SCIENCE CHINA
Technological Sciences

January 2011 Vol.54 No.1: 163–174
doi: 10.1007/s11431-010-4205-z

Modeling of daily pan evaporation using partial least squares regression

ABUDU Shalamu[1,2*], CUI ChunLiang[1], J. Phillip KING[2], Jimmy MORENO[2] & A. Salim BAWAZIR[2]

[1] *Xinjiang Water Resources Research Institute, Urumqi 830049, China;*
[2] *Civil Engineering Department, New Mexico State University, Las Cruces, New Mexico, 88001, USA*

Received March 12, 2010; accepted October 6, 2010

This study presented the application of partial least squares regression (PLSR) in estimating daily pan evaporation by utilizing the unique feature of PLSR in eliminating collinearity issues in predictor variables. The climate variables and daily pan evaporation data measured at two weather stations located near Elephant Butte Reservoir, New Mexico, USA and a weather station located in Shanshan County, Xinjiang, China were used in the study. The nonlinear relationship between climate variables and daily pan evaporation was successfully modeled using PLSR approach by solving collinearity that exists in the climate variables. The modeling results were compared to artificial neural networks (ANN) models with the same input variables. The results showed that the nonlinear equations developed using PLSR has similar performance with complex ANN approach for the study sites. The modeling process was straightforward and the equations were simpler and more explicit than the ANN black-box models.

modeling, daily pan evaporation, partial least squares regression, artificial neural networks, meteorological data

Citation: Abudu S, Cui C L, King J P, et al. Modeling of daily pan evaporation using partial least squares regression. Sci China Tech Sci, 2011, 54: 163–174, doi: 10.1007/s11431-010-4205-z

1 Introduction

Evaporation is an important component of the hydrologic cycle. Hence, its measurement and estimation are needed for water budgeting, the design of reservoirs, and various other hydrological analyses. However, evaporation estimation on the daily time scale using climate variables is always challenging due to the complex nonlinear nature of the evaporation process. The “physically-based” combination type equations [1,2] generally give the best estimates for open water evaporation. However, their practical application is limited since the calibrations of these models generally require a complete set of meteorological data and a vast amount of computational efforts [3]. To meet the needs of practical application, simpler empirical equations have been developed using limited climate variables. Stephens and Stewart [4] developed a linear model using only temperature and solar radiation as inputs. Priestley and Taylor [5] proposed a radiation-based equation that essentially uses solar radiation and temperature as inputs for daily time scales. Linacre [6] derived a simple formula to predict pan evaporation only from temperature for Australia. Hanson [7] suggested an equation using daily solar radiation and daily mean temperature as inputs to model Class-A daily pan evaporation data for three locations in Idaho. In addition to the above empirical equations, the multiple linear regression (MLR) based equations have also been developed for estimating evaporation from climate variables. Kovoor and Nandagiri [8] attempted the application of multivariate statistical tech-

*Corresponding author (email: shalamu@yahoo.cn) niques such as principal components regression (PCA) and partial least squares regression (PLSR) in pan evaporation modeling to solve collinearity issues that exist in climate predictor variables when developing regression models. The results indicated that these approaches produced similar performance although more parsimonious regression equations can be developed using PLSR. Shirsath and Singh [9] developed daily pan evaporation MLR models using climate data and compared their performance to ANN and climate based models. Li [10] developed evaporation forecasting model using multiple linear regression. However, the linear empirical models and linear regression models that were developed for daily time scales cannot provide accurate estimates due to the highly nonlinear nature of evaporation processes. These limitations of estimating pan evaporation require simple-structured models that can address the inherent nonlinearity of the process using climate variables as inputs.

Recently, there is rapid growing interest in the modeling of pan evaporation using the artificial neural networks (ANN) computing technique due to its capability of identifying complex nonlinear relationships between input and output data sets without the necessity of understanding the nature of the phenomena and without making any underlying assumptions regarding linearity or normality. Previous studies have confirmed the improved performance or comparableness of the ANN models relative to the traditional and multiple linear regression based models in estimating pan evaporation. Sudheer et al. [11] investigated the prediction of Class-A pan evaporation using the ANN technique and found that the ANN could be applied successfully to modeling daily pan evaporation and performed better when compared to the linear Stephens-Stewart model with temperature data as the only input. Bruton et al. [12] developed the ANN models to estimate daily pan evaporation using the measured weather variables as inputs. Their results showed that the ANN model of daily pan evaporation with all available variables as inputs was the most accurate model when compared to multiple linear regression models and the Priestly-Taylor model. Keskin and Terzi [13] compared the ANN models with the Penman model in daily lake evaporation in Turkey and concluded that the ANN approach performed better than the Penman method. Other research studies were reported in literature on the increased performance of applying ANN to estimating pan evaporation when compared to other approaches [14–17].

Based on the results of previous research in pan evaporation estimation, the applied ANN models showed an increase in performance when compared to the conventional multiple linear regression models and other traditional methods. However, the physical interpretation of the ANN architecture and the optimum data required for training and for adaptive learning still demand further explorations [18]. One disadvantage of ANN is that the model relationship cannot be explicitly expressed using a mathematical formulation as in other traditional methods. This makes ANN more complex in its implementation and can be impractical for routine estimations of daily pan evaporation. In this study, the nonlinear relationship between climate variables and daily pan evaporation was analyzed. The nonlinear pan evaporation regression equations were developed utilizing the unique feature of PLSR in eliminating collinearity issues in predictor variables. Variable selection was performed using PLSR from a pool of variables that included original climate variables and transformed variables based on the nonlinear relationship with daily pan evaporation. In addition, square root data transformation was applied to daily pan evaporation. Subsequently, MLR and PLSR were performed on the transformed data. The specific objectives of the study were to: (1) develop nonlinear daily pan estimation regression equations using PLSR and MLR through collinearity analysis; (2) compare the estimation performance of the proposed nonlinear equations, linear equations, and ANN models in three dry climate locations in New Mexico, USA and in Shanshan County, Xinjiang, China; (3) and in contrast to the ANN method, propose a more explicit nonlinear pan evaporation equation in the study sites so that the missing evaporation data could be infilled and/or the evaporation records could be extended for better water management in the region. The methodologies, datasets used, and the results pertaining to the performances of the proposed pan evaporation prediction models are presented and discussed.

2 Methodologies

2.1 Regression methods

Principal components regression and partial least squares regression are two forms of multivariate regression that deal with highly intercorrelated independent variables. In other words, they are extensions of multiple linear regression. The matrix form of multiple linear regression (MLR) models is expressed as follows [19,20]:

$$\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{E}^*, \tag{1}$$

where $\boldsymbol{Y}$ is the matrix of dependent variables ($n \times p$); $\boldsymbol{X}$, the matrix of predictor variables ($n \times m$); $\boldsymbol{E}^*$, the residual matrix ($n \times p$); $\boldsymbol{B}$, the matrix of coefficients ($m \times p$); n, the number of observations; p, the number of dependent variables (p=1 in this study; it is daily pan evaporation); m, the number of independent variables

The least squares solution is

$$\hat{B} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}. \tag{2}$$

In PCR, collinearity that exists in the predictor variables can be eliminated by extracting a group of orthogonal predictors through the application of principal components analysis (PCA) on $\boldsymbol{X}$, and then performing MLR on $\boldsymbol{Y}$ using a subset of the resulting components of $\boldsymbol{X}$. The PCA of the matrix ($\boldsymbol{X}$)

decomposes (***X***) into a score matrix (***T***) times a loading matrix (***P***) and a residual (i.e., error) matrix (***E***) [21,22]. It is possible to let the score matrix, ***T***, represent the predictor matrix, ***X***:

$$\boldsymbol{X}=\boldsymbol{TP}'+\boldsymbol{E},\quad \boldsymbol{T}=\boldsymbol{XP}, \tag{3}$$

where ***T*** is the matrix of ***X*** scores ($n \times a$); ***P′***, matrix of ***X*** loadings ($a \times m$); ***E***, residual matrix of ***X***; a, the number of factors used in the regression.

Then, the MLR formula can be written as follows by replacing ***X*** with ***T***:

$$\boldsymbol{Y}=\boldsymbol{TB}+\boldsymbol{E}^{*}. \tag{4}$$

The solution is

$$\hat{B}=(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y}. \tag{5}$$

In contrast to PCR, PLSR is developed based on both principal components of ***X*** and ***Y***. Specifically, PLSR searches for a set of components (also called latent vectors) that explains as much of the covariance between ***X*** and ***Y*** as possible by performing simultaneous decompositions of both ***X*** and ***Y*** [21].

PLSR is a combination of individual outer relations of ***X*** and ***Y***, and an inner relation of linking both ***X*** and ***Y*** matrices. The outer relation for the ***X*** matrix, which is a similar decomposition as PCA, can be expressed as [19]

$$\boldsymbol{X}=\boldsymbol{TP}'+\boldsymbol{E}=\sum \boldsymbol{t}_h\boldsymbol{p}'_h+\boldsymbol{E}. \tag{6}$$

The outer relation for the ***Y*** matrix can be expressed in similar fashion:

$$\boldsymbol{Y}=\boldsymbol{UQ}'+\boldsymbol{F}^{*}=\sum \boldsymbol{u}_h\boldsymbol{q}'_h+\boldsymbol{F}^{*}, \tag{7}$$

where $\boldsymbol{t}_h$ is the column vector of scores for X block; $\boldsymbol{p}'_h$, row vector of loadings for X block; ***U*=**, matrix of ***Y*** scores;

Q′, matrix of ***Y*** loadings; $\boldsymbol{F}^{*}$ = residual matrix of ***Y***; $\boldsymbol{u}_h$= column vector of scores for ***Y*** block, factor h; $\boldsymbol{q}'_h$ = row vector of loadings for ***Y*** block, factor h

The inner relation of ***X*** and **Y** can be expressed by regression of the ***Y*** block score, u, against the ***X*** block score, ***t***, for every component. The simplest model for this relation is linear [19]:

$$\hat{\boldsymbol{u}}_h=b_h\boldsymbol{t}_h \tag{8}$$

where $b_h=\boldsymbol{u}'_h\boldsymbol{t}_h/\boldsymbol{t}'_h\boldsymbol{t}_h$. This b_h is equivalent to the regression coefficients. This simple model (eq. (8)) is not ideal, because the principal components of ***X*** and ***Y*** are calculated separately and therefore may have a weak relationship to one another. The inner relation can be improved by exchanging scores between ***X*** and ***Y*** blocks in an iterative process. Considering the outer and inner relation of ***X*** and ***Y*** blocks, the following mixed relation can be given for ***Y*** where the error, ***F*** (assumed to be independent and identically normally distributed random variables), is minimized:

$$\boldsymbol{Y}=\boldsymbol{TBQ}'+\boldsymbol{F}. \tag{9}$$

There are several algorithms available for obtaining partial least squares estimators, such as nonlinear iterative partial least squares (NIPALS), singular value decomposition (SVD). More detailed information about PLSR can be found in refs. [23,24,21,8]. Geladi and Kowalski [19] also provided a tutorial on the PLSR method. The Statistical Analysis Software® (SAS) version 9.1 was used for the PLSR and MLR model development in this study.

2.2 Artificial neural networks (ANN)

Artificial neural networks are flexible mathematical structures that are capable of identifying complex non-linear relationships between input and output data sets. The most commonly used type of ANN is a feedforward network termed the multilayer perceptron (MLP). Kim and Valdes [25] described three-layered feedforward neural networks and provided a general framework for representing nonlinear functional mapping between a set of input and output variables. The three-layered ANNs are based on a linear combination of the input variables, which are transformed by a nonlinear activation function. Figure 1 describes a typical artificial neural network structure with only one output neuron in the output layer. The output value of ANN for one output neuron can be expressed by the following equation:

$$\hat{y}_p=f_0\left[\sum_{j=1}^{M}w_j f_{\mathrm{h}}\left(\sum_{i=1}^{N}w_{ji}x_{pi}+w_{j0}\right)+w_0\right], \tag{10}$$

where w_{ji} is a weight in the hidden layer connecting the ith neuron in the input layer and the jth neuron in the hidden layer, w_{j0} is the bias for the jth hidden neuron, f_{h} is the activation function of the hidden neuron, w_j is a weight in the output layer connecting the jth neuron in the hidden layer and the neuron in the output layer, w_0 is the bias for the output neuron, x_{pi} is a value of the ith input for pattern p, and f_0 is the activation function for the output neuron. The weights are different in the hidden and output layer and their values can be changed during the process of network training.

The process of training ANNs is usually accomplished

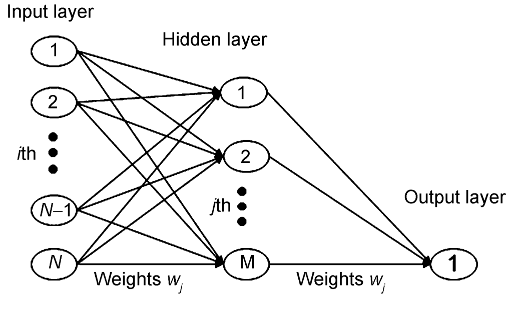

Figure 1 A typical three-layered feedforward neural network structure with one output neuron.

by a backpropagation algorithm, which has been applied successfully to solve difficult and diverse problems. This algorithm is based on the error-correction learning rule. The objective of the backpropagation training process is to adjust the weights of the network to minimize the sum of square errors of the network, which approximates the model outputs to the target values with a selected error goal:

$$E(n)=\frac{1}{2}\sum_{p=1}^{n}\left[y_p(n)-\hat{y}_p(n)\right]^2, \tag{11}$$

where n is the number of observations, $y_p(n)$ is the desired target responses and $\hat{y}_p(n)$ is the actual response of the network at the nth iteration for pattern p. The detailed description of the algorithm is provided in many studies [26,27]. The NeuroSolutions$^{\text{TM}}$ version 5.1 software, a neural network development environment, was used in neural network modeling in this study.

3 Study site and data

Three weather stations, two of them in USA and one in China, were selected in the study for pan evaporation modeling. The two weather stations in USA are located near Elephant Butte Reservoir, New Mexico, USA, and belong to typical arid and semiarid climate. The North Lake Weather Station (called NLWS hereafter) (33°17′50″N, 107°11′38″W) was equipped with an automated Class-A evaporation pan and the South Lake Weather Station (called SLWS hereafter) (33°8′46″N, 107°11′3″W) was equipped with a manually read Class-A evaporation pan. The weather station in China is located in Shanshan County, Turpan Prefecture, Xinjiang Uyghur Autonomous Region, China, and belongs to extreme arid climate. The Shanshan Weather Station (called SSWS hereafter) (42°12′00″N, 90°30′00″W) was equipped with a manually read E601 evaporation pan. There were six years of daily pan evaporation (E, mm/d) data measured from 2002 to 2007 at the NWLS, and three years of data at the SLWS from 2005 to 2007. There were two years of daily pan evaporation (E, mm/d) data measured from 2008 to 2009 at SSWS, but there was no solar radiation data available for 2008 due to lack of instrumentation. Therefore only one year of data (2009) was used in the modeling for SSWS. Missing data were excluded from the data set for modeling purposes. Daily meteorological variables measured at all the three weather stations include maximum air temperature (T_{max}, °C), minimum air temperature (T_{min}, °C), maximum relative humidity (RH_{max}, %), minimum relative humidity (RH_{min}, %), solar radiation (R_{s}, MJ/m^2) and average wind speed (U_{a}, m/s). The SLWS data was divided into two sets: the calibration set (Years 2005–2006, 705 data points) and test set (Year 2007, 351 data points), the total number of data points is 1056. The NLWS data was divided into two sets: the calibration set (Years 2002–2005, 1165 data points) and the test set (Years 2006–2007, 551 data points), the total number of data points is 1716. Due to the shorter length of record, the SSWS data was randomized first and then divided into two sets: the calibration set (136 data points) and the test set (53 data points), and the total number of data points is 188.

4 Results and discussion

4.1 Development of models

To develop nonlinear pan evaporation prediction equations using partial least square regression, two different data transformations were applied to the data. First, a proper data transformation was applied to predictor variables (maximum temperature, minimum temperature, solar radiation, maximum relative humidity, minimum relative humidity and wind speed) based on the analysis of scatter plots between individual climate variable and the daily pan evaporation (Figure 2). The best nonlinear function fitted on the scatter plot was used to transform the climate variables to generate new predictors. Second, a square root transformation was applied to response variable (daily pan evaporation) to get better linear relationship between predictor variables and response variable (as shown in Table 1).

As shown in Figure 2, the relationship between maximum temperature, solar radiation and daily pan evaporation in SLWS could be best approximated by second order function. In contrast, the logarithmic function was the best function to the relationship between maximum relative humidity, minimum relative humidity, wind speed and daily pan evaporation. Similar relationships between climate variables and daily pan evaporation were observed for NLWS. In addition, proper response variable (daily pan evaporation) transformation was investigated through analyzing the scatter plots of the original predictors and daily pan evaporation with various forms of transformation. As a result, the square root transformation of pan evaporation was better linear fit for the relationship between climate variables and evaporation. As shown in Table 1, two different categories of models were developed using the PLSR, MLR and ANN approaches. The first category was linear models that were developed by PLSR and MLR (without transformation) utilizing all the available climate variables (T_{max}, T_{min}, RH_{max}, RH_{min}, R_{s}, U_{a}) as inputs. The second category was nonlinear models that were developed by the MLR, PLSR and ANN modeling methodologies using all climate variables and their derivatives through the transformed variables as inputs.

It is well known that multiple linear regression may produce unreliable prediction results when the predictor variables are highly intercorrelated. This is known as multicollinearity and can be detected by means of variable inflation factors (VIF) in developing regression models. These factors are an indication of how much the variance of the estimated regression coefficients is inflated as compared to the

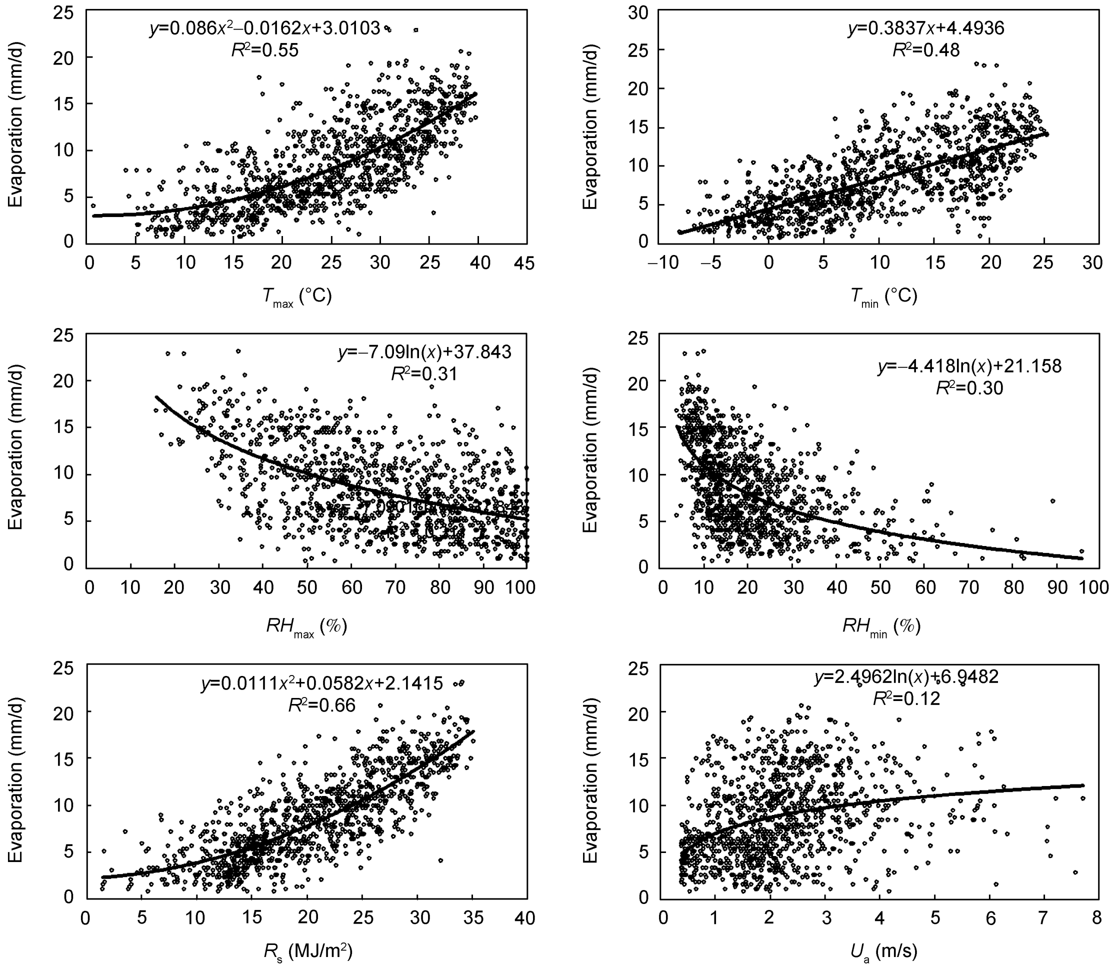

Figure 2 Scatter plot between individual climate variable and daily pan evaporation in South Lake Weather Station, USA.

Table 1 Different modeling approaches in this study using transformed variables

Models		Data transformation	Initial input variables	Response variable	Variable selection method
Linear	MLR-1	no	$T_{\max}$, $T_{\min}$, $RH_{\max}$, $RH_{\min}$, R_s, U_a	E_p	stepwise
	PLSR-1	no	$T_{\max}$, $T_{\min}$, $RH_{\max}$, $RH_{\min}$, R_s, U_a	E_p	variable selection in PLSR
Nonlinear	PLSR-2	predictor variables	$T_{\max}$, $T_{\min}$, $RH_{\max}$, $RH_{\min}$, R_s, U_a, $T^2_{\max}$, $Ln(RH_{\max})$, $Ln(RH_{\min})$, R^2_s, $Ln(U_a)$	E_p	variable selection in PLSR
	MLR-2	predictor variables	Results from variable selection of PLSR-2	E_p	stepwise
	PLSR-3	response variable	$T_{\max}$, $T_{\min}$, $RH_{\max}$, $RH_{\min}$, R_s, U_a	$\sqrt{E_p}$	variable selection in PLSR
	MLR-3	response variable	$T_{\max}$, $T_{\min}$, $RH_{\max}$, $RH_{\min}$, R_s, U_a	$\sqrt{E_p}$	stepwise
	ANN	no	$T_{\max}$, $T_{\min}$, $RH_{\max}$, $RH_{\min}$, R_s, U_a	E_p	no

predictor variables when they are not linearly related. Usually, a VIF value in excess of 10 is often taken as an indication of multicollinearity [20]. In the context of pan evaporation modeling using climate variables (maximum air temperature, minimum air temperature, maximum relative humidity, minimum relative humidity, solar radiation and average wind speed) multicollinearity is likely to exist since the measured variables are inherently related to one another. In this study some of the input variables were highly correlated with one another (Table 2). The highest correlation occurred between maximum and minimum temperature with correlation coefficient of 0.90, 0.89 and 0.71 for SLWS, NLWS and SSWS respectively. These correlations were even higher than the individual correlation coefficients of maximum and minimum temperature with pan evaporation (0.72 for SLWS, 0.78 for NLWS and 0.52 for SSWS).

Table 2 Correlation coefficients between climate variables and daily pan evaporation in the study sites

Location	Variables	T_{max}	T_{min}	RH_{max}	RH_{min}	R_s	U_a	E
SLWS, USA	T_{max}	1.00	0.90	−0.27	−0.32	0.72	−0.01	0.72
	T_{min}		1.00	−0.12	−0.04	0.59	0.11	0.68
	RH_{max}			1.00	0.74	−0.50	−0.22	−0.57
	RH_{min}				1.00	−0.57	−0.17	−0.49
	R_s					1.00	0.21	0.81
	U_a						1.00	0.33
NLWS, USA	T_{max}	1.00	0.89	−0.46	−0.48	0.75	0.12	0.79
	T_{min}		1.00	−0.32	−0.19	0.60	0.22	0.74
	RH_{max}			1.00	0.72	−0.59	−0.29	−0.63
	RH_{min}				1.00	−0.64	−0.19	−0.56
	R_s					1.00	0.26	0.80
	U_a						1.00	0.48
SSWS, China	T_{max}	1.00	0.71	−0.06	−0.32	0.60	−0.39	0.52
	T_{min}		1.00	0.05	0.10	0.23	−0.03	0.57
	RH_{max}			1.00	0.61	−0.05	−0.27	−0.08
	RH_{min}				1.00	−0.47	0.05	−0.25
	R_s					1.00	−0.21	0.55
	U_a						1.00	0.01

To solve the collinearity issue, the stepwise regression was performed in this study to eliminate the variables that were not significant at the 0.05 significance level. The stepwise regression started with all the climate variables (T_{max}, T_{min}, RH_{max}, RH_{min}, R_s, U_a) and used both backward and forward selection. As shown in Table 3, the MLR model was not acceptable in terms of multicollinearity since maximum VIFs of 14.07 and 13.02 were obtained for SLWS and NLWS respectively, which suggested strong multicollinearity among predictors. For SSWS, the maximum VIF is 5.61, indicating that no significant collinearity was observed while developing the MLR equation for SSWS. This may be due to the smaller sample size (n=135) that was used for the analysis. The results of stepwise selection showed that the maximum temperature was not significant for all the weather stations and therefore was dropped from all the final MLR equations developed. Minimum relative humidity was dropped from the MLR equations in SLWS and NLWS, whereas the minimum relative humidity was retained in the MLR equation for SSWS. The four-input (T_{min}, RH_{max}, R_s, U_a) MLR equations did not show any indication of multicollinearity (all VIFs are smaller than 2.16 for SLWS, smaller than 2.15 for NLWS and 1.49 for SSWS). The results are shown in Table 3. The final MLR-1 equations developed for all weather stations and their performance for the calibration period are displayed in Table 4.

Linear partial least squares regression equations (PLSR-1) were developed using variable selection on six original climate variables. There are two major issues when developing PLSR equations. They are the selection of number of components used for the regression equation development and the variable selection based on the contribution of each variable on the final equation. The selection of optimal numbers of the extracted components (factors) is a key issue in developing PLSR, particularly when the models are used for prediction [22]. The PLSR approaches the MLR technique as more components are extracted. However, when there are many predictors, MLR can over fit the observed data. Regression methods with fewer extracted components can provide better predictability of future observations. The rationality of coefficients in the final equation was used to determine the number of components that should be retained in PLSR. The rationality of coefficients included the examination of both sign and magnitude of the coefficients of predictor variables in the final equation [28]. The components of PLSR should be chosen such that the regression coefficients of all variables in the final equation have the same algebraic signs as the correlation coefficients with the dependent variables. Detailed discussions are given in refs. [28–30] regarding rationality of coefficients.

As in MLR, variable selection should be applied when developing the PLSR equations. Some variables can be dropped from the final equations because of their insignificant relationship with the dependent variable or high collinearity with other independent variables. Wold [24] proposed a variable selection technique in PLSR using a Variable Influence on Projection (VIP). The VIP is a weighted sum of squares of the partial least squares weights with the weights calculated from the amount of dependent variable variance of each partial least squares component. The VIP is a statistic that shows the contribution of each independent variable to the model and represents the value of each

Table 3 Variable inflation factors for full models and final models

Variables	Variable inflation factor (VIF)					
	Full model			Final model		
	SLWS	NLWS	SSWS	SLWS	NLWS	SSWS
Intercept	0.00	0.00	0	0.00	0.00	0.00
T_{max}	14.07	13.02	5.61			
T_{min}	11.83	9.91	3.34	1.64	1.56	1.10
RH_{max}	2.41	2.37	2.06	1.44	1.59	
RH_{min}	4.24	3.89	2.89			1.38
R_s	3.09	3.20	2.20	2.16	2.15	1.49
U_a	1.41	1.35	1.71	1.07	1.12	1.05

Table 4 Developed equations for the study sites and their calibration performance

Station	Model	Equation	R^2	*RMSE* (mm/d)	Components used	VIF_{max}
SLWS	MLR-1	$E_p = 4.29 + 0.23T_{min} - 0.07RH_{max} + 0.26R_s + 0.50U_a$	0.808	2.08		2.2 (R_s)
	PLSR-1	$E_p = 4.31 + 0.23T_{min} - 0.07RH_{max} + 0.26R_s + 0.50U_a$	0.808	2.07	3	
	PLSR-2	$E_p = 5.35 + 0.002T_{max}^2 + 0.15T_{min} - 0.06RH_{max} + 0.006R_s^2 + 0.61U_a$	0.816	2.03	4	
	MLR-2	$E_p = 5.69 + 0.21T_{min} - 0.05RH_{max} + 0.007R_s^2 + 0.46U_a$	0.797	2.04		2.1 (R_s^2)
	PLSR-3	$\sqrt{E_p} = 1.89 + 0.007T_{max} + 0.04T_{min} - 0.01RH_{max} - 0.002RH_{min} + 0.04R_s + 0.10U_a$	0.816	2.04	5	
	MLR-3	$\sqrt{E_p} = 2.09 + 0.04T_{min} - 0.01RH_{max} - 0.005RH_{min} + 0.04R_s + 0.09U_a$	0.816	2.03		3.0 (RH_{min})
	ANN	6-7-1, inputs: T_{max}, T_{min}, RH_{max}, RH_{min}, R_s, U_a	0.817	2.02		
NLWS	MLR-1	$E_p = 1.80 + 0.21T_{min} - 0.05RH_{max} + 0.25R_s + 0.87U_a$	0.841	1.77		2.2 (R_s)
	PLSR-1	$E_p = 1.86 + 0.21T_{min} - 0.05RH_{max} + 0.25R_s + 0.88U_a$	0.841	1.77	3	
	PLSR-2	$E_p = 2.31 + 0.003T_{max}^2 + 0.08T_{min} - 0.04RH_{max} + 0.006R_s^2 + 1.02U_a$	0.863	1.64	4	
	MLR-2	$E_p = 1.29 + 0.005T_{min}^2 - 0.03RH_{max} + 0.005R_s^2 + 1.29U_a$	0.843	1.70		2.6 ($R^2{}_s$)
	PLSR-3	$\sqrt{E_p} = 1.31 + 0.02T_{max} + 0.03T_{min} - 0.006RH_{max} - 0.007RH_{min} + 0.03R_s + 0.19U_a$	0.874	1.58	5	
	MLR-3	$\sqrt{E_p} = 1.67 + 0.04T_{min} - 0.005RH_{max} - 0.012RH_{min} + 0.04R_s + 0.17U_a$	0.871	1.60		2.9 (R_s)
	ANN	6-7-1, inputs: T_{max}, T_{min}, RH_{max}, RH_{min}, R_s, U_a	0.880	1.54		
SSWS	MLR-1	$E_p = 1.75 + 0.24T_{min} - 0.07RH_{min} + 0.13R_s + 0.86U_a$	0.523	1.49		1.5 (R_s)
	PLSR-1	$E_p = 1.75 + 0.24T_{min} - 0.07RH_{min} + 0.13R_s + 0.86U_a$	0.523	1.49	3	
	PLSR-2	$E_p = 4.21 + 0.008T_{min}^2 - 0.06RH_{min} + 0.004R_s^2 + 0.87U_a$	0.533	1.47	4	
	MLR-2	$E_p = 4.21 + 0.008T_{min}^2 - 0.06RH_{min} + 0.004R_s{}^2 + 0.87U_a$	0.533	1.47		1.4(R_s^2)
	PLSR-3	$\sqrt{E_p} = 1.63 + 0.04T_{min} - 0.01RH_{min} + 0.02R_s + 0.15U_a$	0.524	1.49	3	
	MLR-3	$\sqrt{E_p} = 1.63 + 0.04T_{min} - 0.01RH_{min} + 0.02R_s + 0.15U_a$	0.524	1.49		1.5(R_s)
	ANN	4-5-1, inputs: T_{min}, RH_{min}, R_s, U_a	0.539	1.46		

predictor in fitting the PLSR model for both predictors and responses. For a selected number of components and input variables, the VIP values of each predictor variable can be calculated and used to examine the strength of the relationship, irregularities, and the contribution of the independent variables in the model. To determine which predictor should be eliminated from the model, the standardized regression coefficient and the VIP of each predictor should be analyzed. An independent variable may have a small coefficient value, but may have a large VIP, which implies that this independent variable is important and contributes significantly to the prediction and therefore, has to be kept in the model. Wold [24] suggested that a VIP value of less than 0.8 is small. If a predictor has a relatively small coefficient (in absolute value) and a small value of VIP (less than 0.8), then it is a prime candidate for deletion. In this study, the variable selection was carried out by analyzing the VIP, standardized regression coefficient of each predictor and the reduction rate of the coefficient of determination (R^2) for the calibration period when dropping unimportant variables from the final equation. In addition the root mean squared error (RMSE) can be used to evaluate the calibration performance of the model.

Using the above selection procedures in PLSR, a linear prediction equation (PLSR-1) and two nonlinear prediction equations (PLSR-2 and PLSR-3) were developed for the weather stations. Based on the variable selection results of PLSR-2 and PLSR-3, two MLR nonlinear equations (MLR-2 and MLR-3) were developed using stepwise variable selection and considering the variable inflation factor (VIF) values in MLR. The final variables entering into the equations have small VIF values (as shown in Table 4). Since VIFs were much smaller than 10, it was assumed that collinearity was not a problem in the prediction of pan evaporation using these equations developed by MLR. As shown in Table 4, there was slight improvement in the calibration performances of the nonlinear approaches through transformation of predictors and response variables. In SLWS, both transformation approaches yielded similar calibration performance (R^2=0.816 and *RMSE*=2.03 mm/d for PLSR-2 and R^2=0.816 and *RMSE*=2.04 mm/d for PLSR-3). The PLSR approach performed slightly better than the MLR approach when using predictor transformation. However, the calibration performance remained the same for the PLSR and MLR methods when using response variable transformation. In NLWS, similar results were observed as in SLWS. The calibration performance of both nonlinear equations improved significantly compared to the linear models. For example, the R^2 increased from 0.841 to 0.874 and *RMSE* decreased from 1.77 mm/d to 1.58 mm/d from PLSR-1 to PLSR-3. When the two data transformations were compared the response variable transformation yielded slightly better calibration results (R^2=0.863 and *RMSE*=1.64 mm/d for PLSR-2 and R^2=0.874 and *RMSE*= 1.58 mm/d for PLSR-3). Compared to NLWS, the calibration performance of nonlinear models in SSWS was not significantly different from the linear models. However, the nonlinear equations developed using predictor variable transformation showed better calibration performance compared to other models. These results may be due to the smaller sample size that was used for the model development.

To evaluate the performance of nonlinear models, the ANN modeling approach was selected for comparison due to its recent successful applications in nonlinear modeling of evaporation processes [13–17]. To compare the results of nonlinear equations with neural network models, the ANN models were developed using all the climate variables for NLWS and SLWS (Table 1). The ANN model input variables for SSWS were kept the same as the developed PLSR and MLR nonlinear models. For the two weather stations in the USA, the calibration data set was divided into a training set (605 samples for SLWS and 1000 samples for NLWS) and cross validation set (100 samples for SLWS and 165 for NLWS). For SSWS, the calibration data set consisted of 135 samples. The training process used the hyperbolic tangent function as the activation function in the hidden layer, linear function in output layer. The NeuroSolutions software automatically scaled and shifted the input data to match the range of the first hidden layer's transfer function. For example, if the hidden layer's activation function is a hyperbolic tangent function, the input data will be scaled and shifted to lie between −1 and 1 [31]. The momentum learning rule was utilized to calculate the weight update. For all the ANN models, the step size and momentum were selected as 1.0 and 0.7 respectively. For the two weather stations in the USA, the training termination criteria employed cross validation techniques that will stop the training when the cross validation error begins to increase. The maximum number of training epochs was set to be 5000. The training was terminated when there was no further improvement in cross validation after 500 epochs. For SSWS, the training was carried out by limiting the number of epochs to 200, mainly because of the smaller sample size (n=135). One hidden layer was used in all the ANN types in this study. Previous experimental results [26] indicate that one hidden layer is enough for most hydrological problems. The nodes in the hidden layer were decided by trial and error. Different numbers of nodes, from 3 to 12, were used to train different networks.

Based on the errors in the training and cross validation set, the best ANN network was selected. As shown in Table 4, the number of nodes in the hidden layer was set to 7 with 6 inputs for both SLWS and NLWS. The number of nodes in the hidden layer was set to 5 with 4 inputs for SSWS. Once the optimum network architectures for different models were selected, each network was trained six to ten times to see if the network performance was stable for different initial values. For each training process, the NeuroSolutions will assign new initial values automatically [31]. The calibration performance of the ANN models is shown in Table 4. As can be seen, the ANN model performed similar to nonlinear equations developed by PLSR for SLWS and SSWS. A slight improvement was seen in the calibration performance for the ANN model in NLWS with the highest R^2 equal to 0.880 and the lowest RMSE equal to 1.54 mm/d, whereas, the best nonlinear equation for NLWS (PLSR-3) has an R^2 of 0.874 and RMSE of 1.58 mm/d. These results indicated that difference between the calibration performance of the nonlinear equations developed with the PLSR and ANN models was marginal for the selected weather stations.

4.2 Application of models for new data

The robustness of the presented models was tested for prediction using new data at the weather stations. The testing data was comprised of one year of data for SLWS (2007, 351 samples), two years of data for NLWS (2006–2007, 551 samples) and randomized 53 samples for SSWS. Model prediction performance was evaluated using the coefficient of determination (R^2) and the root mean squared error (RMSE). The model performances are shown in Table 5. It is evident that the nonlinear models yielded better performance than the linear models overall for the weather stations. The best performed nonlinear model in SLWS was PLSR-3 and resulted in an R^2 of 0.76 and an RMSE of 2.02 mm/d, which is a slightly better performance compared to the linear PLSR model (PLSR-1, R^2=0.74 and *RMSE*=2.15 mm/d). Similar results can be observed for SSWS, where the best performed one was PLSR-2 (R^2=0.57; *RMSE*=1.36 mm/d) and was slightly better than MLR1 linear model (R^2=0.55; *RMSE*=1.39 mm/d). A significant improvement was observed from linear models to nonlinear models at NLWS. The R^2 increased from 0.73 to 0.81 and *RMSE* decreased from 2.21 mm/d to 1.80 mm/d. This supports the assertion that pan evaporation is a nonlinear process and that linear models only partially account for some of the variance that nonlinear models explain. No significant differences in performances were detected when the nonlinear equations developed by PLSR and MLR were compared to ANN models. In fact the nonlinear equation developed by PLSR either performed slightly better than or equivalent to the other models at the weather stations. However, the PLSR equations are explicit and can be implemented readily in contrast to the ANN models which use a complex machine learning technique that cannot be translated into explicit mathematical formulation.

When the two nonlinear equations developed by PLSR and MLR were compared, it was observed that carefully developed nonlinear MLR equations can avoid the multicollinearity problem using stepwise variable selection. Their performance is similar (slightly lower), although the maximum temperature was dropped out from the MLR equation due to multicollinearity. In NLWS, the PLSR-3 model provided slightly better performance than MLR-3. Coefficients of determination were 0.80 and 0.81 and RMSEs were 1.85 mm/d and 1.80 mm/d for MLR-3 and PLSR-3 respectively. The square root transformation of daily pan evaporation gave better performance than the transformation of individual climate variables in SLWS and NLWS for the MLR nonlinear equations. In both weather stations, the square root transformation of daily pan evaporation, followed by the development of regression equations using PLSR produced the same results as the complex ANN models and data-intensive combination type empirical equations. The predictor variable transformation approach (PLSR-2) yielded better results when compared to the other models in SSWS.

The scatter plots of the observed and estimated daily pan evaporation of linear models (MLR-1 and PLSR-1) and nonlinear models (PLSR-3/PLSR-2 and ANN) are given in Figures 3–5 for all the weather stations. The comparison of the MLR-1 and PLSR-1 models indicated that the PLSR-1 model did not show a distinct improvement when compared to the MLR-1 model. This may be due to the fact of both being linear models, the only difference is the structure of the models. It can be seen that the nonlinear models showed better performance when compared to linear models. When ANN and PLSR-3/PLSR-2 were compared, no significant difference could be observed for all the weather stations. The performance of the ANNs and nonlinear PLSR models was quite similar for all the weather stations. However, the nonlinear PLSR models have a much simpler form and can be expressed in an explicit estimation equation form in contrast to the ANN model. The weakness of the ANN models is that they are essentially black box models that cannot be readily replicated. Hence, from the point of parsimony and practical use, the nonlinear PLSR models would be preferable for estimating daily pan evaporation at the study sites.

5 Conclusions

This study presented the application of partial least squares regression (PLSR) in estimating daily pan evaporation. The

Table 5 Comparison of performance of the models for testing data set for all weather stations

Models		SLWS (*n*=351)		NLWS (*n*=551)		SSWS (*n*=53)	
		R^2	*RMSE* (mm/d)	R^2	*RMSE* (mm/d)	R^2	*RMSE* (mm/d)
Linear	MLR-1	0.74	2.15	0.73	2.21	0.55	1.39
	PLSR-1	0.74	2.15	0.73	2.20	0.55	1.39
Nonlinear	PLSR-2	0.76	2.06	0.78	1.98	**0.57**	**1.36**
	MLR -2	0.76	2.06	0.79	1.95	0.57	1.36
	PLSR-3	0.76	**2.02**	0.81	**1.80**	0.55	1.39
	MLR-3	0.76	2.03	0.80	1.85	0.55	1.39
	ANN	**0.77**	2.07	**0.82**	**1.80**	0.56	1.37

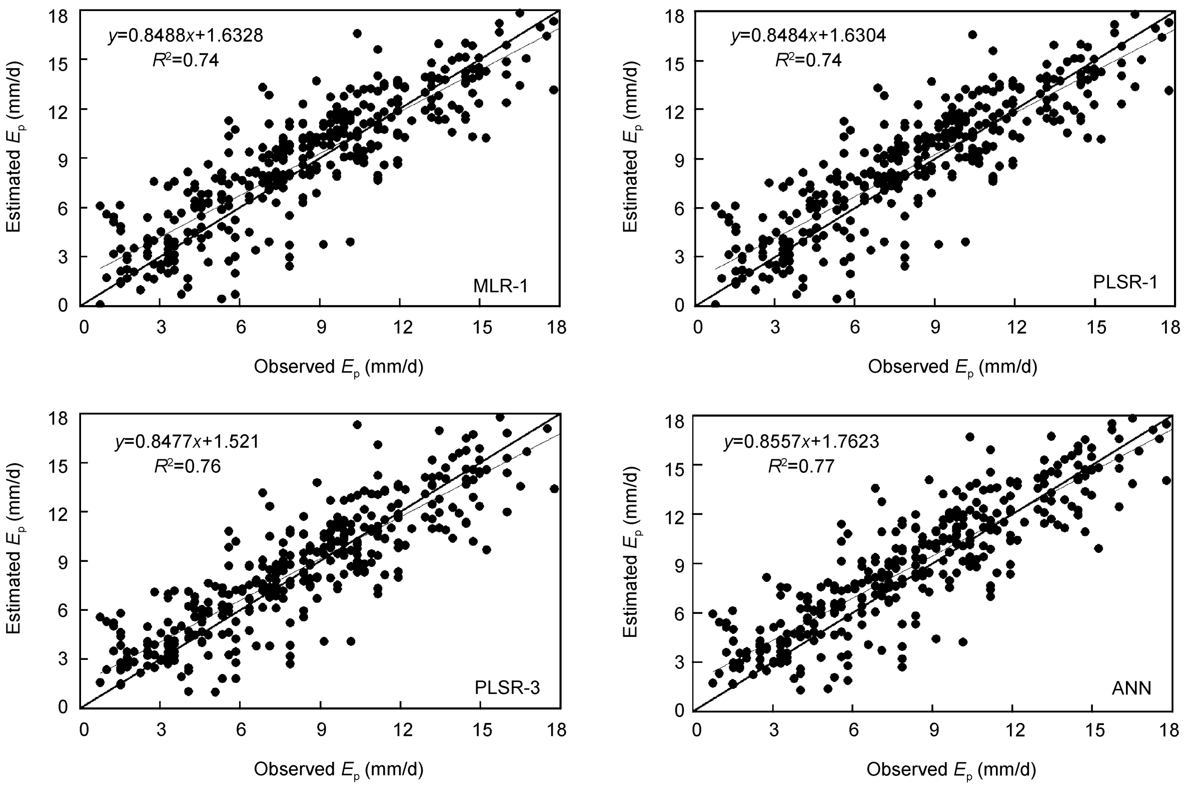

Figure 3 Scatter plots between observed daily pan evaporation using different models for South Lake Weather Station, USA.

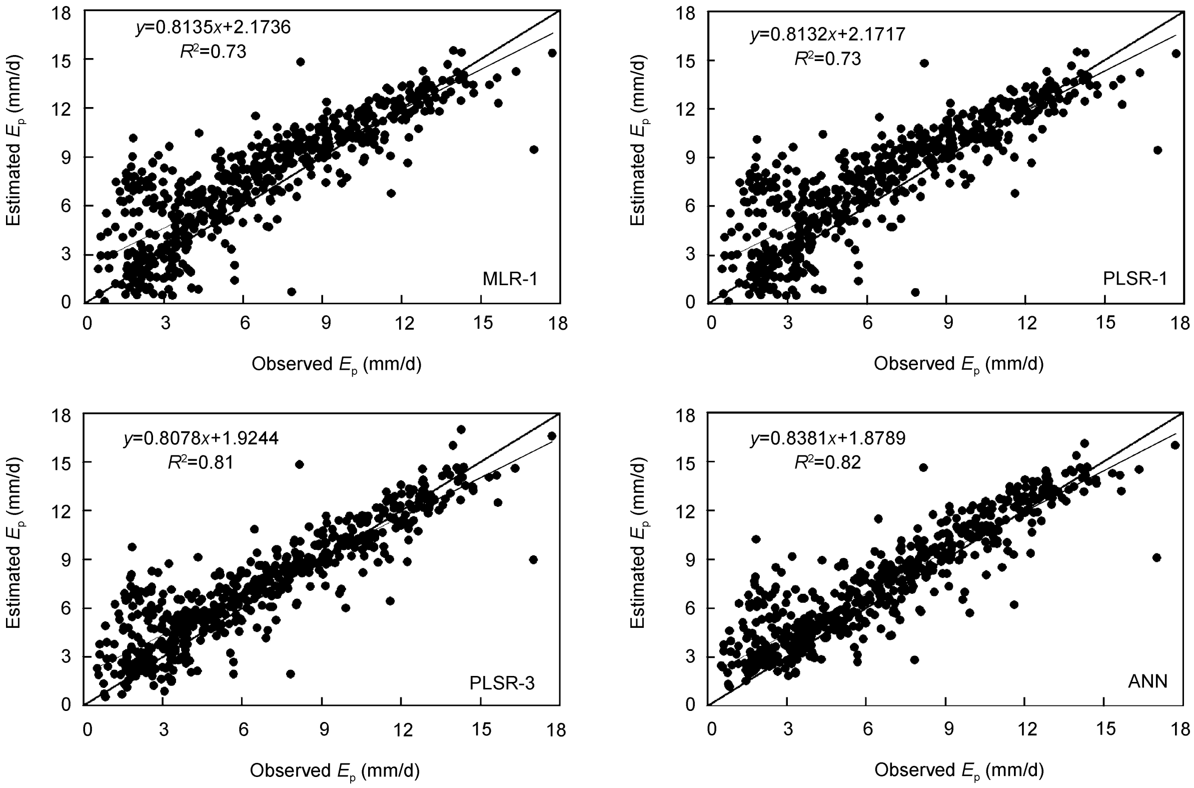

Figure 4 Scatter plots between observed daily pan evaporation using different models for North Lake Weather Station, USA.

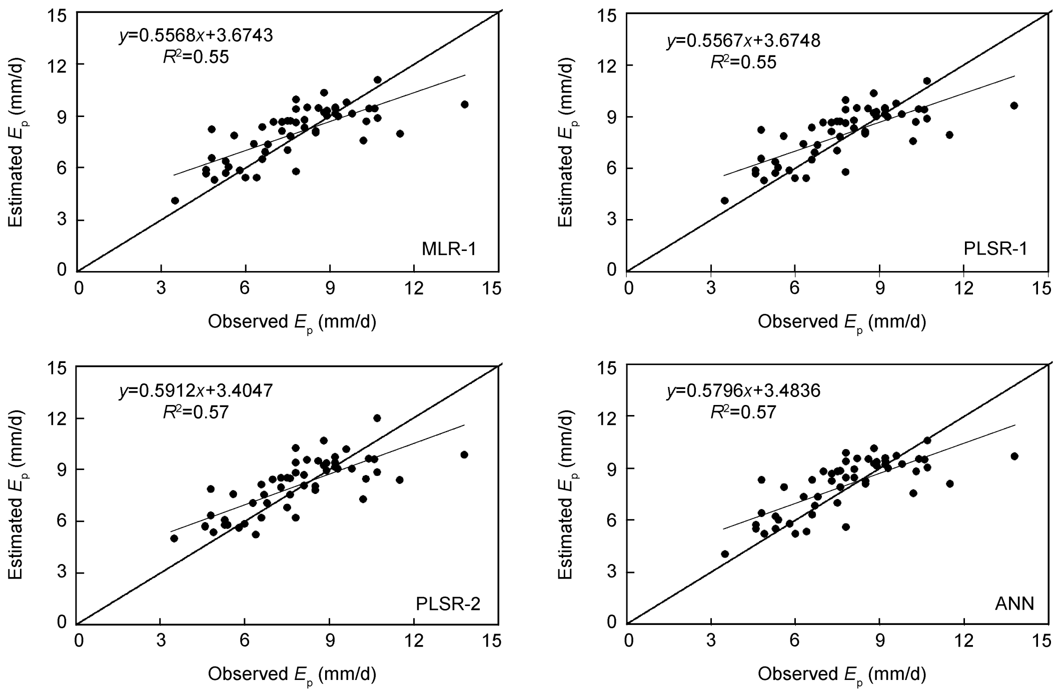

Figure 5 Scatter plots between observed daily pan evaporation using different models for Shanshan Weather Station, China.

nonlinear nature of daily evaporation was modeled by using data transformation and applying the unique feature of PLSR that can deal with highly intercorrelated predictor variables. The modeling results were compared to artificial neural networks (ANN) models that used the same input variables. The results showed that the nonlinear equations developed using PLSR had similar performance with the complex ANN approach for the study sites. The modeling process produced equations that were much simpler and could be expressed explicitly which were not possible with neural networks models. The proper data transformation that reflects the nonlinear relationship between climate variables and daily pan evaporation has been proved to be an effective method in daily pan evaporation modeling in this study. However, the empirical equations in this study were developed for only three weather stations that located in the typical arid and semi-arid climate regions in the United States and China. For other weather stations, new PLSR equations should be calibrated based on the specific conditions of the new study sites. Further studies would be needed to apply the methodology to other study sites, especially in different climate conditions.

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 51069017, 41071026). The authors would like to thank Zhang Jianghui, Bai Yungang, Zhang Shengjiang from Xinjiang Water Resources Research Institute, China for providing data and thank the United States Bureau of Reclamation, especially Mr. Brent Tanzy for providing weather and pan data at Elephant Butte Reservoir. The authors also express their sincere appreciation of the reviewers' valuable suggestions and comments in improving the quality of this paper.

1 Penman H L. Natural evaporation from open water, bare soil and grass. Proceedings, Royal Society of London, 1948, 193: 120-145

2 Kohler M A, Nordenson T J, Fox W E. Evaporation from Pans on Lakes. Weather Bureau Research Paper 38. Washington, DC: US Department of Commerce, 1955

3 Tan S B K, Shuy E B, Chua L H C. Modelling hourly and daily open-water evaporation rates in areas with an equatorial climate. Hydrol Process, 2007, 21(4): 486-499

4 Stephens J C, Stewart E H. A Comparison of Procedures for Computing Evaporation and Evapotranspiration. Publication 62, International Association of Scientific Hydrology. International Union of Geodynamics and Geophysics, Berkeley, CA. 1963. 123–133

5 Priestley C H B, Taylor R J. On the assessment of the surface heat flux and evaporation using large-scale parameters. Monthly Weather Rev, 1972, 100: 81–92

6 Linacre E T. A simple formula for estimating evaporation rates in various climates, using temperature data alone. Agric Met, 1977, 18(6): 409–424

7 Hanson C L. Prediction of Class A pan evaporation in southwest Idaho. J Irrig Drain Eng, 1989, ASCE 115(2): 166–171

8 Kovoor G M, Nandagiri L. Developing regression models for predicting pan evaporation from climatic data—A comparison of multiple least-squares, principal components, and partial least-squares approaches. J Irrigd Drain Eng, 2007, 133(5): 444–454

9 Shirsath P B, Singh A K. A comparative study of daily pan evaporation estimation using ANN, regression and climate based models. Water Resour Manage, 2010, 24: 1571–1581

10 Li Y. Study on water surface evaporation forecasting based on single correlation coefficient. Groundwater, 2010, 32(2): 113–114

11 Sudheer K P, Gosain A K, Rangan D M, et al. Modelling evaporation using an artificial neural network algorithm. Hydrol Process, 2002, 16: 3189–3202

12 Bruton J M, McClendon R W, Hoogenboom G. Estimating daily pan evaporation with artificial neural networks. Trans ASAE, 2000, 43(2): 491–496

13 Keskin M E, Terzi O. Artificial neural network models of daily pan evaporation. J. Hydrol Eng, 2006, 11(1): 65–70

14 Jin L, Shen S H, Luo Y, et al. Study on ANN calculation of water surface evaporation. J Nanjing Institute of Meteorology, 1996, 19(3): 342–347

15 Keskin M E, Terzi O, Taylan D. Fuzzy logic model approaches to daily pan evaporation estimation in western Turkey. Hydrol Sci J, 2004, 49(6): 1001–1010

16 Terzi O, Keskin M E, Taylan E D. Estimating evaporation using ANFIS. J Irrig Drain Eng, 2006, 132(5): 503–507

17 Eslamian S, Gohari S A, Biabanaki M, et al. Estimation of monthly pan evaporation using artificial neural networks and support vector machines. J Appl Sci, 2008, 8(19): 3497–3502

18 American Society of Civil Engineers (ASCE). Artificial neural network in hydrology: hydrologic applications. J Hydrol Eng, 2000, 5: 124–137

19 Geladi P, Kowalski B. Partial least squares regression: A tutorial. Analytica Chimica Acta, 1986, 185: 1–17

20 Neter J, Wasserman W, Kutner M H. Applied Linear Regression Models. 2nd ed. Homewood, Illinois: Richard D. Irwin, Inc., 1989

21 Abdi H. Partial least squares regression. In: Lewis-Beck M, Bryman A, Futing T, eds. The Sage Encyclopedia of Social Sciences Research Methods. Thousand Oaks, CA: Sage, 2003. 1–7

22 Tootle G A, Singh A K, Piechota T C, et al. Long lead-time forecasting of U.S. streamflow using partial least squares regression. J Hydrol Eng, 2007, 12(5): 442–451

23 Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnaiah P R, ed. Multivariate Analysis. New York: Academic Press, 1966. 391–420

24 Wold S. PLS for multivariate linear modeling. In: van de Waterbeemd H ed. QSAR: Chemometric Methods in Molecular Design, Methods and Principles in Medicinal Chemistry. Weinheim: Verlag-Chemie, 1994. 195–218

25 Kim T, Valdes J B. Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. J Hydrol Eng, 2003, 8(6): 319–328.

26 Coulibaly P, Anctil F, Bobée B. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J Hydrol, 2000, 3(4): 244–257

27 Dibike Y B, Solomatine D P. River flow forecasting using artificial neural networks. J Phys Chem Earth, Part B: Hydrol, Oceans Atmos, 2001, 26(1): 1–8

28 McCuen R H, Rawls W J, Whaley B L. Comparative evaluation of statistical methods for water supply forecasting. Water Resour Bull, 1979, 15(4): 935–947

29 Garen D C. Improved techniques in regression-based streamflow volume forecasting. J Water Res PL-ASCE, 1992, 118(6): 654–670

30 McCuen R H. Statistical Methods for Engineers. Englewood Cliffs, N. J.: Prentice Hall, 1985

31 NeuroDimension, Inc. NeuroSolutions Getting Started Manual Version 5, 2009