Paddy Water Environ (2017) 15:711–722
DOI 10.1007/s10333-016-0571-x

ARTICLE

Daily pan evaporation modeling in climatically contrasting zones with hybridization of wavelet transform and support vector machines

Leeladhar Pammar[1] · Paresh Chandra Deka[1]

Received: 2 May 2016 / Revised: 14 October 2016 / Accepted: 3 December 2016 / Published online: 22 May 2017

Abstract The estimation of evaporation has been under surveillance, which is being carried out by many researchers toward applications in the fields related to hydrology and water resources management. Due to complexities associated with its estimation, research has employed several modes via direct and indirect methods to estimate. Accurate estimations are still the thrust area of research in these fields. The pan evaporation estimations with the help of data modeling techniques have provided better results in the recent past. The advancement in the field of data modeling has introduced several techniques which can best fit the data type and provide accurate estimations. The novel gamma test (GT) was used to decide the best input–output combination. Parameter optimization was carried out by grid search. The developed models gave better estimations of pan evaporation, but exhibited some limitations with nonlinearity, and sparse and noisy data. These limitations paved way for data pre-processing techniques such as wavelet transform. This study made an attempt to explore hybrid modeling using discrete wavelet transform (DWT) and support vector machines (SVR) for pan evaporation estimation. Two stations representing contrasting climatic zones namely 'Bajpe' and 'Bangalore' located in the state of Karnataka, India, are selected in this study. The meteorological datasets recorded at these stations are analyzed using gamma test and grid search to use the best input–output combinations for the models. The modeled pan evaporation estimations are very promising toward ever demanding accuracy expected in the associated fields.

Keywords Support vector machine · Wavelet transformation · Grid search · Gamma test · Regression · Pan evaporation

Introduction

Evaporation is an element of the hydrologic cycle, which can be generally estimated by the indirect methods such as mass transfer, energy budget, and water budget methods. The direct method such as pan evaporation (PE) is widely used to estimate the evaporation of lakes and reservoirs (Finch and Calver 2008). The setback associated with the direct methods is the subsequent application of coefficients based on the measurements from a small tank to large bodies of open water and also the accuracy of estimation from experimental models. Therefore, many researchers tried estimating the evaporation through the indirect methods using the climatic variables, but the difficulty experienced in the indirect methods is the requirement of data, which is not easily available particularly in developing countries (Burt et al. 2005).

The evaporation process is strongly nonlinear in nature. Few researchers emphasize the estimation of accurate evaporation in the research field using modeling techniques (Xu and Singh 2001). For efficient use of data for estimation of evaporation, data modeling may be considered as a suitable option. Several data modeling approaches have been utilized for estimation of evaporation. The data modeling techniques such as neural networks, adaptive network-based fuzzy inference system (ANFIS), and other techniques experienced some drawbacks such as

✉ Leeladhar Pammar
lohitnitks@gmail.com

[1] Department of Applied Mechanics and Hydraulics, National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India

complexity in their implementation, risk of over-fitting, and degraded performance with sparse data (Kim et al. 2012; Nourani and Sayyah 2012; Jothiprakash and Kote 2011). This allowed the emergence and use of support vector machine (SVM) in a variety of field applications (Deswal and Pal 2008).

In the last few years, drawbacks experienced in modeling hydrological features such as evaporation, evapotranspiration, and rainfall using artificial intelligence (AI) were better addressed by SVM (Kumar et al. 2007; Jain et al. 2008; Tabari et al. 2012; Deswal and Pal 2008, Jain et al. 2008). It is found to be working well in comparison to neural network and other similar techniques. There were few literature studies regarding hybridization of wavelet transform and other data-driven techniques in evaporation modeling (Guimarães Santos and Silva 2014).

In modeling hydrologic time series, sometimes signals are highly nonlinear and exhibit seasonal irregularity. Under such circumstances, SVM alone may not be able to cope with nonlinear data, if pre-processing of input and output data is not performed. In this context, wavelet transform may be utilized for data pre-processing. The wavelet transform is a strong mathematical signal processing tool with the ability of analyzing nonlinear and nonstationary data. It can produce both time and frequency information with a higher resolution. This provides an opportunity for better potential techniques such as SVMs to blend with wavelet transform to match up with the growing demands.

In this research, it is attempted to investigate the performance of discrete wavelet transform–support vector regression (DWT–SVR) hybrid techniques models for modeling daily pan evaporation and to compare this with the performance of conventional SVR models. The models were developed on the daily weather attributes of evaporation recorded at two stations representing different climatic zones. This provides a better comparison of efficiencies of models to provide accurate results even under varied conditions.

Data collection and preparation

The present work is based on two climatically contrasting stations to determine the efficiency of developed models in providing accurate estimation of evaporation. The meteorological stations selected in this study are Bajpe representing humid climatic condition and Bangalore representing semi-arid condition as per Thornthwaite's classification (Ramachandra et al. 2004), located in Karnataka, India. The daily recorded weather attributes used in the model building are mean air temperature (T), wind speed (W), rainfall (P), mean relative humidity (Rh), sunshine hours (Sh), and pan evaporation (E).

The Bajpe station is located close to Arabian Sea in Dakshina Kannada district. The geographic coordinates of this place are 12°57′N and 74°53′E. The average rainfall is about 3600 mm with an altitude of 103 m above mean sea level. The weather attributes used in the study involve 7-year daily pan evaporation data recorded in the period of 2000–2006. The 5-year (2000–2004) data were used as training dataset and the remaining data, i.e., (2005–2006), were used as testing dataset.

Bangalore station is situated 350 Kms away from the Arabian Sea with the geographic coordinates of approximately 13°39′N and 77°22′E. The altitude is about 900 m above mean sea level with an average annual rainfall of 860 mm. The 7-year (1975–1981) data were used as training dataset and the remaining data, i.e., (1982–1984), were used as testing dataset.

Table 1 shows the statistical analysis of daily recorded data of the Bajpe station (2000–2006) as sample. It also indicates the influencing attributes of pan evaporation. The statistical analysis carried out at the study stations indicates that temperature and sunshine hours showed positive correlation with pan evaporation, while wind speed, rainfall, and relative humidity showed negative correlation. The observation also shows that relative humidity for the variance of rainfall is too high for both the stations. It indicates that the influence of rainfall on pan evaporation is not uniform throughout the periods, but during rainfall period, pan evaporation will be minimum.

Model input selection using gamma test

The selection of proper combinations of inputs is very essential for finding better input–output patterns. Various techniques are available to decide the best pair of inputs to be used in modeling output. One such technique is Genetic algorithm (GA) to identify possible candidate variables for inclusion in the model (Espinoza et al. 2005). IA sensitivity analysis also can be conducted, enabling the ability to determine which variables to be included in the model. With 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 (Kaheil et al. 2008). It is achieved by the estimation of variance of the noise Var (r) computed from the raw data using efficient, scalable algorithms. The novel technique of GT 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

Table 1 Statistical analysis of the weather data taken on weekly basis. (period 2000–2006)

S. No.	Attribute	X_{max}	X_{min}	Standard deviation S_d	Coefficient of variation C_v	Correlation with pan evaporation
1	Mean temperature (°c)	32.75	12.5	1.66	0.06	0.64
2	Wind speed (m/s)	7	4	0.51	0.12	−0.22
3	Rainfall (mm)	291.4	0	22.77	2.44	−0.55
4	Humidity (%)	100	35	11.32	0.14	−0.67
5	Sunshine hours (No's)	11.5	0	3.59	0.59	0.40
6	Pan evaporation (mm)	12.2	0	1.78	0.40	1.00

combinations to achieve a particular target output (Cortes and Vapnik 1995). The GT is designed to efficiently solve overtraining problem, as one of the serious weaknesses associated with almost all nonlinear modeling techniques, by providing 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 (Cortes and Vapnik 1995). The gamma test is a tool for nonlinear modeling and analysis, which can examine the input/output pattern in a numerical dataset. More importantly, the gamma test estimates the part of the output variance which cannot be accounted for by any smooth model based on the inputs, even though this model is unknown. This tool is handy, because of its rapid processing of data, especially in large databases which consisting of thousands of points of datasets, while a single run of the GT takes a few moments (Jones 2004). Genetic algorithm was performed in different dimensions by varying the number of inputs to the model, which clearly presented the response of the data model to some different combinations of input datasets. Input combinations and gamma test results generated for Bajpe station are displayed in Table 2. Similar results were obtained for Bangalore station as well.

Importance of parameter optimization

The parameters used for model building influence the effectiveness of the nonlinear SVR. Among them, the major parameters are the cost constant C, the radius of the insensitive tube ε, and the kernel parameters (Drucker et al. 1999). These parameters mutually influence each other; hence, varying the value of one parameter brings changes in the other linked parameters also.

The parameter C identifies the smoothness/flatness of the approximation function. The smaller value of C leads to a poor approximation, thereby resulting in under-fitting of training data. On the other hand, greater C value overfits the training data and sets its objective to minimize only the empirical risk, making way for more complex learning.

The parameter optimization process is very tedious and requires several times by trial and error (Raghavendra and Deka 2014). Nevertheless, the model will be of benefit for being more efficient and reliable. The grid search results obtained using Bajpe station are displayed in Table 3.

The parameter denotes smoothening the complexity of the approximation function and controls the width of the ε-insensitive zone used for fitting the training data. Ultimately, the number of support vectors is based on parameter ε, and both the complexity and the generalization capability of the approximation function are dependent on its value. It also governs the precision of the approximation function. Smaller values of ε lead to more number of support vectors and result in a complex learning machine. Greater ε values result in more flat estimates of the regression function. Since the present study includes various combinations of inputs, for the epsilon (ε) values of 0.01, the prime parameters C and γ were optimized.

In sequential minimal optimization support vector regression (SMO-SVR), two methods were employed for finding optimal parameter values, a grid search and a cross-validation. Grid search attempts to find the values of each parameter across the specified search range using geometric steps. Generally, grid search needs abundant data

Table 2 Gamma test results for selection of input combinations

S. No.	Input combination	Gamma value	Standard error (SE)	V ratio
1	T	0.116	0.0010	0.467
2	$T + W$	0.113	0.0013	0.455
3	$T + W + P$	0.0741	0.0048	0.296
4	$T + W + P + \mathrm{Rh}$	0.0730	0.0038	0.288
5	$T + W + P + \mathrm{Rh} + \mathrm{Sh}$	0.0724	0.0022	0.289

Table 3 Optimized parameters for combinations of input parameters (Bajpe station)

Input combination	Obtained C parameter	Obtained γ parameter	Number of support vectors (out of training instances 1826)	Correlation coefficient with RBF kernel	
				Training	Testing
T	20	2	1627	0.719	0.650
$T + W$	5	10	1604	0.728	0.682
$T + W + P$	1	1	1578	0.839	0.803
$T + W + P$ + Rh	2	10	1567	0.841	0.832
$T + W + P$ + Rh + Sh	3	1	1552	0.832	0.839

for computations; therefore it is not economical computationally, as the model is evaluated at various points within the grid for each parameter. If cross-validation parameter selection is employed, then V-fold cross-validation is used by the search to estimate the optimal parameters using the error computed from the training data. The derived parameters were later used as inputs for sequential minimal optimization (SMO-Reg) kernel SVR functions for further computations. For this study, a value of $\varepsilon = 0.001$ was set as it fits well with the desired parameters.

Methodology

Support vector regression

Support vector machines are used in classification or regression methods, which have been derived from statistical learning theory. SVMs are good at producing accurate and robust classification results on a sound theoretical basis, even when input data are nonmonotone and nonlinearly separable (Vapnik 1995).

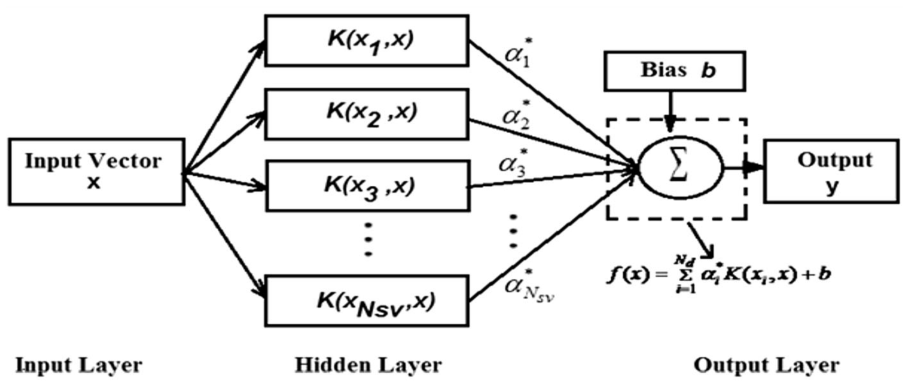

Fig. 1 General structure of support vector machines

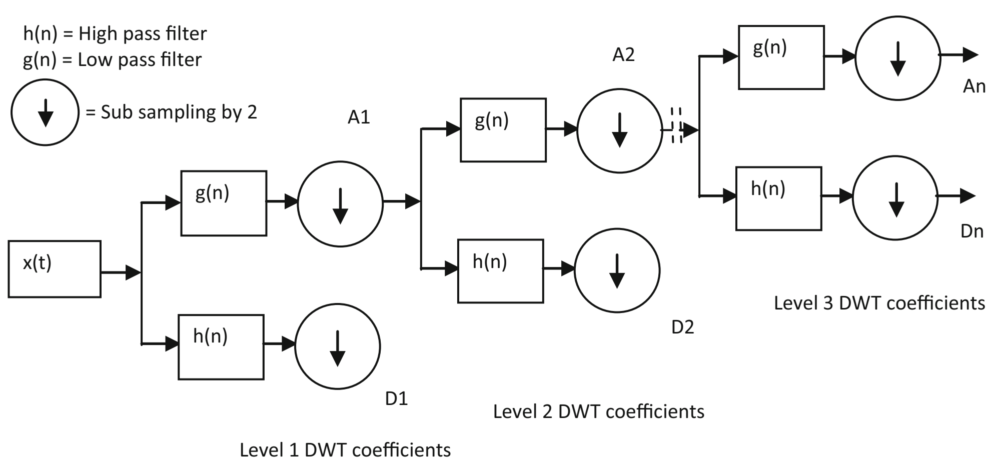

Fig. 2 Wavelet decomposition tree

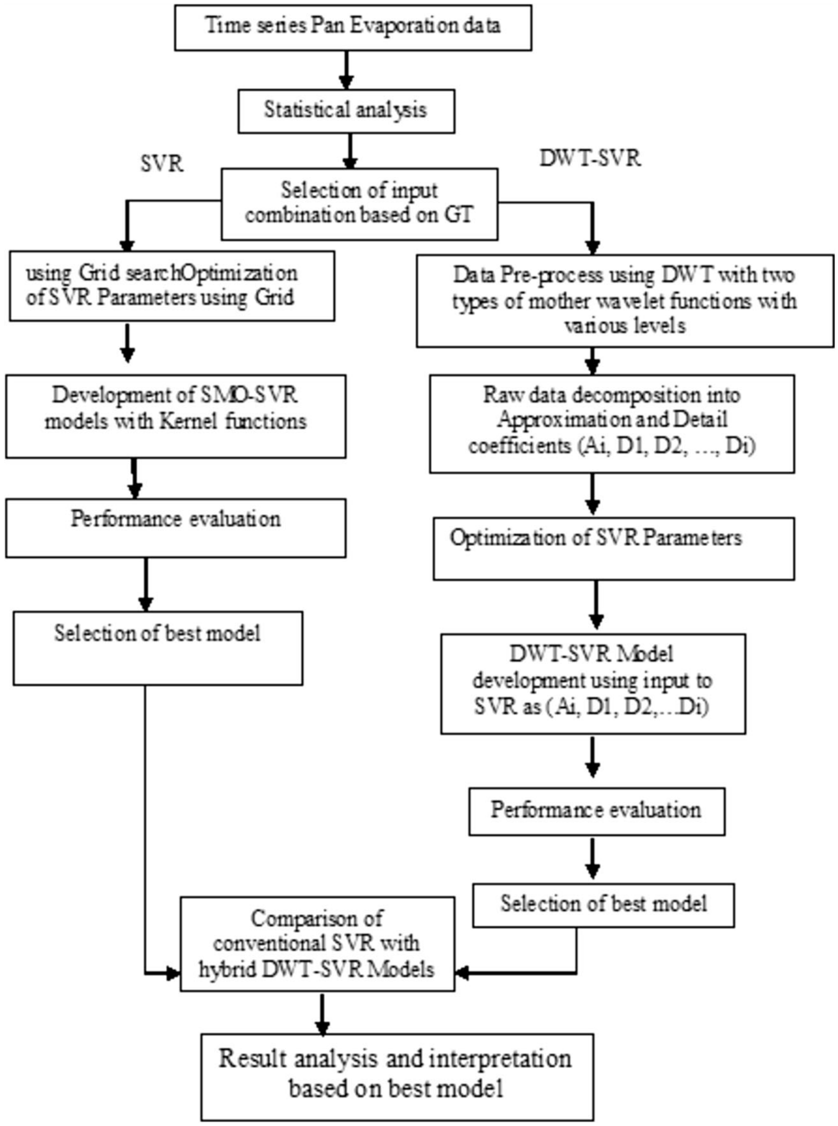

Fig. 3 Flowchart of model development

Table 4 Statistical indices of SMO-SVR models for combined input combinations

Kernel	RMSE (mm)	MAE (mm)	CC (mm)
Polynomial			
Train	1.032	0.759	0.794
Test	0.997	0.770	0.834
RBF			
Train	0.941	0.687	0.832
Test	0.990	0.763	0.839
PuK			
Train	0.906	0.650	0.845
Test	0.981	0.761	0.838

SVM is developed on the basis of statistical learning theory. It is considered to be an approximation implementation of the method of structural risk minimization with a good generalization capability. Advanced algorithm of SVM has been proven to be robust and efficient for classification (Vapnik 1995), regression (Vapnik 1995; Kaheil et al. 2008), forecasting, and prediction. The standard SVM algorithm being used currently was proposed by Cortes and Vapnik (Cortes and Vapnik 1995). The advantage of SVM approach is twofold. The algorithm is simple to understand, and it is so powerful that the predictive accuracy of this approach overpowers many other methods, such as neural networks, nearest neighbor, and

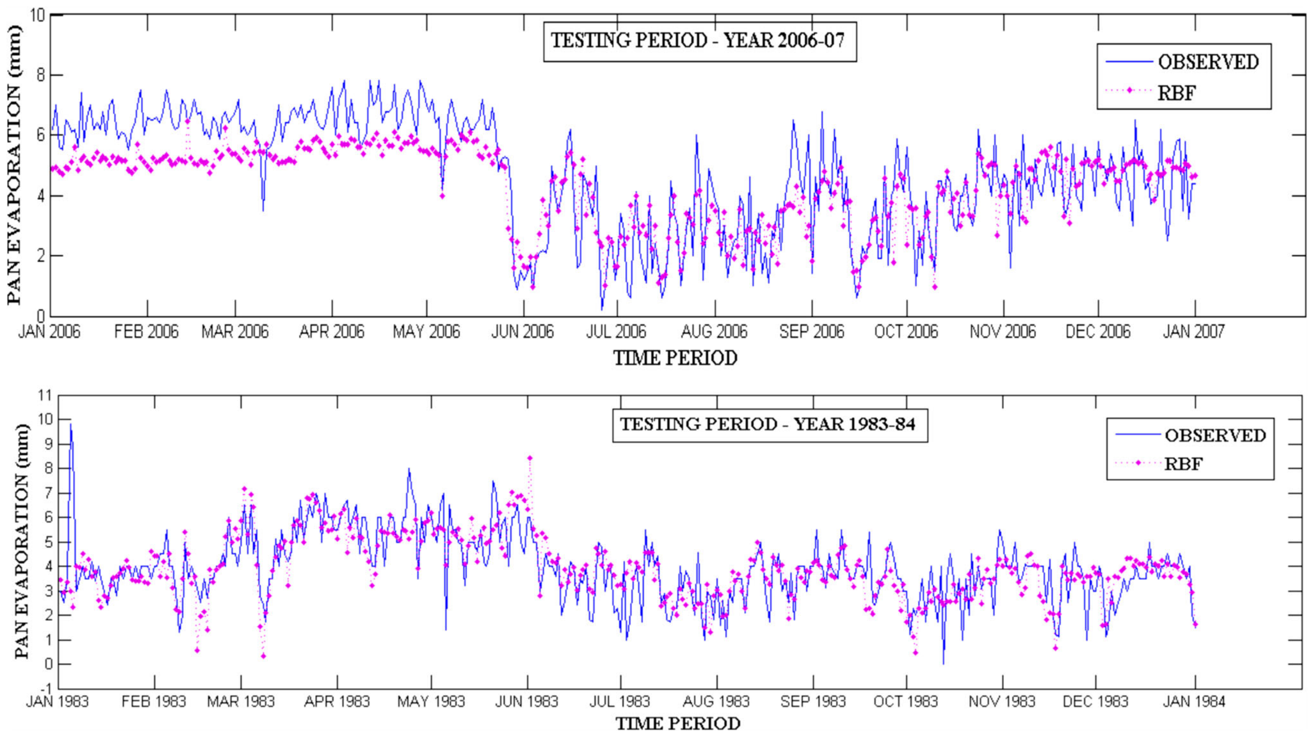

Fig. 4 Daily pan evaporation estimation using SVR–RBF kernel for testing period 2006–2007 of Bajpe station and 1983–1984 of Bangalore station

also decision tree. The modeling techniques like SVMs have shown their potential to reproduce the unknown relationship that exists between a set of input and the output variables of the system. SVM has gained the popularity over other modeling techniques because of their great advantage of minimizing both model complexity and prediction error simultaneously. The general structure of support vector machines is displayed in Fig. 1.

SVM differs from other classification methods significantly. Its intent is to create an optimal separating hyperplane between two classes to minimize the generalization error (i.e., error for the unseen test pattern), thereby maximizing the margin, and thus reducing the overtraining efforts ANN suffers from overtraining, in case training was performed too long or where training examples are rare.

The selection of a proper kernel function plays a vital role in SVM-based classification or regression problems. A number of kernels are discussed in the literature (Vapnik 1995), and therefore the one which gives the best generalization for a given dataset has to be analyzed. In this work, it is attempted to explore and compare the performances of three competent kernels such as polynomial, radial basis function (RBF), and Pearson VII function-based universal kernel (PuK) which are capable of producing desirable results.

Principle of wavelet transformation

As a pre-processing tool, wavelet transforms provide useful decompositions of original time series, so that data that have been pre-processed improve the ability of a forecasting model by capturing information on various resolution levels (Adamowski and Adamowski 2008). In addition, it has also been found that pre-processing data with wavelet transforms can lead to models that better represent the true features of the underlying system by eliminating noise (Adamowski and Adamowski 2008).

DWT operates two sets of functions viewed as high-pass (wavelet function) and low-pass (scaling function) filters. The original time series are passed through high-pass and low-pass filters and down-sampled by two (i.e., throwing away every second data point). Figure 2 shows the wavelet decomposition tree.

Model development methodology

The present work carried out for evaporation estimation makes use of advanced soft computing techniques such as SVR and DWT–SVR. As discussed in the Introduction section, need for accurate approach to provide reliable estimations of evaporation is active field in hydrological modeling. SVM has emerged as one of the robust and

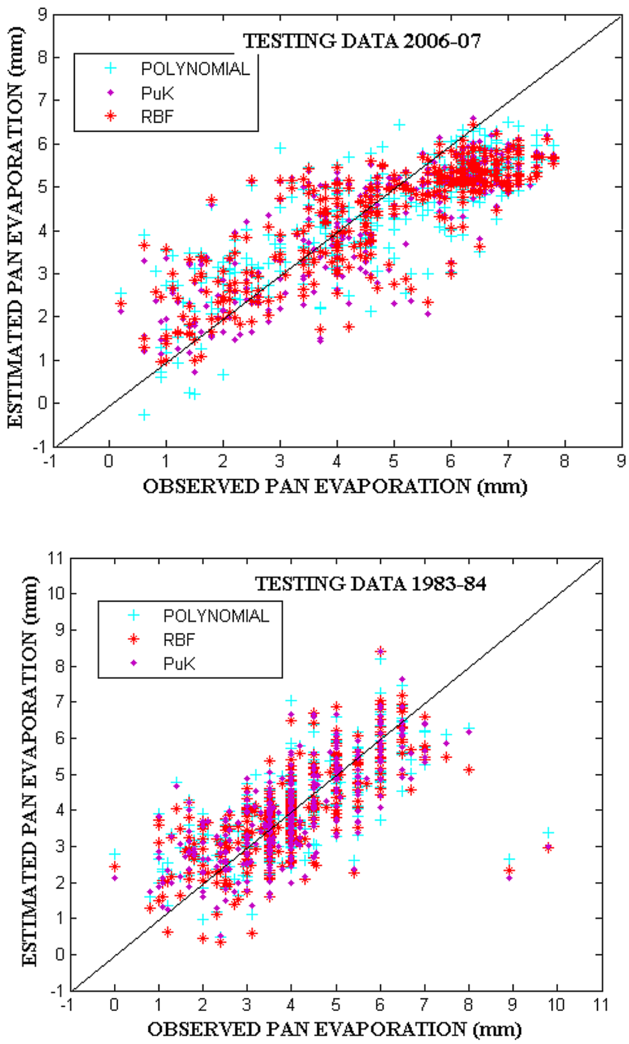

Fig. 5 *Scatter plot* between SVR estimated and observed daily pan evaporation testing data for Bajpe station (2006–2007) and Bangalore station (1983–1984)

Table 5 Statistical indices of wavelet–SVR models for combined input combinations

Kernel	RMSE (mm)	MAE (mm)	CC (mm)
Polynomial			
Train	0.443	0.356	0.926
Test	0.554	0.427	0.950
RBF			
Train	0.448	0.363	0.924
Test	0.525	0.410	0.953
PuK			
Train	0.318	0.225	0.962
Test	0.518	0.400	0.921

accurate techniques, which when coupled with wavelet known to be the best pre-processing technique. It would be interesting to know the model performances between conventional SVR and hybrid DWT–SVR models. The detailed methodology adopted in model development and analysis is shown in Fig. 3.

Evaluation criteria

In the forecasting of evaporation amount, models were set up for the inputs which consisted of different combinations of hydrological variables. In the performance evaluation of SVR and wavelet–SVR models formed using Radial Basis and polynomial kernel, PUK kernel functions such as RMSE, MAE, and CC were selected to gauge their accuracy and to aid comparison as follows:

Root mean square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (X - Y)^2}{N}}$$

Mean absolute error (MAE)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |Y - X|$$

Correlation coefficient (CC)

$$\mathrm{CC} = \frac{\sum_{i=1}^{N} \left(X - \overline{X}\right).\left(Y - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N} \left(X - \overline{X}\right)^2 . \sum_{i=1}^{\mathrm{N}} \left(Y - \overline{Y}\right)^2}}$$

Results and discussion

Once the input and output combination was finalized, support vector machine regression technique was employed initially for model building followed by analysis without pre-processing of data. As discussed in the previous chapters, SMO–SVR methodology was employed with three types of kernels, i.e., polynomial, radial basis function, and PuK kernels which are competitive enough to provide judgmental results. The parameter selection was later conducted using grid search several times by trial and error to arrive at desired SMO–SVR parameters. This would enhance the model efficiency in providing the optimal results. The combination including all the five parameters with evaporation as the output yielded optimum performance. The results obtained with all the three types of kernels highlight that for evaporation to take place, these listed parameters must act unitedly rather than individually.

In the next stage, the wavelet support vector regression (DWT–SVR) models are obtained by hybridizing two techniques, DWT and SVR. The pre-processing of data was carried out using DWT. The DWT–SVR model was then developed using coefficients generated from decomposed original data from DWT and recompiling the data to feed

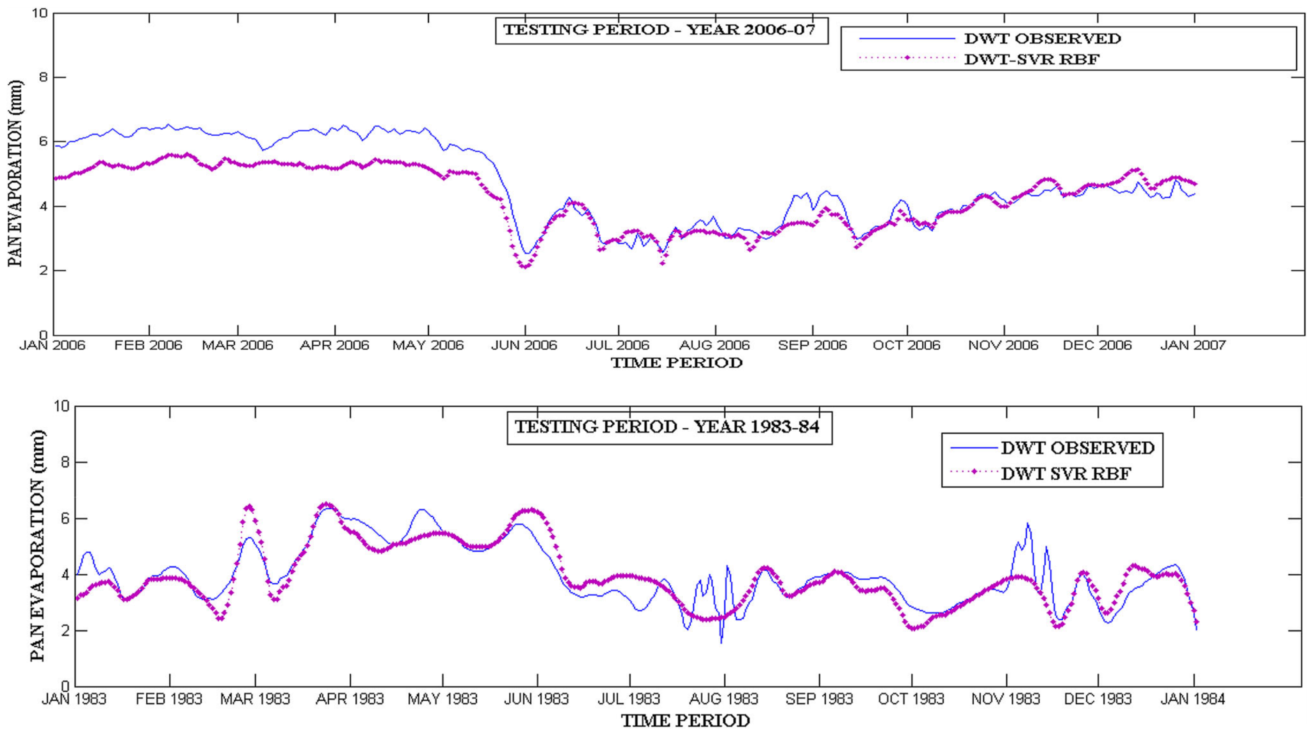

Fig. 6 Daily pan evaporation estimation using DWT–SVR RBF kernel for testing period 2006–2007 of Bajpe station and 1983–1984 of Bangalore station

in SMO–SVR. In this model, signals are split into a detail and an approximation. The approximation obtained from first level is further split into new detail and approximation and this process is repeated.

Discrete wavelet transformation (DWT) is used with Daubechies wavelet order 4 (db4), and Haar mother wavelet functions were employed as pre-processing tools. The selected mother wavelet 'db4' is the simplest wavelet having only three wavelet filter coefficients with exact reconstruction possibilities. To obtain the decomposed wavelet coefficients, various decomposition levels have been tried (L1–L5), but only Daubechies wavelet mother function with level 3 showed better results, i.e., enhanced performance, when fed as an input to SVM on a trial-and-error basis.

Results with support vector regression (SVR)

Although gamma test could identify the best data pattern of input–output combinations, reconfirmation was carried out by testing against support vector regression with the same pattern of various input combinations as done in the gamma test. Since this was just for assurance, SMO–SVR methodology was employed with only one type of kernel, i.e., radial basis kernel which is competitive enough to provide judgmental results. The parameters' selection was also conducted through grid search several times by trial and error to determine desired parameters to enhance the model efficiency to produce better results as discussed in the previous sections. The combination including all the five parameters with evaporation as an output yielded the optimum output. The results of SVR kernel functions are displayed in Table 4.

The choice of suitable kernel will benefit the process of data separation, creating an opportunity to separate data in the feature space, despite being nonseparable in the original input space. A final combination of input–output patterns that were later gauged against SMO having two more competent kernel functions, as listed in Table 4, was adopted to produce support vector regression models.

Comparison of the kernel functions

The accuracy of kernels relies on the selection of the model parameters. The best fitting of models depends on the number of support vectors generated during model building. As the number of influential parameters combines together, model superiority increases. It is also seen in the table that kernel functions played their roles to make the model superior and robust. For the all the five input combination scenarios, RBF kernel function showed slightly better performance in comparison to the other two kernel

Table 6 Comparison of RBF kernel function results with different models and combinations

Input combination	SVR RBF training			SVR RBF testing			SVR–Wavelet RBF training			SVR–Wavelet RBF testing		
	CC (mm)	RMSE (mm)	MAE (mm)	CC (mm)	RMSE (mm)	MAE (mm)	CC (mm)	RMSE (mm)	MAE (mm)	CC (mm)	RMSE (mm)	MAE (mm)
T	0.719	1.187	0.889	0.650	1.324	1.046	0.811	0.680	0.540	0.733	0.897	0.707
$T + W$	0.728	1.165	0.866	0.682	1.282	1.013	0.848	0.617	0.484	0.828	0.817	0.626
$T + W + P$	0.839	0.997	0.737	0.803	1.107	0.864	0.889	0.536	0.419	0.886	0.740	0.545
$T + W + P$ + Rh	0.841	0.934	0.681	0.832	0.879	0.643	0.915	0.470	0.375	0.944	0.541	0.425
$T + W + P$ + Rh + Sh	0.832	0.941	0.687	0.839	0.990	0.763	0.924	0.448	0.363	0.953	0.525	0.410

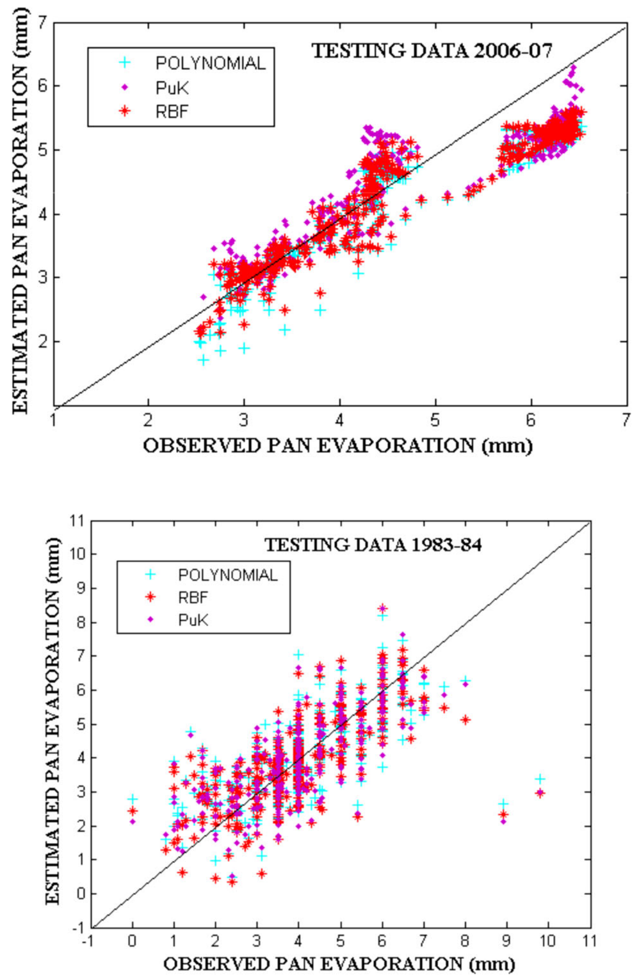

Fig. 7 Scatter plot between DWT and SVR estimated and observed daily pan evaporation testing data for Bajpe station (2006–2007) and Bangalore station (1983–1984)

functions in both training and testing periods. However, PUK showed near similar results to RBF kernel function. Similar pattern of results also observed for Bangalore station confirms that the model accuracy depends on the combination of influential attributes leading evaporation to occur. The drawback experienced in the results of Bangalore station is the sudden decline in the performance from training to testing data phase. The reasons attributed for such a decline may be due to the variations in the ranges of variables and noisy data.

Figure 4 represents the SVR–RBF kernel for testing period of 2006–2007 of Bajpe station and 1983–1984 of Bangalore station, as some samples of the SVR results pattern. In these plots, RBF-estimated values are on par for middle and lower range pan evaporation values which constitute the majority of dataset points. However, RBF underestimated the observed peak values. Considering the station-wise results, RBF showed better performance in testing period of Bangalore station than Bajpe for the

Table 7 Optimal SVR parameters for DWT–SVR models (Bajpe station)

SMO–SVR RBF	Mother wavelet functions							
	DB4 level 1	DB4 level 2	DB4 level 3	DB4 level 4	DB4 level 5	Haar level 3	Haar level 4	Haar level 5
C	7.00	5.50	9.00	14.00	8.75	8.00	10.00	12.00
Gamma	3.75	4.30	5.10	6.45	4.75	3.75	5.25	6.50
Epsilon	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 8 Statistical indices of DWT–SVR RBF models with various mother wavelets (testing period)

Mother wavelets	DWT–SVR RBF kernel (Bajpe station)				DWT–SVR RBF kernel (Bangalore station)			
	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE
Db4 level 1	0.625	0.381	0.867	0.963	0.524	0.373	0.937	0.993
Db4 level 2	0.579	0.385	0.875	0.978	0.419	0.282	0.948	0.995
Db4 level 3	0.525	0.410	0.953	0.991	0.334	0.223	0.959	0.997
Db4 level 4	0.615	0.422	0.854	0.981	0.340	0.512	0.914	0.993
Db4 level 5	0.621	0.357	0.845	0.984	0.472	0.335	0.906	0.994
Haar 3	0.517	0.404	0.934	0.987	0.446	0.332	0.957	0.997
Haar 4	0.691	0.429	0.830	0.989	0.449	0.335	0.951	0.995
Haar 5	0.712	0.466	0.902	0.974	0.560	0.374	0.917	0.992

summer season, i.e., from March to June. This indicates the deviation between the observed and RBF-modeled values in these plots.

However, for rest of the seasons, RBF estimations are uniform for both the stations. Polynomial and Puk kernel results have shown better performance, but ultimately, RBF overshadowed the performance of these two kernels.

Further, this analysis continues through the scatter plots of 1-year testing period of Bajpe and Bangalore stations, as shown in Fig. 5. Referring to the scatter plots in Fig. 5, polynomial kernel-estimated values scatter away from the 45° line and most of them found to be underestimated. RBF kernel matches the observed values with better accuracy than Puk kernel. Most of the RBF kernel-estimated points are closer to 45° line. The similar trend was found in the remaining dataset results.

Results with decomposed wavelet transform support vector regression (DWT–SVR)

As discussed in the previous chapter, parameter optimization plays a vital role in deciding model accuracy. As data points change from observed data points to modeled SVR data points, there is necessity to fine tune the SVR kernel parameters to enhance the accuracy of DWT–SVR models. The fine-tuned parameters computed for the Bajpe dataset are presented in Table 5.

For wavelet analysis, DWT is used with two different mother wavelet functions. The selected mother wavelet 'db4' is the simplest wavelet having only three wavelet filter coefficients with exact reconstruction possibilities. To obtain the decomposed wavelet coefficients, various decomposition levels have been tried with various mother wavelet functions. Among the two mother wavelet functions employed, only Daubechies wavelet mother function with level 3 showed optimum results when fed as inputs to SVM.

The model performances of DWT–SVR RBF results for Bajpe station (2006–2007) and Bangalore station (1983–1984) are plotted with DWT observed as shown in Fig. 6.

Comparison among the SVR and DWT–SVR models

As identified in several research works, hybrid models will definitely make better impact than individual models can do. A comparison between the SVR models and DWT–SVR models can be analyzed with the help of displayed results shown in Tables 4 and 5. The modeled results reveal that RBF kernel function has shown superior performance for both the stations.

Table 9 Statistical indices of DB4 level 3 DWT–SVR models with combinations of kernel functions (testing PERIOD)

Station	DWT–SVR polynomial kernel				DWT–SVR RBF kernel				DWT–SVR PUK kernel			
	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE
Bajpe	0.554	0.427	0.950	0.984	0.525	0.410	0.953	0.991	0.518	0.400	0.921	0.989
Bangalore	0.447	0.321	0.947	0.995	0.334	0.223	0.959	0.997	0.505	0.320	0.951	0.993

In the testing period datasets, DWT–SVR model results with RBF kernel are plotted in Fig. 6. The trend remained the same, RBF could estimate appropriately for the higher pan evaporation values but it underestimated for the low-range pan evaporation values. Considering the station-wise results, RBF showed superior performance for Bajpe station, whereas for Bangalore results, the gap between the observed and estimated values found to be higher for the low-range values of pan evaporation. Further, in the estimated values for Bajpe station, the gap is higher for the initial month's data period, i.e., from January 2006 to May 2006 where there is low-range pan evaporation trend. This suggests that RBF underestimates the observed values for low-range values as shown in Table 6

Figure 7 represents the scatter plots of testing datasets, respectively, for the Bajpe and Bangalore stations. The DB 4–3 mother wavelet functions with three types of SVR kernel-modeled values are plotted in the figures. The testing data show more scattered model values of polynomial and PuK kernels with respect to the 45° line. At the same time, RBF estimates are near accurate and closer to the trend line. Considering station-wise results, it is common that RBF estimations show better accuracy as the values stick closer to the trend line, whereas PuK- and polynomial kernel-estimated values are scattered (Table 7). Considering Bangalore results in scatter plot, PuK kernel values show overestimation of observed values and polynomial kernel values show underestimation of the observed values, Tables 8 and 9. This confirms the superiority of RBF kernel over other kernels.

Conclusions

In this research work, it is attempted to model pan evaporation by employing hybrid model consisting of SVM and wavelet decomposition using daily pan evaporation values. The work includes comparing the performance of hybrid model with conventional model of single SVM. The study was carried out on two climatically contrasting zones which are Bajpe and Bangalore stations. GT test advocates combining all the listed weather attributes used in this study for evaporation to take place. Parameter optimization found usefulness in detecting optimum parameters. Together all these enhance the accuracy of SVR and DWT–SVR models.

Based on the result analysis, the following conclusions are drawn:

- Based on the evaluation of developed models, SVR kernel-based models faced some limitations in modeling pan evaporation data with trends, seasonal patterns, discontinuities, and other complex behaviors.

- In modeling pan evaporation, hybrid models of DWT–SVR are found superior to conventional SVR models.
- DWT–SVR-estimated pan evaporation values were more accurate for the humid station, i.e., Bajpe, than semi-arid station selected in this study, i.e., Bangalore.

Acknowledgement The authors are grateful to Engineer in-charge and staff members of meteorological station near Bajpe for their courtesy to provide access to data for the research work.

References

Adamowski J, Adamowski K (2008) Development of a real-time river flood forecasting transfer function-noise model with a Kalman filter for snowmelt driven floods. J Environ Hydrol 16:1–11

Burt CM, Mutziger AJ, Allen RG, Howell TA (2005) Evaporation research: review and interpretation. J Irrig Drain Eng 131(1):37. doi:10.1061/(ASCE)0733-9437

Cortes C, Vapnik V (1995) Support vector networks. Machine Learn 20:273–297

Deswal S, Pal M (2008) Artificial neural network based modeling of evaporation losses in reservoirs. Int J Math Phys Eng Sci 39(3):177–181

Drucker H, Wu D, Vapnik VN (1999) Support vector machines for spam categorization. IEEE Trans Neural Netw 10:1048–1054

Espinoza F, Minsker B, Goldberg D (2005) Adaptive hybrid genetic algorithm for groundwater remediation design. J Water Resour Plan Manag 131(1):14–24. doi:10.1061/(ASCE)0733-9496(2005)

Finch J, Calver A. (2008) Methods for the quantification of evaporation from lakes. http://nora.nerc.ac.uk/14359

Guimarães Santos CA, Silva GBL (2014) Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models. Hydrol Sci J 59(2):312–324. doi:10.1080/02626667.2013.800944

Jain SK, Nayak PC, Sudheer KP (2008) Models for estimating evapotranspiration using artificial neural networks and their physical interpretation. Hydrol Process 22:2225–2234. doi:10.1002/hyp

Jones A (2004) New tools in non-linear modelling and prediction. CMS 1(2):109–149

Jothiprakash V, Kote AS (2011) Improving the performance of data-driven techniques through data pre-processing for modelling daily reservoir inflow. Hydrol Sci J 56(1):168–186. doi:10.1080/02626667.2010.546358

Kaheil YH, Rosero E, Gill MK, McKee M, Bastidas LA (2008) Downscaling and forecasting of evapotranspiration using a synthetic model of wavelets and support vector machines. IEEE Trans Geosci Remote Sens 46(9):2692–2707

Kim S, Shiri J, Kisi O (2012) Pan evaporation modeling using neural computing approach for different climatic zones. Water Resour Manag 26(11):3231–3249. doi:10.1007/s11269-012-0069-2

Kumar DN, Reddy MJ, Maity R (2007) Regional rainfall forecasting using large scale climate teleconnections and artificial intelligence techniques. J Intell Syst 16(4):307–322. doi:10.1515/JISYS.2007.16.4.307

Nourani V, Sayyah FM (2012) Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes. Adv Eng Softw 47(1):127–146. doi:10.1016/j.advengsoft.2011.12.014

Raghavendra NS, Deka PC (2014) Support vector machine applications in the field of hydrology: a review. Appl Soft Comput 19:372–386. doi:10.1016/j.asoc.2014.02.002

Ramachandra TV, Kamakshi G, Shruthi BV (2004) Bioresource status in Karnataka. Renew Sustain Energy Rev 8:1–47. doi:10.1016/j.rser.2003.09.001

Tabari H, Kisi O, Ezani A, Hosseinzadeh TP (2012) SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. J Hydrol 444–445:78–89. doi:10.1016/j.jhydrol.2012.04.007

Vapnik VN (1995) The nature of statistical learning theory, vol 8. Springer, Verlag. doi:10.1109/TNN.1997.641482

Xu CY, Singh VP (2001) Evaluation and generalization of temperature-based methods for calculating evaporation. Hydrol Process 15:305–319. doi:10.1002/hyp.11