

Evaporation Estimation Using Support Vector Machines Technique

A. Moghaddamnia, M. Ghafari, J. Piri, and D. Han

Abstract—This paper is aimed at presenting a preliminary study on evaporation, as a major component of the hydrologic cycle that plays a key role in water resources development and management in arid and semi-arid regions. Debate remains as to the optimal method for estimating evaporation from free water surfaces of reservoirs. This paper investigates the abilities of Support Vector Machines (SVMs) technique to improve the accuracy of daily evaporation estimation in the Chahnimeh reservoirs of Zabol in the southeast of Iran. This paper describes SVMs technique for simulating evaporation, whose performance compared favorably based on performance criteria such as root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE) and coefficient of determination (R^2) have been employed for comparison of the results obtained from this technique with some empirical methods. In this paper, the Gamma Test (GT) has been used for the first time in modeling one of the key hydrological components: evaporation estimation modeling.

Keywords—Evaporation, support vector machine, gamma test, chahnimeh reservoirs.

I. INTRODUCTION

EVAPORATION takes place whenever there is a vapour pressure deficit between a water surface and the overlying atmosphere and sufficient energy is available. The most common and important factors affecting evaporation are solar radiation, temperature, relative humidity, vapour pressure deficit, atmospheric pressure, and the wind speed. Evaporation losses should be considered in the design of various water resources and irrigation systems. In areas with little rainfall, evaporation losses can represent a significant part of the water budget for a lake or reservoir, and may contribute significantly to the lowering of the water surface elevation [1]. Therefore, accurate estimation of evaporation loss from the water body is of primary importance for monitoring and allocation of water resources, at farm scales as well as at regional scales. A large number of experimental formulae exist for evaporation estimation. These methods all have their strengths and weaknesses. Ryan and Harleman made detailed comparisons

of existing evaporation formulae. They concluded that the discrepancies between these formulae were not significant [2].

Evaporation estimations are important in almost every aspect of water resources engineering including water supply, distribution, management, irrigation, agriculture, and hydrological practices. There are direct and different indirect methods 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, include temperature-based formulas [3]; radiation-based approximations [4]; humidity-based formulas [5]; combination formulas, which include allowance for humidity and wind speed [6]; or even more intensive evaluations of an energy balance at the evaporation surface [7]. These and similar methods have been used and compared for evaporation estimation by many researchers [8], [9], [10], [11], [7], [13]. 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 other. It is not possible to consider simultaneously all the factors affecting evaporation by any of the aforementioned approaches over a period of time, because there is 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. Among 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 [13]. The existing methods can be categorized as direct and indirect methods. Evaporation pans are commonly used to estimate evaporation from lakes and reservoirs as direct methods.

In this paper, methods of Support Vector Machines (SVMs) are proposed for estimating evaporation from a water surface of reservoirs. The modeling of estimating evaporation from surface reservoirs is a very active field of study and definitely there still is a lot of work to be done. In the initial stages, modeling of evaporation variables was done using the traditional statistical models. In recent years, modern techniques have been proposed as efficient modeling tools.

A. Moghaddamnia, Assistant Professor of Hydrology, is with Department of Watershed and Range Management, Faculty of Natural Resources, University of Zabol, Iran (e-mail: ali.moghaddamnia@gmail.com).

M. Ghafari, MSc. student of Combating desertification, Department of Range and Watershed Management, Faculty of Natural Resources, University of Zabol, Iran.

J. Piri is with Graduate of Irrigation, Zabol, Iran.

D. Han is with Civil and Environmental Engineering Department, University of Bristol, UK.

Here is a large pool of these techniques, and hence there is always a need to investigate which technique is the most efficient for a particular application. The modeling of estimating evaporation from surface reservoirs is a very active field of study and definitely there still is a lot of work to be done. In the initial stages, modeling of evaporation variables was done using the traditional statistical models. In recent years, modern techniques have been proposed as efficient modeling tools. Here is a large pool of these techniques, and hence there is always a need to investigate which technique is the most efficient for a particular application.

Khan and Coulibaly conducted a comparative study between support vector machines, artificial neural networks and the traditional seasonal autoregressive model (SAR) in the forecasting of lake water levels. They observed that the support vector machine is generally compatible with the other two models, but when it comes to long-term forecasting, the support vector machine displays better performance [14]. Mukherjee et al conducted a study to predict chaotic time series using support vector machines. The performance of support vector machines stood out when compared to other approximation methods such as polynomial and rational approximation, local polynomial techniques and artificial neural networks [15]. Other forecasting applications that employed support vector machines include the work of Mohandes et al in the prediction of wind speed. They observed that the performance of the support vector machines is comparable to that of artificial neural networks [16].

Tripathi proposed a support vector machine (SVM) approach for statistical downscaling of precipitation at monthly time scale. It is shown that SVMs provide a promising alternative to conventional artificial neural networks for statistical downscaling, and are suitable for conducting climate impact studies. The SVMs has found wide application in the field of pattern recognition and time series analysis [17].

However, there are many problems with them. Many factors can introduce errors in pan evaporation measurement, such as debris in the water, animal activity in and around the pan, pan size, materials employed to construct the pan, exposure of the pan, strong winds, and measurement of water depth in the pan [18], [19]. Indirect methods include those that use meteorological data to estimate evaporation from other meteorological variables through empirically developed methodologies or statistical and stochastic approaches in addition to mass-balance based formulations. Both direct and indirect methods have been used for evaporation estimation studies by many researchers. However, many of indirect methods are not applicable in this study due to the limitation in data availability that among them only three empirical methods is presented in Table I.

TABLE I
EMPIRICAL FORMULAE USED FOR EVAPORATION ESTIMATION

Formula Name	Equation
Hefner	$E = 0.028 \times U \times (e_s - e_a)$
Lincare	$E = f(T, T_{dew}, Latitude)$

where, E: evaporation rate (mm/day), e_s : saturation vapour pressure (mm of Hg), e_a : actual vapour pressure (mm of Hg), U: average wind velocity (km/hr) at a height of 2 meters above the lake or surrounding land areas, T and T_{dew} : temperature and dew point.

II. SUPPORT VECTOR MACHINES

Please submit your manuscript electronically for review as e-mail attachments. The foundation of the subject of Support Vector Machines (SVMs) has been developed principally by Vapnik and his collaborators [20], [21]. 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 [22], [23]. It is this difference that provides SVM with a greater ability to generalize, which is the goal in statistical learning. SVM has 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 [21].

A. Statistical Learning Theory

In statistical learning theory [20], [21], the problem of learning an input-output relationship from a data set is generally viewed as that of choosing, from the given set of functions $f(x, \alpha)$, $\alpha \in \Lambda$ (where $x \in R_n$ is a random vector drawn independently from a fixed but unknown probability distribution function $P(x)$ and Λ is a set of parameters), the one that best approximates the output value y to every input vector x , according to a conditional distribution function $P(y | x)$, also fixed but unknown. The selection of the desired function is based on a training set of l independent and identically distributed observations $(x_1, y_1), \dots, (x_l, y_l)$ drawn according to $P(x, y) = P(x)P(y | x)$.

If one considers the expected value of the loss due to classification or estimation errors, given by the risk functional,

$$R(\alpha) = \int L(y, f(x, \alpha)) dP(x, y) \quad (1)$$

where $L(y, f(x, \alpha))$ is the discrepancy between the measured response y and the response $f(x, \alpha)$ provided by the learning machine, the goal is to find the function $f(x, \alpha_0)$ that minimizes this risk functional $R(\alpha)$ in the situation where the only available information is the training set.

B. Support Vector Regression

SVMs can also be applied to regression problems by the introduction of an alternative loss function that is modified to include a distance measure [24]. Let the observed variable y be has real value, and let $f(x, \alpha), \alpha \in \Lambda$ be a set of real functions that contains the regression function $f(x, \alpha_0)$. Considering the problem of approximating the set of data, $\{(x_1, y_1), \dots, (x_l, y_l), x \in \mathbb{R}^N, y \in \mathbb{R}\}$ with a linear function, $f(x, \alpha) = (w \cdot x) + b$, the optimal regression function is given by minimizing the empirical risk,

$$R_{emp}(w, b) = \frac{1}{l} \sum_{i=1}^l |y_i - f(x_i, \alpha)|_{\varepsilon} \quad (2)$$

with the most general loss function with ε -insensitive zone described as,

$$|y - f(x, \alpha)|_{\varepsilon} = \begin{cases} \varepsilon & \text{if } |y - f(x, \alpha)| \leq \varepsilon; \\ |y - f(x, \alpha)| & \text{otherwise} \end{cases} \quad (3)$$

The objective is now to find a function $f(x, \alpha)$ that has at most a deviation of ε from the actual observed targets y_i for all the training data, and at the same time is as flat as possible. This is equivalent to minimising the functional,

$$\Phi(w, \xi^*, \xi) = \|w\|^2 / 2 + C (\sum \xi^* + \sum \xi_i) \quad (4)$$

where C is a pre-specified value and ξ^*, ξ are slack variables representing upper and lower constraints on the outputs of the system (Fig. 1), as follows,

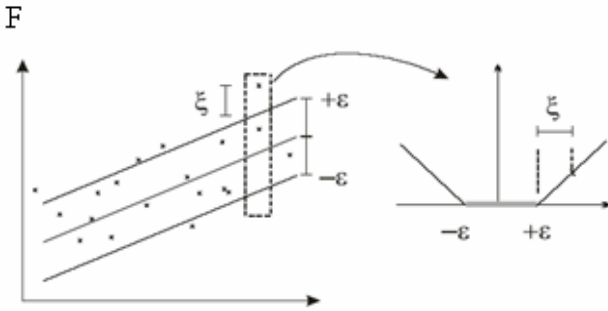


Fig. 1 Pre-specified accuracy ε and a slack variable ζ in SV regression

$$\begin{aligned} y_i - ((w \cdot x_i) + b) &\leq \varepsilon + \xi_i & i = 1, 2, \dots, l \\ ((w \cdot x_i) + b) - y_i &\leq \varepsilon + \xi_i^* & i = 1, 2, \dots, l \\ \xi_i^* &\geq 0 & \text{and } \xi_i \geq 0 \end{aligned} \quad (5)$$

Now the Lagrange function is constructed from both the objective function and the corresponding constraints by introducing a dual set of variables, as follows,

$$\begin{aligned} L = & \|w\|^2 / 2 + C \left(\sum_{i=1}^l (\xi_i + \xi_i^*) \right) - \\ & \sum_{i=1}^l \alpha_i [\varepsilon + \xi_i - y_i + (w \cdot x_i) + b] \\ & - \sum_{i=1}^l \alpha_i^* [\varepsilon + \xi_i^* + y_i - (w \cdot x_i) - b] - \sum_{i=1}^l (\eta_i \xi_i + \eta_i^* \xi_i^*) \end{aligned} \quad (6)$$

It follows from the saddle point condition that the partial derivatives of L with respect to the primary variables (w, b, ζ_i, ζ_i^*) have to vanish for optimality. Substituting the results of this derivation into Equation (4) yields the dual optimisation problem,

$$\begin{aligned} W(\alpha^*, \alpha) = & -\varepsilon \sum_{i=1}^l (\alpha_i^*, \alpha_i) + \sum_{i=1}^l y_i (\alpha_i^*, \alpha_i) - \frac{1}{2} \\ & \sum_{i=1}^l \sum_{j=1}^l (\alpha_i^*, \alpha_i) (\alpha_j^*, \alpha_j) (x_i \cdot x_j) \end{aligned} \quad (7)$$

that has to be maximised subject to the constraints:

$$\sum \alpha_i^* = \sum \alpha_i; \quad 0 \leq \alpha_i^* \leq C \quad \text{and} \quad 0 \leq \alpha_i \leq C \quad \text{for } i = 1, 2, \dots, l$$

Once the coefficients α_i^* and α_i are determined from Equation (7), the desired vectors can now be found as,

$$w_0 = \sum_{\text{support vectors}} (\alpha_i^*, \alpha_i) x_i \quad \text{and} \quad (8)$$

$$\text{therefore } f(x) = \sum_{\text{support vectors}} (\alpha_i^*, \alpha_i) (x_i \cdot x) + b_0$$

where $b_0 = -(1/2)w_0 \cdot [x_r + x_s]$.

Once again, when linear regression is not appropriate, as in the case of most engineering applications, a non-linear mapping kernel K is used to map the data into a higher-dimensional feature space where linear regression is performed. Once the optimum values α_{i0} and α_{i0}^* are obtained, then the regression function is given by: [25],

$$\begin{aligned} f(x) &= w_0 \cdot x + b_0 \\ \text{where} \\ w_0 \cdot x &= \sum_{\text{support vectors}} (\alpha_i^{0*} - \alpha_i^0) K(x_i, x) \quad \text{and } b_0 = \\ & -(1/2) \sum_{\text{support vectors}} (\alpha_i^{0*} - \alpha_i^0) [K(x_r, x_i) + K(x_s, x_i)] \end{aligned} \quad (9)$$

Recently, it has been extended to the domain of regression problems [26]. Dibike et al. presented some results showing

that Radial Basis Function (RBF) is the best kernel function to be used in SVM models [25].

III. GAMMA TEST, V-RATIO AND M-TEST

The Gamma test was firstly reported by Končar [27] and Agalbjörn, et al. [28], and later enhanced and discussed in detail by many researchers [29], [30], [31], [32], [33], [34].

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

$$\{(\mathbf{x}_i, y_i), 1 \leq i \leq M\} \quad (10)$$

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

$$y = f(\mathbf{x}_1 \dots \mathbf{x}_m) + r \quad (11)$$

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 r 's distribution is zero (since any constant bias can be subsumed into the unknown function f) and that the variance of the noise $\text{Var}(r)$ 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 Gamma Test is based on $N[i, k]$, which are the k th ($1 \leq k \leq p$) nearest neighbors $x_{N[i, k]}$ ($1 \leq k \leq p$) for each vector x_i ($1 \leq k \leq p$). Specifically, the Gamma Test is derived from the Delta function of the input vectors,

$$\delta_M(k) = \frac{1}{M} \sum_{i=1}^M \left| \mathbf{x}_{N(i, k)} - \mathbf{x}_i \right|^2 \quad (1 \leq k \leq p) \quad (12)$$

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

$$\gamma_M(k) = \frac{1}{2M} \sum_{i=1}^M \left| y_{N(i, k)} - y_i \right|^2 \quad (1 \leq k \leq p) \quad (13)$$

where $y_{N(i, k)}$ is the corresponding y -value for the k th nearest neighbor of x_i in Eq. (12). In order to compute Γ a least squares regression line is constructed for the p points $(\delta_M(k), \gamma_M(k))$.

$$\gamma = A\delta + \Gamma \quad (14)$$

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

$$\gamma_M(k) \rightarrow \text{Var}(r) \text{ in probability as } \delta_M(k) \rightarrow 0 \quad (15)$$

Calculating the regression line gradient can also provide helpful information on the complexity of the system under investigation. A formal mathematical justification of the method can be found in Evans and Jones [35].

The graphical output of this regression line (Eq.14) 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 modeling technique for unknown smooth functions of continuous variables [35]. Second, the gradient offers an indication of model's complexity (a steeper gradient indicates a model of greater complexity)

The Gamma test is a non-parametric method and the results apply regardless of the particular techniques used to subsequently build a model of f . We can standardize the result by considering another term V ratio, which returns a scale invariant noise estimate between zero and one. The Vratio is defined as,

$$V_{ratio} = \frac{\Gamma}{\sigma^2(y)} \quad (16)$$

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 Vratio close to zero indicates that there is a high degree of predictability of the given output y .

We can also determine the reliability of Γ statistic by running a series of Gamma test for increasing M , to establish the size of data set required to produce a stable asymptote. This is known as M-test. M-test result would help us to avoid the wasteful attempts of fitting the model beyond the stage where the MSE on the training data is smaller than $\text{Var}(r)$, which may lead to over fitting. The M-test also helps us to decide how much data we require to build a model with a mean squared error which approximates the estimated noise variance. In practice, the Gamma test can be achieved through winGamma™ software implementation [33]. Corcoran, et al. (2003), applied the Gamma Test 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 modeling [36].

IV. DATASETS

Chahnimeh reservoirs are located in the Sistan region that is one of arid regions generally characterized by water scarcity

and low per capita water allocation. The Sistan region is located in the Southeast of Iran, one of the driest regions of Iran and famous for its "120 day wind" (bād-e sad-o-bist-roz), a highly persistent dust storm in the summer which blows from north to south with velocities of nearly 20 knots. Hirmand River, originated from Afghanistan, is bifurcated into two branches when it reaches the Iranian border, namely Parian and Sistan. Sistan is the only water supply known in Sistan and Baluchistan province. It is the main stream of Hirmand River, which flows through Sistan plain and discharges into the natural swamp of Hamun-e-Hirmand. Sistan plain is essentially an inland delta with its major watercourses leading to a series of lakes.

The Sistan delta 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. The open water evaporation is very high and is estimated at 3200 mm/year. Strong winds in the region are quite unique and are an important contributing factor for the high evaporation. The Chahnimeh reservoirs are a series of natural depressions used primarily to store water for irrigation. However, they also play an important part in attenuating floods. During periods of high flows, water is diverted to these reservoirs via an intake and canal which has a capacity of up to 1000 m³ /s.

For better control over the distribution of the water reaching the Sistan irrigated plain, the Chahnimeh reservoirs were constructed on the Iranian side immediately downstream from the Hirmand fork, where the Helmand river separates into the Sistan and the Common Parian rivers. These reservoirs included three parts (Chahnimeh) and have been constructed for public water supply with a fourth reservoir under preparation. The present capacity of the Chahnimeh reservoirs is sufficient to guarantee a reliable supply for both the Sistan area as well as the agreed upon delivery for Zahedan, one of most important cities in southeast of Iran. The use of water for irrigated agriculture in Sistan is mainly restricted by the variability of the supply and not by the total supply; the use of Chahnimeh reservoirs for irrigation will improve the performance of irrigated agriculture.

The daily weather variables of automated weather station namely, Chahnimeh Station of Zabol (latitude 61°40' - 61°49' W, longitude 30°45' - 30°50' N) operated by the IR Sistan Baluchistan Regional Water (IR SBRW) are used in this study. The data sample 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 vapour 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). For the station of interest, the first nine years (1983–2004) data were used for training modes and the remaining data were used for validation.

V. RESULTS AND DISCUSSION

A. Data Analysis and Model Input Selection based on the Gamma Test

The Gamma Test is able to provide the best mean square error that can possibly be achieved using any nonlinear smooth models. In this study, different combinations of input data were explored to assess their influence on the Evaporation estimation modeling (Table II). In the research we using Genetic Algorithm for finding best combinations when using T_{mean}, T_{max}, RH_{AM}, RH_{PM}, RH_{mean}, W and SR combination we can have minimum value of Gamma (Γ). There were 2ⁿ - 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 modeling methods for unseen smooth functions of continuous variables.

TABLE II
GAMMA TEST RESULTS ON THE EVAPORATION ESTIMATION DATA SET

Parameters	Different combinations			
	All Variables without W	All Variables without T _{max} , RH _{noon}	All Variables without T _{max} , RH _{noon} , Ed	All Variables without T _{max} , SR
Gamma	0.007878	0.004112	0.00408	0.004868
Gradient	0.092253	0.054348	0.067984	0.011857
Standard Error	0.000225	0.000035	0.000098	0.000115
V-Ratio	0.14477	0.07555	0.07495	0.08943
Near Neighbours	10	10	10	10
Unique Points	4018	4018	4018	4018
Mask	1111111011	0111011111	0111011101	1011111110

In Table II, we can see some very interesting variations of the best MSE (Γ) with different input combinations that signer on the Fig. 2.

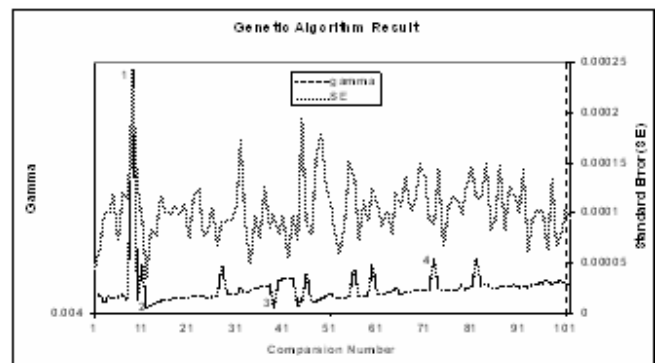


Fig. 2 Variation of Gamma statistic (Γ) for the data corresponding to different combination of input data sets

A model with low MSE and low gradient is considered as the best scenario for the modeling. V ratio is the measure of

degree of predictability of given outputs using available inputs. The smaller value of V_{ratio} was observed when we considered combination with T_{mean} , T_{max} , RH_{mean} , RH_{AM} , RH_{PM} , W and SR .

The quantity of available input data to predict the desirable output was analyzed using the M-test. The M-Test results would help us 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 Fig. 3.

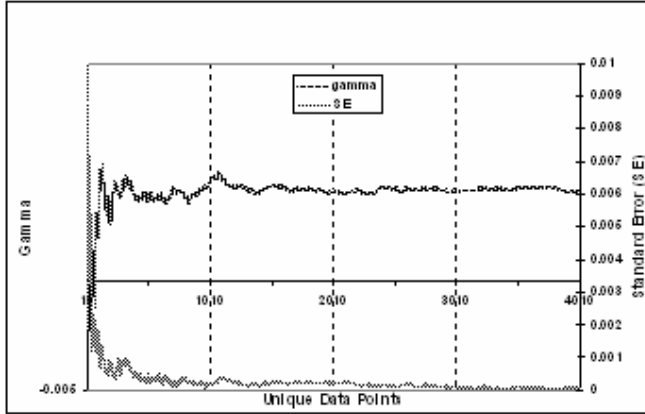


Fig. 3 Relation between Standard Error and Gamma Test for identified Tran data length

The test produced an asymptotic convergence of the Gamma statistic to a value of 0.00408 at around 4018 data points (i.e. $M=4018$). The variation of the Standard Error (SE) corresponding to the data points is shown in the Fig. 3. In the figure we can note that, the SE corresponding to $M=4018$ is very small at ~ 0.00009 , which shows the precision and accuracy of the Gamma statistic. We also performed Genetic Algorithm in different dimensions varying the number of inputs to the model (Table II), which clearly presented the response of the data model to some different combination of inputs data sets. The Genetic Algorithm analysis results in different scenarios are shown in the Fig. 2. The embedding 0111011101 model (a eight input and one output set of I/O pairs) was identified as the best structure because of its low noise level (Γ value), the rapid decline of the Genetic Algorithm SE graph (Fig. 2), low V_{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 (Fig. 3) for the best combination ("No T_{min} , RH_{noon} and Ed scenario).

B. Results and Conclusion

Among the seven meteorological variables considered, it is clear that some would play more important roles than others and it is important that only the significant ones are used as inputs for the final model. In this study, various combinations of these variables were examined to evaluate the impact of each variable. Root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE) and determination coefficient (R^2) criteria were used as evaluation criteria. The RMSE represents the deviation between simulated values and observed values.

$$RMSE = \left[\frac{\sum_{i=1}^n (p_i - o_i)^2}{N} \right]^{1/2} \quad (17)$$

The lower MAE values indicate more accurate estimations.

$$MAE = \frac{\sum_{i=1}^n |p_i - o_i|}{N} \quad (18)$$

MSE provide different types of information about the predictive capabilities of the model. The MSE measures the goodness-of-fit relevant to high evaporation values [37].

$$MSE = \frac{1}{N} \sum_{i=1}^N (o - p)^2 \quad (19)$$

R^2 measures are the ratio of the variability of the modeled values to the variability of the original data values.

$$R^2 = 1 - \left(\frac{\sum_{i=1}^n (o_i - p_i)^2}{\sum_{i=1}^n (o_i - \bar{o})^2} \right) \quad (20)$$

where o_i and p_i are the observed and predicted evaporation at time i , respectively; \bar{o} is the mean of the observed evaporation; and N is the number of data points.

TABLE III
COMPARISON BETWEEN ONE-CLASS SVM, EPSILON-SVR AND NU-SVR FOR EVAPORATION ESTIMATION

SVMs		One-Class SVM	epsilon-SVR	nu-SVR
Train	RMSE (%)	44.012	2.926	2.926
	MAE (mm/day)	41.102	2.122	2.141
	MSE (mm ² /day ²)	1937	8.563	8.562
	R^2	-21.7	0.899	0.899
Validation	RMSE (%)	43.36	2.277	2.349
	MAE (mm/day)	40.31	1.669	1.693
	MSE (mm ² /day ²)	1880	5.154	5.518
	R^2	-21.7	0.937	0.933

The architectures of the SVM model used for the Chahnimeh station, and also the results obtained from employing statistical criteria of performance are given in Table III.

At first, we compared SVM three type including one-class SVM, epsilon-SVR and nu-SVR that results are shown in the Table III. In this study, epsilon-SVR model shows better results. Also among four kernel functions including linear, polynomial, radial basis function (RBF), and sigmoid, using RBF kernel indicates better performance as given in Table IV. The table represents the four SVM models used in this research as well as their corresponding performance criteria of RMSE, MAE, MSE, and R^2 .

TABLE IV
COMPARISON BETWEEN ONE-CLASS LINEAR, POLYNOMIAL, RBF AND SIGMOID KERNEL FOR EVAPORATION ESTIMATION

Kernel Function Type		linear	polynomial	RBF	sigmoid
Training	RMSE (%)	3.407	3.140	2.926	14.22
	MAE (mm/day)	2.637	2.258	2.122	11.11
	MSE (mm ² /day ²)	11.611	9.859	8.563	202.2
	R ²	0.863	0.884	0.899	-1.37
Validation	RMSE (%)	2.644	2.494	2.277	12.90
	MAE (mm/day)	2.032	1.482	1.669	10.06
	MSE (mm ² /day ²)	6.992	6.221	5.184	166.5
	R ²	0.915	0.924	0.937	-1.01

It is concluded that the best input combination should include some variables in the order of their importance, that is, W, SR, RH_{mean}, T_{max}, RH_{AM}, RH_{PM}, and T_{mean}.

The combinatory architectures of the SVM models used for the Chahnimeh weather station, and also the results obtained from statistical criteria for models performance are given in Table III.

The models were trained after splitting the data into training datasets, and validation datasets. Prior to execution of the models, standardization, X_{i1} , on the data, X_i ($i=1, 2, \dots, n$) was done according to the following equation such that all data values fell between 0 and 1,

$$x_i^1 = (x_i - x_{\min}) / (x_{\max} - x_{\min}) \quad (21)$$

where the X_i is actual value; and the X_{\max} and X_{\min} are maximum and the minimum of the measurement values, respectively.

The One-class SVM, epsilon-SVR and nu-SVR, results are similar to the pan measurements with $R^2 = -21.7, 0.937$ and 0.933 , and RMSE = 43.36, 2.277 and 2.349, respectively.

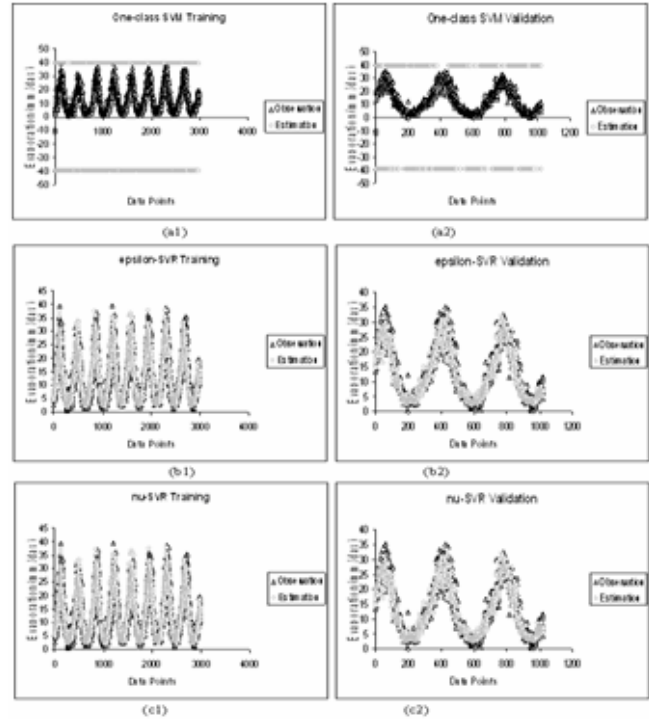


Fig. 4 Curves of observed and estimated evaporations using one-class SVM (a), epsilon-SVR (b) and nu-SVR(c). (1=Train & 2=Validation)

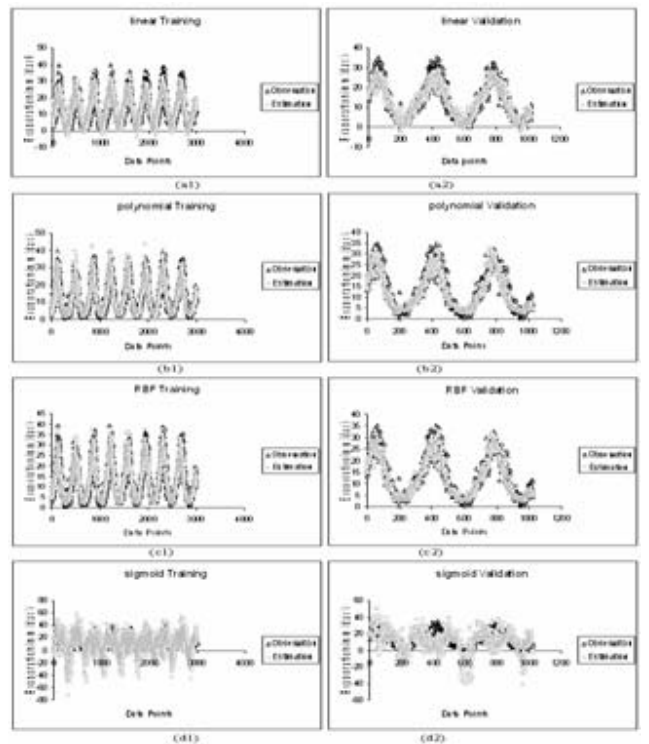


Fig. 5 Graphs of observed and estimated evaporations by linear (a), polynomial (b),RBF(c) and sigmoid(d) kernels. (1=Train & 2=Validation)

Chahnimeh reservoir lies in the southeast part of Iran, and its hydrological balance necessitates the estimation of

evaporation rates from a set of measured meteorological factors. Finally, the three types of epsilon-SVR kernel functions for evaporation estimation were compared with pan evaporation values, graphically and numerically. RBF kernel function in comparison with other kernels is better. The results obtained from employing linear, polynomial, RBF and sigmoid kernel functions for estimating evaporation indicate R^2 equal to 0.915, 0.924, 0.937 and -1.01, and RMSE equal to 2.644, 2.494, 2.277 and 12.90, respectively (Fig. 4). Fig. 5 shows graphs of observed and simulated evaporations based on linear, polynomial, RBF, and sigmoid kernel functions. Also Fig. 6 shows the curves of pan evaporation values versus evaporation values estimated by using empirical methods.

In this study, we used RBF kernel function for SVM models, because it represents the better performance than the other kernel functions.

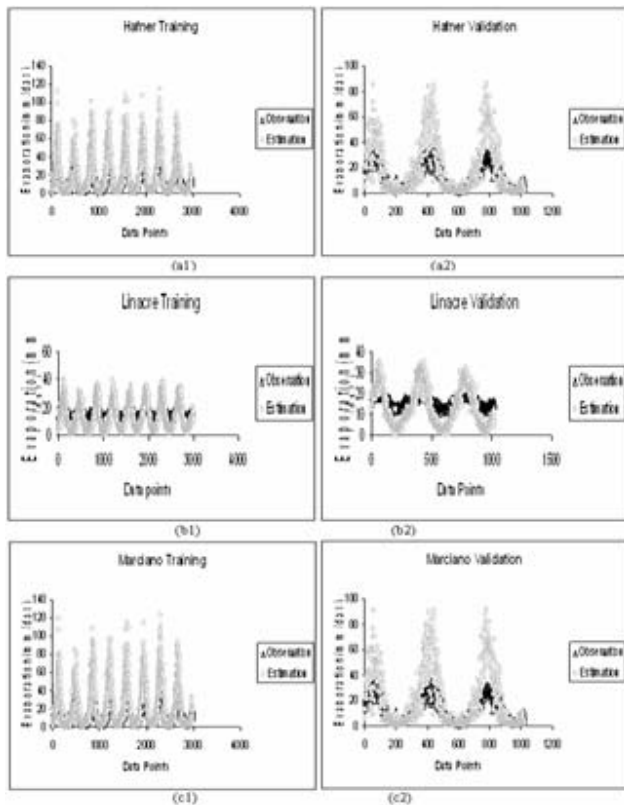


Fig. 6 Curves of observed and estimated evaporations by Hafner (a), Lincare (b) and Marciano (c). (1=Train & 2=Validation)

C. Conclusion

Also in four types of kernels in epsilon-SVR, RBF kernel was able for evaporation estimate of Chahnimeh reservoirs. Estimation results in comparison with empirical methods results illation those empirical methods in area are not able. Table V gives evaporation values estimated by using empirical formulae based on the performance criteria.

TABLE V
RESULT OBTAINED FROM EMPIRICAL METHODS FOR
EVAPORATION ESTIMATION

Empirical methods		Hafner	Lincare	Marciano
Training	RMSE (%)	43.23	63.54	43.00
	MAE (mm/day)	4.72	7.31	4.54
	MSE (mm ² /day ²)	33.34	66.78	31.54
	R ²	0.62	0.24	0.64
Validation	RMSE (%)	48.27	61.45	45.82
	MAE (mm/day)	5.32	6.90	5.01
	MSE (mm ² /day ²)	39.70	59.95	35.87
	R ²	0.49	0.23	0.54

This article describes a new approach to estimate daily evaporation from meteorological data sets with the Gamma Test in combination with nonlinear modeling techniques. The study successfully demonstrated the informative capability of the Gamma Test in the selection of relevant variables in the construction of non-linear models for daily (global) evaporation estimations. In this study, we used eight relevant variables for estimating the daily evaporation. The quantity of data required to construct a reliable model was determined using the M-Test, which has identified M=4018 as the sufficient data scenario.

The methodology described in the study might have significant implications for other types of hydrological modeling. If the innate errors in the input data exceed the model's capability, it would be very difficult for the model to perform, no matter how good the model itself is. For this purpose, the Gamma Test employed in this research would have a massive potential to help hydrologists in solving the uncertainty issues in hydrological modeling process. We hope this study will stimulate a further exploration of methodology mentioned in natural sciences and hydrology.

REFERENCES

- [1] R.H. McCuen, Hydrologic Analysis And Design, Prentice Hall, EnglewoodCliffs, 1998, NewJersey.
- [2] P.J. Ryan and O.R.F. Harleman, An Analytical and Experimental Study of Transient Cooling Pond Behavior. R.M. Parsons Laboratory, MIT, Technical Report, 1973, No. 161
- [3] C. W. Thornthwaite, Na approach toward a rational classification of climate. Geographical Reviews, 38, 1948, 55-94.
- [4] L. Turc, Estimation of irrigation water requirements, potential evapotranspiration: A simple climatic formula evolved up to date. Ann. Agron., 12, 1961,13-49.
- [5] V. A. Romanenko, Computation of the autumn soil moisture using a universal relationship for a large area, Proc., Ukrainian Hydrometeorological Research Institute, No. 3, Kiev, 1961.
- [6] H. L. Penman, Natural evaporation from open water, bare soil, and grass. Proc. R. Soc. London, 193, 1948, 120-145.
- [7] R. S. McKenzie and Craig, A. R, Evaluation of river losses from the Orange River using hydraulic modeling." J. Hydrol., 241(1-2), 2001, 62-69.
- [8] W. Abtew, Evaporation estimation for Lake Okeechobee in south Florida. J. Irrig. Drain. Eng., 127(3), 2001, 140-147.
- [9] C. L. Hanson, Prediction of Class A pan evaporation in southwest southwest Idaho. J. Irrig. Drain. Eng., 115(2), 1989. 166-17.

- [10] H. V. Knapp, Yu, Y. S and Pogge, E. C, Monthly evaporation for Milford Lake in Kansas. *J. Irrig. Drain. Eng.*, 110(2), 1984, 138– 148.
- [11] K. Warnaka and Pochop, L, Analyses of equations for free water evaporation estimates." *Water Resour. Res.*, 24(7), 1988, 979–984.
- [12] B. J. Choudhury, Evaluation of empirical equation for annual evaporation using field observations and results from a biophysical model. *J. Hydrol.*, 216(1-2), 1999, 99–110.
- [13] V. P. Singh and Xu, C. Y, Evaluation and generalization of 13 mass-transfer equations for determining free water evaporation, *Hydrol. Proc.*, 11, 1997, 311-324.
- [14] M. Khan and P. Coulibaly, Application of Support Vector Machine in Lake Water Level Prediction, *J. Hydrologic. Engrg.*, vol. 11, no. 3, 2006, pp. 199-205.
- [15] Mukherjee, S, E. Osuna and F. Girosi, Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines, *IEEE NNSP'97*, 1997, pp 24–26.
- [16] M. Mohandes., T. O. Halawani, S., Rehman and A.A. Hussain, Support vector machines for wind speed prediction, *Renewable Energy*, vol. 29, no. 6, 2004, pp 939–947
- [17] S. Tripathi., Srinivas, V. V., S. Nanjundiah, R, Downscaling of precipitation for climate change scenarios: A support vector machine approach. *Journal of Hydrology*. 330, 2006, 621– 640.
- [18] R.K. Farnsworth and E.S. Thompson, Mean Monthly, Seasonal, and Annual Pan Evaporation for the United States," NOAA Technical Report NWS 34, Washington, D.C., 1982, 82 p.
- [19] N.J. Rosenberg., Blad, B.L. and Verma, S.B, *Microclimate: The biological environment*. 2.ed. New York: John Wiley & Sons, 1983, 495p.
- [20] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
- [21] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
- [22] E. Osuna., Freund, R. and Girosi, F, An improved training algorithm for support vectormachines. In *Proc. of the IEEE Workshop on Neural Networks for Signal Processing VII*, New York, 1997, 276-285.
- [23] S. Gunn, *Support Vector Machines for Classification and Regression*, Image Speech & Intelligent Systems Group, University of Southampton, United Kingdom, 1998.
- [24] A. Smola, *Regression Estimation with Support Vector Learning Machines*, Technische Universitat Munchen, 1996.
- [25] Y.B. Dibikey., Velickov S., Sololatine D.P and Abbott M.B, Model induction with support vectre machine: Introdaction and application. *ASCE Journal of Computing in Civil Engineering*. 15, 2001, 208-216.
- [26] V. Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*, MIT Press, Cambridge, Massachusetts, London, England, 2001.
- [27] N. Končar, *Optimisation methodologies for direct inverse neurocontrol*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London, 1997.
- [28] S. Agalbjörn, Končar N, Jones AJ, A note on the gamma test. *Neural Computing and Applications*, 5(3), 1997, 131–133. ISSN 0-941-0643.
- [29] N.A. Chuzhanova, Jones AJ, Margetts S, Feature selection for genetic sequence classification. *Bioinformatics* 14(2), 1998, 139–143.
- [30] A. G. Oliveira, *Synchronisation of chaos and applications to secure communications*. Ph.D. thesis, Department of Computing, Imperial College, University of London, U.K, 1999.
- [31] APM. Tsui, *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, 1999.
- [32] APM. Tsui, Jones AJ, de Oliveira AG, The construction of smooth models using irregular embeddings determined by a gamma test analysis. *Neural Computing and Applications* 10(4), 2002, 318–329. 10.1007/s005210200004.
- [33] P.J. Durrant, *winGamma: A non-linear data analysis and modelling tool with applications to flood prediction*. PhD thesis, Department of Computer Science, Cardiff University, Wales, UK, 2001.
- [34] A. J. Jones, Tsui A, de Oliveira AG, Neural models of arbitrary chaotic systems: construction and the role of time delayed feedback in control and synchronization. *Complexity International Vol 09*, 2002.
- [35] D. Evans, Jones AJ, A proof of the gamma test. *Proceedings of Royal Society. Series A* 458(2027), 2002, 2759–2799
- [36] J. Corcoran, Wilson I, Ware J, Predicting the geo-temporal variation of crime and disorder. *International Journal of Forecasting*, 19, 2003, 623–634. doi:10.1016/S0169-2070(03)00095-5.
- [37] N. Karunanithi., Grenney, W.J., Whitley, D., Bovee, K, Neural networks for river flow prediction. *Journal of Computing in Civil Engineering* 8(2), 1994, 201–220.
- [38] D. Han and Yang, Z, River flow modelling using support vector machines. In *XXIX IAHR Congress, Beijing, China, 17–21 September*, pp. 494–499, 2001, Qinghua University Press, China.