optimal mapping from $X$ to $y$. The rank $r$ of the matrix $X$ is the number of linearly independent rows, which will affect the existence or uniqueness of the solution for $m$. If the matrix $X$ is square and non-singular then the unique solution to $Xm = y$ is $m = X^{-1} y$. If $X$ is not square or singular, we modify $Xm = y$ and attempt to find a vector $m$ which minimizes $|Xm - y|^2$, which was proved by Penrose (1955) where the unique solution to this problem is provided by $m = X^{-1} y$, where $X^{-1}$ is a pseudo-inverse matrix [33,32]. In this research, a kd-tree (short for k-dimensional tree) is used to organize the input training data, with a time-complexity in the order of $O$ (MlogM). A kd-tree is a space-partitioning data structure for organizing points in a k-dimensional space so that the LLR algorithms could be implemented using a minimum number of direct evaluations. More theoretical aspects of kd-tree can be found in [7,18].

\[\text{ISSN: 1790-5095} \quad 109 \quad \text{ISBN: 978-960-474-160-1}\]
2.2 Support Vector Machines (SVMs)

The foundation of SVMs has been developed principally by Vapnik and his collaborators [46,47]. Their formulation embodies the Structural Risk Minimization (SRM) principle, which has been shown to be superior to the more traditional Empirical Risk Minimization (ERM) principle employed by many of the other modeling techniques [30,11]. It is this difference that provides SVMs with a greater ability to generalize, which is the goal in statistical learning. SVMs have been proved to be effective in classification by many researchers in many different fields such as electric and electrical engineering, civil engineering, mechanical engineering, medical, financial and others [47]. SVMs can also be applied to regression problems by the introduction of an alternative loss function that is modified to include a distance measure [38]. Recently, it has been extended to the domain of regression problems [20]. Dibike et al. (2001) presented some results showing that Radial Basis Function (RBF) is the best kernel function to be used in SVM models [6]. Based on studies conducted by other researchers [12,26,3,4], only two kernel functions (linear and radial basis) have been explored further in hydrology applications. In this study, Radial Basis Function (RBF) function was used to estimate evaporation, this is due to this fact that RBF kernel function was superior to other kernel functions. There is a large amount of literature on SVMs. For more background on this topic, readers are referred to the references.

2.3 Artificial Neural Networks (ANNs)

ANNs is one of Artificial Intelligence techniques that mimic the behavior of the human brain; it attempts to describe a nonlinear relationship between the input and output of a complex system using historic process data. The idea of ANNs dates back to early 1940’s when McCulloch and Pitts developed the first computational representation of a neuron [27]. Later Rosenblatt proposed the idea of perceptrons in which single layer feed forward networks of McCulloch-Pitts neurons could carry out various computational tasks with the help of weights and training algorithm [35]. A very brief introduction on ANNs is given here and further details can be found in the textbook and the relevant references. 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 one is the most commonly used ANNs, 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 that employs gradient descent and gradient descent with momentum that are often too slow for practical problems because they require low learning rates for stable learning. Algorithms like Conjugate gradient, quasi-Newton, and Levenberg–Marquardt (LM) are considered as some of the faster algorithms, which 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 [29]. In practice nowadays, it is usually most effective to use two hidden layers [18]. In the recent decades, ANNs has received massive attention in hydrology and water sciences; despite it suffers from the disadvantage of a lack of explanation of their outcomes. In this study, two neural networks training algorithms of Broyden-Fletcher-Goldfarb-Shanno (BFGS) [10], and Conjugate Gradient (CG) were used to estimate evaporation with a two layer architecture embeded in WinGamma software.

2.4 Study Area

Chahnimeh reservoirs are located in the Sistan plain located in the Southeast of Iran, one of the driest regions of Iran. The Sistan plain has a very hot and dry climate. In summer, the temperature exceeds 50°C. Rainfall is about 60 mm/year and occurs only in autumn and winter and the open water evaporation is estimated at about 3200 mm/year. Strong winds in the region are an important contributing factor for the high evaporation. The daily weather variables used in this study consisted of eleven years (1983–2005) of daily records of air minimum temperature (T_{min}), air mean temperature (T_{mean}), air maximum temperature (T_{max}), wind speed (W), saturation vapor pressure deficit (Ed), mean relative humidity (RH_{mean}), 6:30 AM relative humidity (RH_{AM}), 12 noon relative humidity (RH_{noon}), 6:30 PM relative humidity (RH_{PM}), solar radiation (SR) and pan evaporation (E). The first nine years (1983–2004) data were used as training dataset and the remaining data were used as validation dataset.

2.5 Model Inputs Selection

Genetic algorithm (GA) can be used identify possible candidate variables for inclusion in model. In other words, it enables to determine which variables to include in modeling as sensitivity
analysis. Owing to the advancements in modern computing technology and development of a novel algorithm from the computing science community called the Gamma Test (GT), it has been possible to make significant progresses in tackling these problems [2,24]. It is achieved by the estimation of variance of the noise \( \text{var}(r) \) computed from the raw data using efficient, scalable algorithms. The novel technique of Gamma Test enables us 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, prior to model construction. This technique can be used to find the best embedding dimensions and time lags for time series analysis. This information would help us determine the best input combinations to achieve a particular target output. Overtraining is considered as one of the serious weaknesses associated with almost all nonlinear modeling techniques including ANNs, which lead to excellent results on the training data but very poor results on the unseen test data. The GT is designed to efficiently solve overtraining problem, as one of the serious weaknesses associated with almost all nonlinear modeling techniques, by giving an estimate of how closely any smooth model could fit the unseen data. In practice, the Gamma test can be achieved through winGamma™ software implementation [7]. One of the tools available within winGamma, genetic algorithm model identification, was used to inform and refine the choice of input variables. The Genetic Algorithm (GA) searches the space of all masks to find good embeddings. The parameters which can be used to control this search include population size, gradient fitness, intercept fitness, length fitness, and run time [19]. GA is particularly useful when the number of inputs is too large to make a full embedding practical. Genetic Algorithm was performed in different dimensions varying the number of inputs to the model, which clearly presented the response of the data model to some different combination of inputs data sets. The embedding model including a seven input and one output set of I/O pairs) was identified as the best structure because of its low noise level (\( \Gamma \) value), the rapid decline of the Genetic Algorithm SE graph, low \( V_{\text{ratio}} \) value (indicating the existence of a reasonably accurate smooth model), the regression line fit with slope \( A = 0.0680 \) (low enough as a simple non-linear model with a minimum complexity) and good fit with SE 0.00009. These values altogether can give a clear indication that it is quite adequate to construct a nonlinear predictive model using around 4018 data points with an expected MSE around 0.00408. Training data length identified as 2955 with least gamma value 0.00402 and SE 0.00009 for the best combination.

3 Modeling Results
For estimating evaporation, the LLR, SVM, CGNN and BFGSNN models were trained after splitting the datasets, and then validated using the validation dataset. Prior to execution of the models, standardization on the data was done such that all data values fell between 0 and 1. The results obtained from evaluating performance of the LLR, SVM, CGNN and BFGSNN models, in terms of statistical criteria including root mean square error (RMSE), mean absolute error (MAE), and determination coefficient \( R^2 \) are given in Table 1. As can be seen in the table, the SVM and CGNN models represent more consistent estimates with the pan evaporation measurements. In addition, the results obtained from SVM, CGNN and BFGSNN but LLR are almost similar to the pan measurements with \( R^2 = 0.934, 0.934, 0.933 \) and 0.878, and RMSE=2.284, 2.291, 2.310 and 3.121, respectively.

Table 1: Comparison of the models used in this study

<table>
<thead>
<tr>
<th>Models</th>
<th>RMSE (mm/day)</th>
<th>MAE (mm/day)</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>2.932</td>
<td>2.126</td>
<td>0.900</td>
</tr>
<tr>
<td>LLR</td>
<td>3.428</td>
<td>1.841</td>
<td>0.863</td>
</tr>
<tr>
<td>CGNN</td>
<td>2.972</td>
<td>2.199</td>
<td>0.897</td>
</tr>
<tr>
<td>BFGS</td>
<td>2.939</td>
<td>2.132</td>
<td>0.899</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>2.284</td>
<td>1.677</td>
<td>0.934</td>
</tr>
<tr>
<td>LLR</td>
<td>3.121</td>
<td>1.715</td>
<td>0.878</td>
</tr>
<tr>
<td>CGNN</td>
<td>2.291</td>
<td>1.713</td>
<td>0.934</td>
</tr>
<tr>
<td>BFGS</td>
<td>2.310</td>
<td>1.730</td>
<td>0.923</td>
</tr>
</tbody>
</table>

Fig.1: Curves of observed and estimated evaporations using SVM, LLR, CGNN and BFGSNN (validation).

Fig. 1 shows the curves of daily pan evaporation versus simulated evaporation using LLR, SVM, BFGSNN, and CGNN models. As shown in the figure, based on graphically inspecting of these curves, the best and worst accordance exist between observations and results obtained from SVM and LLR, respectively. Fig.2 shows scatter plots of the estimated using the models and observed daily evaporation during the validation period.
4 Conclusion
Based on the results obtained from GA, the best combination on model inputs includes $T_{\text{mean}}$, $T_{\text{max}}$, $R_{\text{HAM}}$, $R_{\text{PM}}$, $R_{\text{mean}}$, $W$ and $SR$. In addition, both BFGSNN and CGNN models performed reasonably well in modeling the validation data but SVM outperformed them in terms of performance criteria (minimum overall RMSE value). In the meantime, the LLR technique was not able to provide more reliable estimations compared with SVM, CGNN and BFGSNN models. It would be very interesting to explore these models further in other reservoirs around the world to confirm whether similar results could be repeated.

Fig. 2: Scatter plot of observed and estimated evaporations based on the models used in this study (validation).

References:


[38] Smola, A., Regression Estimation with Support Vector Learning Machines, Technische Universität Munchen, 1996.


