Developing Regression Models for Predicting Pan Evaporation from Climatic Data—A Comparison of Multiple Least-Squares, Principal Components, and Partial Least-Squares Approaches

Gicy M. Kovoor[1] and Lakshman Nandagiri[2]

Abstract: Regression models for predicting daily pan evaporation depths from climatic data were developed using three multivariate approaches: multiple least-squares regression (MLR), principal components regression (PCR), and partial least-squares (PLS) regression. The objective was to compare the prediction accuracies of regression models developed by these three approaches using historical climatic datasets of four Indian sites that are located in distinctly different climatic regimes. In all cases (three approaches applied to four climatic datasets), regression models were developed using a part of the data and subsequently validated with the remaining data. Results indicated that although performances of the regression models varied from one climate to another, more or less similar prediction accuracies were obtained by all three approaches, and it was difficult to identify the best approach based on performance statistics. However, the final forms of the regression models developed by the three approaches differed substantially from one another. In all cases, the models derived using PLS contained the smallest number of predictor variables; between two to three out of a possible maximum of six predictor variables. The MLR approach yielded models with three to six predictor variables, and PCR models included all six predictor variables. This implies that the PLS regression models are the most parsimonious in terms of input data required for estimating e_{pan} from climate variables, and yet yield predictions that are almost as accurate as the more data-intensive MLR and PCR models.

DOI: 10.1061/(ASCE)0733-9437(2007)133:5(444)

CE Database subject headings: Climatic data; Correlation; Eigenvalues; Eigenvectors; Evapotranspiration; Regression analysis; Regression models; Validation; Least squares method.

Introduction

The multiple least-squares regression (MLR) technique is a popular data analysis and synthesis tool used in several fields of science and technology. The MLR approach has found widespread use, even in agronomic and irrigation studies; most notably in the development of empirical, albeit simple equations for predicting various evaporation/evapotranspiration characteristics using inputs of more routinely measured climatic variables. Doorenbos and Pruitt (1977) describe several such popular regression models developed for predicting reference crop evapotranspiration (ET_0), an important variable in procedures for computing irrigation water requirements of agricultural field crops, from standard ground-based climatological measurements. Hargreaves and Allen (2003) provide a historical review highlighting the application of MLR techniques in the development of a variety of empirical equations for estimation of ET_0. Even though more "physically based" combination-type equations for estimation of ET_0 have been subsequently developed (e.g., Allen et al. 1998), location-specific MLR-based ET_0 prediction equations continue to be developed primarily to circumvent the higher input data requirements of the combination-type methods (e.g., Irmak et al. 2003a; Nandagiri and Kovoor 2006). The simplicity and easy applicability of the MLR approach has resulted in its widespread use, even in the development of empirical equations for various climatic parameters that are involved in the estimation of ET_0 (e.g., Kotsopoulos and Babajimopoulos 1997; Irmak et al. 2003b). Another common application of the MLR technique has been in the development of regression models for estimating pan coefficients from climate/site characteristics for converting pan evaporation measurements into equivalent values of ET_0 (e.g., Snyder et al. 2005). MLR has also been used to develop models for predicting daily pan evaporation from climatic variables (e.g., Bruton et al. 2000).

However, the MLR approach is known to yield unreliable results in the presence of strong correlations between the predictor variables (multicollinearity). In the presence of multicollinearity, use of the ordinary least-squares criterion to estimate the parameters of the response function results in instability and variability of the regression coefficients (Newbold et al. 2003). When the predictor variables exhibit multicollinearity, regression coefficients derived using MLR techniques may result in large variances and signs that cannot be explained through physical

[1] Research Scholar, Dept. of Applied Mechanics and Hydraulics, National Institute of Technology Karnataka, Surathkal, Srinivasnagar P.O., Mangalore, Karnataka-575025, India. E-mail: gkovoor@lycos.com

[2] Professor, Dept. of Applied Mechanics and Hydraulics, National Institute of Technology Karnataka, Surathkal, Srinivasnagar P.O., Mangalore, Karnataka-575025, India. E-mail: lnand@rocketmail.com

Note. Discussion open until March 1, 2008. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on March 15, 2006; approved on April 19, 2007. This paper is part of the ***Journal of Irrigation and Drainage Engineering***, Vol. 133, No. 5, October 1, 2007. ©ASCE, ISSN 0733-9437/2007/5-444–454/$25.00.

reasoning (e.g., Draper and Smith 1981; Neter et al. 1996; Fekedulegn et al. 2002). In the context of MLR applications in evapotranspiration studies, multicollinearity is likely to be significant since the set of climate variables used as predictors (e.g., air temperature, humidity, windspeed, radiation) are known to exhibit high degrees of mutual correlation. Also, application of MLR approach considering all predictor variables as being important leads to the problem of model "over fit," which in turn may result in lower predictive capability of the regression model. In order to overcome this problem, "step wise" regression techniques may be employed to eliminate predictor variables that do not contribute significantly in explaining the observed variations of the response variable.

On the other hand, principal components regression (PCR) (McCuen and Snyder 1986; Haan 1995) is a multivariate statistical technique designed to handle the problem of multicollinearity and produce stable and meaningful estimates for regression coefficients. In this approach, the original predictor variables are transformed into a new set of orthogonal or uncorrelated variables called principal components of the correlation matrix. The transformation ranks the new orthogonal variables in order of their importance and thereby permits elimination of the least important principal components. Subsequently, MLR techniques are employed between the response variable and the reduced set of principal components. Because the principal components are orthogonal, they are pairwise independent, thus, ensuring absence of multicollinearity. Once the regression coefficients for the reduced set of orthogonal variables have been calculated, they are mathematically transformed into a new set of coefficients that correspond to the original or initial correlated set of variables. The PCR approach has found applications in ecological and climate studies (Fekedulegn et al. 2002; Huth 2002), but our review of literature did not indicate any previous attempts at using PCR in development of equations for evaporation/evapotranspiration estimation.

A still more recent multivariate regression technique that generalizes and combines the features from PCR and MLR is the partial least-Squares (PLS) regression approach (Abdi 2003). The method originated in social sciences and became popular in chemometrics, i.e., computational chemistry (Geladi and Kowalski 1986). The ability of PLS to extract correlation between input and output data, that is itself highly collinear, allows it to deal with problems that would be inappropriate for MLR or PCR. As in the case of PCR, PLS regression also produces factor scores as linear combinations of the original predictor variables, so that there is no correlation between the factor score variables used in the predictive regression model. However, while PCR produces the weight matrix reflecting the covariance structure between the predictor variables, PLS regression produces the weight matrix reflecting the covariance structure between the predictor and response variables.

From the above discussion, it is evident that the PCR and PLS approaches appear to have the potential to offer advantages over the conventional MLR approach in developing explanatory or predictive models from multivariate datasets. However, few earlier studies seem to have used these multivariate tools in the development of evaporation/evapotranspiration estimation models and evaluated their relative performances with a common dataset of climatic observations. Therefore, in the present study, we consider the case of developing models for predicting daily pan evaporation from climatic variables using the three multivariate statistical approaches and compare the relative prediction accuracies of the developed models. Historical climatic data obtained from four climate stations located in distinctly different climatic regimes of India were used in the comparative analysis. The objectives of the study were to: (1) investigate the applicability of the MLR, PCR, and PLS approaches in the development of regression models for estimation of pan evaporation; (2) compare the predictive capabilities of regression models developed through application of these three approaches on the same datasets; and (3) evaluate possible differences in predictive capabilities of the approaches in different climatic regimes. In the following sections of this paper, theoretical aspects relating to the three multivariate regression approaches, details of the climate datasets used, and results pertaining to the performances of the developed prediction models are discussed.

Multivariate Regression Methods

Multiple Least-Squares Regression

The general form of the multiple linear regression model is given by

$$Y = b_0 + b_1 X_1 + b_2 X_2 + , \ldots , + b_q X_q \tag{1}$$

in which Y=response or criterion variable; X_i $(i=1,2,\ldots,q)$ are the predictor variables; q=number of predictor variables, and b_i $(i=0,1,2\ldots,q)$ are the regression coefficients. Procedures for determination of the regression coefficients through application of the least-squares principle are well documented in standard texts (e.g., McCuen and Synder 1986; Haan 1995) and will not be repeated here.

However, as mentioned earlier, implementation of MLR considering all the predictor variables may lead to over fit and consequent reduction in predictive capability. To overcome this, a stepwise procedure was applied to arrive at the final form of the regression model involving only those predictor variables that can explain observed variabilities in the response variable. The objective of stepwise regression is to develop an "optimal" prediction equation by using statistical criteria to eliminate superfluous predictor variables. Based on the sequence of selecting the predictor variables, the stepwise procedure may be either the forward regression method or the backward regression method (McCuen and Synder 1986). In the forward regression method, the predictor variable having the highest correlation with the criterion variable is entered first. The next variable with the highest partial correlation is then entered. Partial correlation is the correlation of each independent variable with the dependent variable after removing the linear effect of variables already in the model. A test of significance is done at each level, and computation ends when all statistically significant variables have been included. The test of significance was done using F-statistics. If the value of F-statistics is small (<0.05), then the independent variable does a good job explaining the variation in the dependent variable. In the backward approach, one begins with an equation that includes all the predictor variables and sequentially deletes variables, with the variable contributing the least explained variance being deleted first.

In the present study, the forward regression method was used. The entire procedure was implemented using SPSS software, which was available at the Department of Community Medicine, MAHE, Manipal, India.

Table 1. Details of Climate Stations

Station	State	Latitude (N)	Longitude (E)	Altitude (m a.m.s.l)	Climate	Data period
Jodhpur	Rajasthan	26° 18′	73° 01′	224.00	Arid	1984–1987
Hyderabad	Andhra Pradesh	17° 32′	78° 16′	545.00	Semiarid	1988–1990
Bangalore	Karnataka	13° 00′	77° 37′	899.00	Subhumid	1982–1985
Pattambi	Kerala	10° 48′	76° 12′	253.60	Humid	1985–1988

Principal Components Regression

In the PCR approach, the set of correlated predictor variables is first converted into a set of orthogonal factors with the help of principal component analysis (PCA) (McCuen and Snyder 1986). Since the new sets of factors are orthogonal to each other, each of these factors contributes independently to Y. Thus if ζ_k represents the set of orthogonal factors, and X_i are the original variables that are correlated, then instead of stating the linear model as a relationship of Y to X_i, it is stated as a relationship of Y to ζ_k. Thus

$$Y = \alpha_1\zeta_1 + \alpha_2\zeta_2 + \alpha_3\zeta_3 + , \ldots , + \alpha_k\zeta_k \tag{2}$$

The new factor ζ_k is given by

$$\zeta_k = \sum_{j=1}^{P} l_{j_k} x_j k = 1,2, \ldots , P \tag{3}$$

where l_{jk}=direction are the cosines between the original and rotated axis that is given by the eigenvector matrix of the correlation matrix of x_i. Thus

$$\left.\begin{aligned} \zeta_1 &= l_{11}x_1 + l_{21}x_2 + l_{31}x_3 + \cdots \\ \zeta_2 &= l_{12}x_1 + l_{22}x_2 + l_{32}x_3 + \cdots \end{aligned}\right\rangle \tag{4}$$

The contribution of ζ_1 to y is then given by $\alpha_1\zeta_1$, which results in

$$y_1 = \alpha_1 l_{11} x_1 + \alpha_1 l_{21} x_2 + \alpha_1 l_{31} x_3 + \cdots \tag{5}$$

Similarly, the contribution of ζ_2 to y is given by $\alpha_2\zeta_2$, which gives

$$y_2 = \alpha_2 l_{12} x_1 + \alpha_2 l_{22} x_2 + \alpha_2 l_{32} x_3 + \cdots \tag{6}$$

The α_k terms in the above equations are given by

$$\alpha_k = (1/\lambda_k)(l_{ik} r_{x_1 y} + l_{2k} r_{x_2 y} + , \ldots , + l_{ik} r_{x_i y}) \tag{7}$$

Each component simply represents a relation of one independent or orthogonal element in the x_i to y. The full relationship of all the independent elements of the x_i to y is given by the sum of all the nontrivial components. Thus

$$\begin{aligned} Y = y_1 + y_2 + y_3 + \cdots &= (\alpha_1 l_{11} + \alpha_2 l_{12} + \cdots)x_1 \\ &+ (\alpha_1 l_{21} + \alpha_2 l_{22} + \cdots)x_2 + \cdots \end{aligned} \tag{8}$$

The λs are the measure on a variance scale of the information content of the components. A detailed explanation of the methodology is given by McCuen and Snyder (1986). In the present study, the eigenvalue-eigenvector analysis was performed using SPSS software. The PCA from SPSS gives the factors as the component matrix and eigenvectors are calculated from the relationship

$$\text{eigenvector} = \frac{\text{factor}}{\lambda} \tag{9}$$

where λ = eigenvalue for the particular factor.

Partial Least-Squares Regression

Partial least-squares (PLS) regression is based on linear transition from a large number of original descriptors to a new variable space based on a small number of orthogonal factors (latent variables). In other words, factors are mutually independent (orthogonal) linear combinations of original descriptors. Unlike in the case of PCR, latent variables are chosen in such a way as to provide maximum correlation with dependent variable, thus ensuring that the PLS model contains the smallest necessary number of factors. For the sake of brevity, only a conceptual description of the PLS method is given herein and the reader may refer to Abdi (2003) for a complete description of the mathematical theory underlying the approach.

Broadly, PLS works by extracting one set of latent variables for the set of manifest independent variables and another set of latent variables is extracted simultaneously for the set of manifest response (dependent) variables. The extraction process is based on decomposition of a cross-product matrix involving both the independent and response variables. Once all the latent variables have been extracted, the exact number of variables that gives the best prediction of the response variable has to be determined. This is done by a strict test of the predictive significance of each PLS component, and the optimum number of components is identified. Once the optimum number of components is identified, the PLS regression coefficients for this number of components is extracted. An analysis of these coefficients may show that all the variables are not significant. As a next step, the variables that are found to contribute significantly to the prediction of the response variable (i.e., those variables which have a regression coefficient $\geqslant 0.5$) are identified, and the entire process is repeated with only these variables. This gives us a new set of regression coefficients for the marked variables that are finally used to express the regession equation by the PLS regression method.

In the present study, the PLS analysis was carried out using a 30-day trial version of the commercial software Unscrambler developed by CAMO (http://www.camo.com/).

Methodology

Climate Data

Table 1 lists details of the climate stations considered in the analysis. These stations are drawn from a network of over 550 surface observatories operated and maintained by the India Meteorological Department (IMD), Government of India. The stations were selected to represent the major climate types prevalent in India (Subrahmanyam 1983): arid (Jodhpur), semiarid (Hyderabad), subhumid (Bangalore) and humid (Pattambi).

All stations are equipped with standard ground-based instruments; class A pan evaporimeter, alcohol and wet-bulb thermometers, sunshine recorder, cup anemometer, and mercury

Table 2. Matrix of Intervariable Correlation Coefficients

		Variables						
Station		$T_{\max}$	$T_{\min}$	$\mathrm{RH}_{\max}$	$\mathrm{RH}_{\min}$	u_2	n/N	e_{pan}
Jodhpur	$T_{\max}$	1.000	0.827	−0.032	−0.077	0.268	−0.088	0.776
(N_d=969)	$T_{\min}$		1.000	0.350	0.367	0.505	−0.413	0.679
	$\mathrm{RH}_{\max}$			1.000	0.843	0.350	−0.523	−0.103
	$\mathrm{RH}_{\min}$				1.000	0.332	−0.648	−0.189
	u_2					1.000	−0.388	0.597
	n/N						1.000	−0.007
	e_{pan}							1.000
Hyderabad	$T_{\max}$	1.000	0.604	−0.786	−0.567	0.277	0.213	0.906
(N_d=696)	$T_{\min}$		1.000	−0.327	0.240	0.530	−0.467	0.473
	$\mathrm{RH}_{\max}$			1.000	0.608	−0.191	−0.234	−0.810
	$\mathrm{RH}_{\min}$				1.000	0.177	−0.754	−0.627
	u_2					1.000	−0.389	0.414
	n/N						1.000	0.304
	e_{pan}							1.000
Bangalore	$T_{\max}$	1.000	0.533	−0.429	−0.607	−0.250	0.435	0.710
(N_d=912)	$T_{\min}$		1.000	0.021	0.108	0.143	−0.226	0.431
	$\mathrm{RH}_{\max}$			1.000	0.538	0.216	−0.370	−0.368
	$\mathrm{RH}_{\min}$				1.000	0.314	−0.617	−0.444
	u_2					1.000	−0.442	0.071
	n/N						1.000	0.246
	e_{pan}							1.000
Pattambi	$T_{\max}$	1.000	0.182	−0.328	−0.727	0.231	0.500	0.652
(N_d=850)	$T_{\min}$		1.000	0.203	0.235	0.115	−0.231	−0.021
	$\mathrm{RH}_{\max}$			1.000	0.510	−0.390	−0.353	−0.409
	$\mathrm{RH}_{\min}$				1.000	−0.299	−0.633	−0.672
	u_2					1.000	0.313	0.390
	n/N						1.000	0.597
	e_{pan}							1.000

thermometers. Readings are taken twice a day at 08.30 h and 17.30 h Indian Standard Time. Records are transmitted from the stations to the IMD Data Centre at Pune where data archives are maintained. Data is scrutinized and subjected to quality checks prior to supply to users.

Historical data were procured from IMD for the periods shown against each station in Table 1. Unfortunately, good quality data were unavailable for a common period for all the stations. For each station, the data set used in this study comprised daily values of maximum air temperature ($T_{\max}$), minimum air temperature ($T_{\min}$), maximum relative humidity ($\mathrm{RH}_{\max}$), minimum relative humidity ($\mathrm{RH}_{\min}$), actual hours of sunshine (n), 24-h wind speed (u_z) at 3 m height and pan evaporation depth (e_{pan}). Individual data records were subjected to further screening, and integrity checks were performed on the climatic variables as per procedures described in Allen et al. (1998) (results not presented here for brevity). After discarding obvious outliers and accounting for missing records using techniques suggested by Allen et al. (1998), the number of days (N_d) for which complete records were available for each station is: Jodhpur 1453, Hyderabad 1044, Bangalore 1368, and Pattambi 1275. Two-thirds of this data set, i.e., Jodhpur 969, Hyderabad 696, Bangalore 912, and Pattambi 850 was used in the development (calibration) of the regression models, and the remaining data set was set apart for validation of the developed models.

Application

In order to meet the objectives of this study, the MLR, PCR, and PLS techniques were used separately to develop regression models relating daily pan evaporation (e_{pan}) to various climatic variables. Separate regression models were fitted to historical data records of climatic variables at the four lysimetric locations (Jodhpur, Hyderabad, Bangalore, and Pattambi) of interest to this research. In all cases (three regression methods applied to climate datasets of four stations), recorded daily pan evaporation (e_{pan}) values were considered to be the response (dependent) variable and corresponding daily averages of $T_{\max}$, $T_{\min}$, $\mathrm{RH}_{\max}$, $\mathrm{RH}_{\min}$, u_2, and n/N were considered to be the predictor (independent) variables. In addition, all regression models were developed using 67% (Jodhpur 969, Hyderabad 696, Bangalore 912, and Pattambi 850) of available daily records (calibration phase) and subsequently tested using the remaining 33% (Jodhpur 484, Hyderabad 348, Bangalore 456, and Pattambi 425) of records (validation phase). Performances of models developed in the validation phase were evaluated by computing standard error of estimate (SEE), coefficient of determination (R^2), and standard deviations of estimates (σ), statistics considering predicted and observed values of e_{pan}. The best model was one with the smallest values of SEE and σ, and the highest value of R^2.

Multiple linear regression equations of the form specified by Eq. (1) were developed using the forward stepwise regression module available in SPSS software. For the development of models based on PCR, eigenvalue-eigenvector analysis was carried out using SPSS. The rotated component matrix obtained from SPSS was subsequently exported to Microsoft Excel spreadsheet software, in which the components are added one by one to obtain the optimum number of components. In a subsequent step, regression coefficients were computed.

PLS regression modeling was carried out using Unscrambler software program. The data set for the independent and dependent variables for each station was fed into Unscrambler. The optimum number of components is identified by the software, and for this number of components, the regression coefficients are extracted. The variables with regression coefficients greater than 0.5 were marked, and the PLS regression analysis was repeated with only the marked variables. The final PLS regression coefficients were obtained from the output.

Results and Discussions

Factor Analysis

Table 2 shows the matrix of correlation coefficients between the variables considered in the analysis for all four stations. It is interesting to note from these results the high degree of correlation between a few of the predictor variables and the response variable. While this is a desirable feature from the viewpoint of developing regression models using the least-squares approach, the existence of equally high degrees of correlation between the predictor variables themselves highlights the complexity of the problem. For instance, at the arid Jodhpur site, a significantly high correlation of 0.776 exists between $T_{\max}$ and e_{pan}, but an even higher degree of correlation (0.827) exists between $T_{\max}$ and $T_{\min}$. Similarly, high correlation is observed between $\text{RH}_{\max}$; $\text{RH}_{\min}$ and $\text{RH}_{\min}$: n/N. At the Hyderabad site too, $T_{\max}$ has the highest correlation of 0.906 with e_{pan}, but this predictor variable has correlation coefficients of -0.786 with $\text{RH}_{\max}$ and 0.604 with $T_{\min}$. A similar pattern is observed at the subhumid site Bangalore with correlation coefficients of e_{pan}: $T_{\max}=0.710$, $T_{\max}$: $\text{RH}_{\min}=-0.607$, $T_{\max}$: $T_{\min}=0.533$. With regard to the humid location Pattambi, from among the predictor variables, the highest correlation with e_{pan} is exhibited by $\text{RH}_{\min}$ (-0.671), but correlations of $\text{RH}_{\min}$: $T_{\max}=-0.727$ and $\text{RH}_{\min}$: $n/N=-0.633$ may be observed. Due to the existence of such strong degrees of multicollinearity, the relative importance of the predictor variables cannot be understood clearly from the correlation matrices alone. Therefore, the datasets were subjected to PCA and factor analysis (FA) (McCuen and Snyder 1986), and factors that are a linear combination of the meteorological variables were extracted. The amount of variance explained by each component is shown in Table 3.

Table 3. Total Variance Explained by Components Extracted from Factor Analysis

		Initial eigenvalues		
Station	Component	Total	% of variance	Cumulative %
Jodhpur	1	3.227	46.100	46.100
	2	2.285	32.649	78.749
	3	0.708	10.110	88.859
	4	0.495	7.066	95.925
	5	0.151	2.160	98.085
	6	0.078	1.111	99.196
	7	0.056	0.804	100.000
Hyderabad	1	3.543	50.619	50.619
	2	2.300	32.862	83.481
	3	0.570	8.146	91.628
	4	0.342	4.887	96.515
	5	0.146	2.088	98.603
	6	0.058	0.835	99.438
	7	0.039	0.562	100.000
Bangalore	1	3.088	44.113	44.113
	2	1.736	24.795	68.908
	3	0.835	11.931	80.839
	4	0.593	8.477	89.315
	5	0.332	4.747	94.062
	6	0.305	4.356	98.418
	7	0.111	1.582	100.000
Pattambi	1	3.417	48.809	48.809
	2	1.217	17.383	66.193
	3	0.923	13.182	79.375
	4	0.599	8.557	87.932
	5	0.361	5.153	93.086
	6	0.328	4.680	97.766
	7	0.156	2.234	100.000

At all stations, it can be seen that the first four components (out of maximum possible seven) explain more than 85% of the total variability. These factors were then rotated using a component transformation matrix into factors that are orthogonal, thereby ensuring that the multicollinearity effect was eliminated. The Varimax method of rotation was implemented in stages. From these results, those solutions that provided the best interpretation of the relative importance of the predictor variables on the response variable for each dataset are presented in Table 4.

At the arid Jodhpur site, four components were found to explain about 96% of the variance. The rotated component matrix (Table 4) shows that all the variables are loaded heavily on the four components. However, the first factor, which has the highest loading for e_{pan} (Table 4), shows that the temperature related variables are the most important. The second important factor with the next highest loading for e_{pan} is factor 3, which indicates a higher loading for u_2. The results are found to match the findings of Nandagiri and Kovoor (2006), where a similar analysis was done with the computed values of FAO-56 Penman-Monteith ET_0 (Allen et al. 1998) as the dependent variable.

At the semiarid site Hyderabad, the two-factor solution explains about 83% of the variance (Table 3). The first factor with highest loading for e_{pan} is found to exhibit high loadings for $T_{\max}$, $\text{RH}_{\max}$, and $\text{RH}_{\min}$ (Table 4). The next component also gives relatively high loading for the remaining three variables, $T_{\min}$, u_2, and n/N.

At the subhumid Bangalore site, the three-factor solution explains about 81% of the variance (Table 3). The second factor with highest loading for e_{pan} has temperature related variables as the highest loading variables (Table 4). The next factor with highest loading for e_{pan} has $\text{RH}_{\max}$ and $\text{RH}_{\min}$ as the variables with higher loadings on this factor, and the third factor has u_2 as the highest loading variable. However, it should be noted that $T_{\max}$ and $\text{RH}_{\min}$ have a correlation of -0.607 and $\text{RH}_{\max}$ and $\text{RH}_{\min}$ have a correlation of 0.538, which is significant.

Table 4. Rotated Component Matrix for the Four Stations

		Rotated component matrix			
		Components			
Station	Variables	1	2	3	4
Jodhpur	$T_{\max}$	0.9800	−0.0789	0.0581	−0.0153
(N_d=969)	$T_{\min}$	0.8770	0.3100	0.2090	−0.2180
	$\mathrm{RH}_{\max}$	0.0272	0.9490	0.1260	−0.1370
	$\mathrm{RH}_{\min}$	−0.0121	0.8910	0.0689	−0.3630
	u_2	0.2220	0.2290	0.9220	−0.1750
	n/N	−0.0955	−0.3720	−0.1420	0.9060
	e_{pan}	0.7590	−0.2110	0.5670	0.0887
Hyderabad	$T_{\max}$	0.9310	0.2180		
(N_d=696)	$T_{\min}$	0.3780	0.8330		
	$\mathrm{RH}_{\max}$	−0.8830	−0.0536		
	$\mathrm{RH}_{\min}$	−0.7590	0.5850		
	u_2	0.2510	0.7390		
	n/N	0.4300	−0.8070		
	e_{pan}	0.9590	0.1650		
Bangalore	$T_{\max}$	−0.5190	0.7630	−0.2590	
(N_d=912)	$T_{\min}$	0.2310	0.9230	0.0830	
	$\mathrm{RH}_{\max}$	0.7830	−0.0765	0.0116	
	$\mathrm{RH}_{\min}$	0.8440	−0.1150	0.2600	
	u_2	0.1180	0.0233	0.9430	
	n/N	−0.6620	−0.0566	−0.5230	
	e_{pan}	−0.5420	0.6870	0.1770	
Pattambi	$T_{\max}$	0.7820	0.3360	0.3740	−0.1780
(N_d=850)	$T_{\min}$	−0.1380	0.9540	0.0151	−0.1000
	$\mathrm{RH}_{\max}$	−0.6450	0.2140	0.4560	0.5260
	$\mathrm{RH}_{\min}$	−0.8870	0.1290	−0.1940	0.1100
	u_2	0.5190	0.2850	−0.7110	0.3060
	n/N	0.7770	−0.1800	0.1370	0.4040
	e_{pan}	0.843	0.1340	0.1120	0.1060

At the humid Pattambi site, even though the three-factor solution can provide an interpretation of the relative importance of variables, the three factors together explain only 79% of the variance (Table 3). Therefore, the results for this site were analyzed using the four-factor solution, which explains a total variance of 88%. The first factor, which has the highest loading for e_{pan}, has $\mathrm{RH}_{\min}$, $T_{\max}$, and n/N as the heavily loaded variables (Table 4). Except $T_{\min}$, the other variables $\mathrm{RH}_{\max}$ and u_2 are also found to give relatively high loadings on the first factor.

Stepwise Regression Analysis

Results of factor analysis clearly showed that not all predictor variables are significant at all stations. This finding justified the need to use a stepwise procedure in the development of multiple linear regression models between the response variable and the predictor variables. Application of the stepwise regression procedure in SPSS software program yielded the results summarized in Table 5. Shown therein are the regression coefficients associated with the predictor variables that were included in the final forms of the regression models. Additionally shown in Table 5 are the R^2 and SEE (mm/d) values obtained during the calibration phase. It can be seen that the number of predictor variables included in the final multiple linear regression models for e_{pan} varies from 3 for Bangalore to the maximum of 6 for Hyderabad.

At the arid Jodhpur site, $T_{\max}$, which had the highest correlation with e_{pan}, was the first variable to be entered. In the next step, u_2, which has the highest partial correlation, was entered. In a similar manner, variables $\mathrm{RH}_{\min}$, $T_{\min}$, and n/N gained successive entry into the model. $\mathrm{RH}_{\max}$, which was the only remaining variable, was not entered since its partial correlation was very low. In addition, since this variable had a significant value of 0.473, which was greater than the cutoff of 0.05, it was not considered to be statistically significant. A high value of R^2 and low value of SEE are indicative of goodness of fit in the calibration phase.

At the semiarid Hyderabad site, stepwise regression resulted in all the predictor variables finding a place in the final model. The fact that in the factor analysis all the variables were heavily loaded on the two factors (Table 4), provides an explanation to this result. Again, an R^2 value of 0.922 and SEE of 0.857 mm/d indicate an extremely good model fit.

In contrast to the Hyderabad site, at the subhumid Bangalore site, only three variables $T_{\max}$, u_2, and $\mathrm{RH}_{\max}$ were entered into the final stepwise regression equation. Even though the R^2 value of the final model is only 0.58, entry of the remaining variables did not improve goodness of fit since the partial correlations of the remaining variables were $T_{\min}$=0.009, $\mathrm{RH}_{\min}$=−0.056, and n/N=0.016, and their significant values were 0.782, 0.089, and 0.621, respectively.

Table 5. Models Derived Using Stepwise Multiple Least-Squares Regression

Station	Variables	Regression coefficients	R^2	SEE (mm/d)
Jodhpur	Constant	−7.7110		
(N_d=969)	$T_{\max}$	0.2920		
	$T_{\min}$	0.1420		
	$\mathrm{RH}_{\max}$		0.866	1.473
	$\mathrm{RH}_{\min}$	−0.0773		
	u_2	0.0213		
	n/N	1.8600		
Hyderabad	Constant	−4.5520		
(N_d=696)	$T_{\max}$	0.3440		
	$T_{\min}$	0.0623		
	$\mathrm{RH}_{\max}$	−0.0399	0.922	0.857
	$\mathrm{RH}_{\min}$	−0.0268		
	u_2	0.0080		
	n/N	2.1100		
Bangalore	Constant	−5.9620		
(N_d=912)	$T_{\max}$	0.4510		
	$T_{\min}$			
	$\mathrm{RH}_{\max}$	−0.0202	0.580	1.256
	$\mathrm{RH}_{\min}$			
	u_2	0.0044		
	n/N			
Pattambi	Constant	−4.5340		
(N_d=850)	$T_{\max}$	0.2580		
	$T_{\min}$			
	$\mathrm{RH}_{\max}$		0.576	1.390
	$\mathrm{RH}_{\min}$	−0.0344		
	u_2	0.0107		
	n/N	2.1590		

Table 6. Models Derived Using Principal Components Regression

Station	No. of components	Variables	Regression coefficients
Jodhpur		Constant	2.1282
(N_d=969)		$T_{\max}$	0.1664
		$T_{\min}$	0.1196
	4	$\mathrm{RH}_{\max}$	−0.0130
		$\mathrm{RH}_{\min}$	−0.0124
		u_2	0.0012
		n/N	−1.0250
Hyderabad		Constant	3.4253
(N_d=696)		$T_{\max}$	0.1388
		$T_{\min}$	0.0694
	2	$\mathrm{RH}_{\max}$	−0.0330
		$\mathrm{RH}_{\min}$	−0.0201
		u_2	0.0020
		n/N	0.6047
Bangalore		Constant	2.7225
(N_d=912)		$T_{\max}$	0.1159
		$T_{\min}$	0.1132
	3	$\mathrm{RH}_{\max}$	−0.0124
		$\mathrm{RH}_{\min}$	−0.0113
		u_2	−0.0004
		n/N	0.3509
Pattambi		Constant	3.0114
(N_d=850)		$T_{\max}$	0.1113
		$T_{\min}$	−0.0059
	3	$\mathrm{RH}_{\max}$	−0.0258
		$\mathrm{RH}_{\min}$	−0.0215
		u_2	0.0064
		n/N	1.2327

At the humid site (Pattambi), four variables $\mathrm{RH}_{\min}$, $T_{\max}$, n/N, and u_2 were entered sequentially resulting in a final model with R^2=0.576 and SEE=1.39 mm/d. The other two variables were not entered due to poor partial correlation ($T_{\min}$=0.02 and $\mathrm{RH}_{\max}$=−0.061) with significant values of 0.554 and 0.076, respectively.

Principal Components Regression Analysis

Components regression analysis was performed on the same climate data sets used in the multiple linear stepwise regression analysis. The coefficients of the new variates in the linear model for each of the components extracted and the corresponding correlation coefficients were then computed. Since the full relationship of all the independent elements of X_i to Y is given by the sum of all the nontrivial components, these components are added one by one, keeping in mind the total variance explained at each step and the nature of the relationship exhibited by the regression coefficient of each variable developed as compared to the Y correlations. Table 6 shows the final regression coefficients developed for the original variables for each of the four stations.

At the arid Jodhpur site, the sum of four components is taken. This gives a total eigenvalue (λ) of 6.7 out of a possible maximum of 7 and the regression coefficients for $T_{\max}$, $T_{\min}$, and u_2 are positive with values equal to 0.1664, 0.1196, and 0.0012, respectively. However, the coefficients for $\mathrm{RH}_{\max}$, $\mathrm{RH}_{\min}$, and n/N are negative with values equal to −0.0130, −0.0124, and −1.0250, respectively (Table 6), which is in agreement with the Y correlations shown in Table 2. A relatively higher regression coefficient associated with $T_{\max}$ indicates the importance of this variable, whereas the lower value (negative) of n/N is indicative of the smaller significance of the radiation term in the arid Jodhpur site.

At the semiarid Hyderabad site, the sum of just two components explains a variance of about 84%, and hence the regression coefficients are taken from the sum of the first and second components. Only $\mathrm{RH}_{\max}$ and $\mathrm{RH}_{\min}$ have negative coefficients, the same variables that exhibit negative correlations with the response variable (Table 2). Here again, n/N has the highest regression coefficient (0.6047), followed by $T_{\max}$, with a value of 0.1388 (Table 6), indicative of their importance in determining the magnitude of the response variable.

At the subhumid Bangalore site, the number of components increases to three and the total variance explained by these three components is 81%. However, when the Y correlation is considered, only $\mathrm{RH}_{\max}$ and $\mathrm{RH}_{\min}$ have negative correlations, whereas u_2 also exhibits a negative regression coefficient. However, the fact that the coefficient for this variable is only −0.00041, permits its acceptance. On the other hand, even though the sum of two components yields coefficients consistent with the nature of the Y correlation, the fact that these two factors explain only 69% of the variance does not permit acceptance of this combination. At this site, the highest regression coefficient was associated with n/N (0.3509), whereas the coefficients for the temperature vari-

Table 7. Models Derived Using Partial Least-Squares Regression

Station	Variables	Regression coefficients
Jodhpur	Constant	−4.2219
(N_d=969)	$T_{\max}$	0.2651
	$T_{\min}$	0.2832
	$\mathrm{RH}_{\max}$	
	$\mathrm{RH}_{\min}$	−0.0729
	u_2	
	n/N	
Hyderabad	Constant	−14.8696
(N_d=696)	$T_{\max}$	0.7163
	$T_{\min}$	−0.0754
	$\mathrm{RH}_{\max}$	
	$\mathrm{RH}_{\min}$	
	u_2	
	n/N	0.0465
Bangalore	Constant	−6.5322
(N_d=912)	$T_{\max}$	0.4047
	$T_{\min}$	0.0528
	$\mathrm{RH}_{\max}$	
	$\mathrm{RH}_{\min}$	
	u_2	
	n/N	
Pattambi	Constant	−10.3158
(N_d=850)	$T_{\max}$	0.3764
	$T_{\min}$	
	$\mathrm{RH}_{\max}$	
	$\mathrm{RH}_{\min}$	
	u_2	
	n/N	3.4038

ables $T_{\max}$ and $T_{\min}$ are almost the same (0.1159 and 0.1132) (Table 6). From this result, it may be inferred that e_{pan} estimates are considerably more influenced by the actual number of sunshine hours, and to a lesser degree by air temperature in the subhumid climate.

At the humid location (Pattambi), it is again the sum of three components that was found to give the best combination of regression coefficients. While the three components together explain about 80% of the total variance, $T_{\min}$, $\mathrm{RH}_{\max}$, and $\mathrm{RH}_{\min}$ have negative coefficients, and these variables are the ones that exhibit negative Y correlation (Table 2). At this site, it is found that n/N has the highest regression coefficient (Table 6), indicating the importance of this variable in predicting e_{pan}.

Partial Least-Squares Regression

Partial least-squares regression was also implemented on the same climate data sets using the Unscrambler software program. The regression coefficients obtained from PLS regression are given in Table 7.

The optimum number of components at the arid Jodhpur site was found to be three. These three components explain about 86% of the variance and root mean square error (RMSE) was found to be 1.5096 mm/d for the calibration and 1.5179 mm/d in the cross-validation. Hence, in the first step of PLS regression, the coefficients computed with the three components were extracted. $T_{\max}$, $T_{\min}$, and $\mathrm{RH}_{\min}$ were the variables that were found to have coefficients greater than 0.05. The coefficients obtained by recalculation with these variables are shown in Table 7.

At the semiarid site Hyderabad, the optimum number of components was identified as six, and this explained about 92% of the variance. The RMSE values for calibration and cross-validation were 0.8565 mm/d and 0.8690 mm/d, respectively. Recalculation was done with $T_{\max}$, $T_{\min}$, and n/N, which gave coefficients greater than 0.05 with the six components.

At the subhumid Bangalore site, only the temperature variables were found to contribute significantly when the regression coefficients with optimum number of five components were extracted. These five components explained about 58% of the variance and RMSE values were found to be 1.2492 mm/d and 1.2538 mm/d for calibration and validation, respectively.

At the humid Pattambi site, $T_{\max}$ and n/N were the only two variables that were included in the second step of calculation. The optimum number of components was identified as six, which explained about 57% of the variance.

Validation of Regression Models

The stepwise regression equations, the component regression equations, and the partial least-squares regression equations were validated using one-third of the total data set that was set aside for this purpose. The performances of the regression equations developed were compared by the following statistics: Standard error of estimate (SEE) statistic, standard deviations of the estimates (σ), and coefficient of determination (R^2). These statistics for the validation phase of each of the approaches are given in Table 8.

The scatter plots between the measured e_{pan} values and those computed by final models derived from the stepwise regression method, the component regression method, and the partial least-squares regression methods are shown in Figs. 1–3.

From these results, several interesting observations with regard to the performances of individual methods in different climates and also with regard to relative prediction accuracies between methods, can be made.

Table 8. Performance of Developed Regression Models during Validation

	Stepwise MLR			PCR			PLS		
Station	SEE (mm/d)	R^2	σ (mm/d)	SEE (mm/d)	R^2	σ (mm/d)	SEE (mm/d)	R^2	σ (mm/d)
Jodhpur (N_d=484)	1.3708	0.8608	3.4306	2.4207	0.6852	1.7493	2.2844	0.6188	3.2297
Hyderabad (N_d=348)	0.7941	0.8988	2.4748	1.3029	0.8799	1.3365	1.0564	0.8182	2.3599
Bangalore (N_d=456)	1.2067	0.3816	1.4819	1.0542	0.3720	0.7014	1.1108	0.3656	1.2673
Pattambi (N_d=425)	1.4320	0.4515	1.3499	1.4637	0.4645	0.7897	1.5842	0.3612	1.2710
Mean	1.2009	0.6482	2.1843	1.5604	0.6004	1.1442	1.5090	0.5410	2.0320

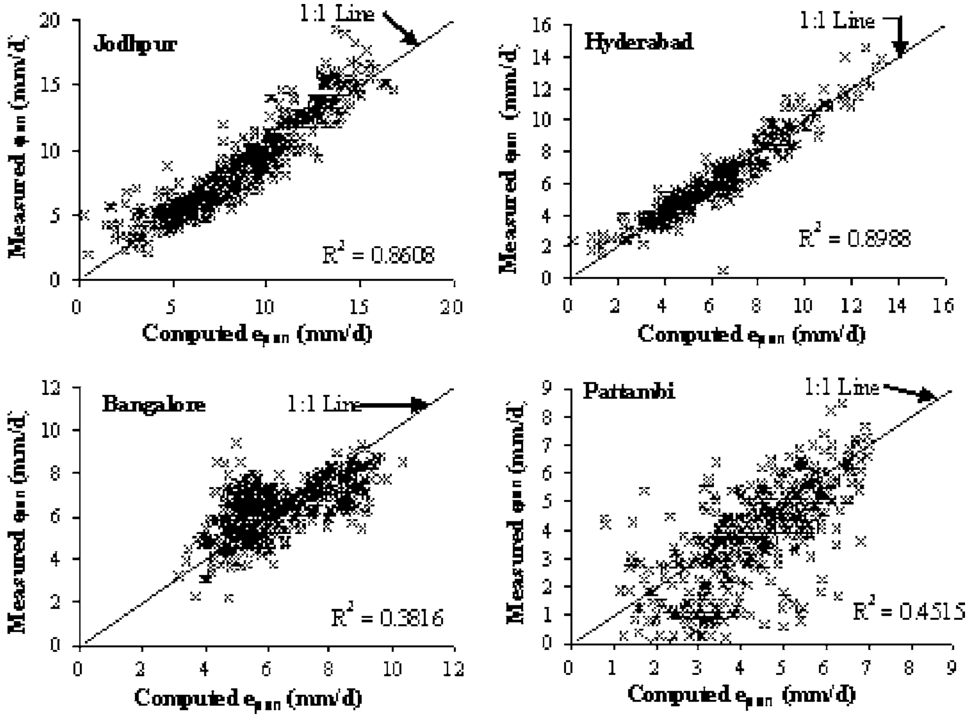

Fig. 1. Comparison of observed daily pan evaporation with those computed using MLR models

The multiple linear regression models for e_{pan} developed through the stepwise approach appear to yield the best results at the semiarid Hyderabad site considering the values of SEE and R^2 (0.7941 and 0.8988 mm/d, respectively). In terms of SEE alone, the performance of the MLR models are more or less similar at the other three sites (Table 8). However, R^2 values are considerably lower at the subhumid Bangalore and humid Pattambi sites, and larger scatter can be observed in Fig. 1 for these locations. However, σ values are significantly lower at these sites. Overall, the performances of the MLR models across the sites in the validation phase are more or less similar to their performances in the calibration phase (Table 5).

Validation results for regression models developed using the PCR approach are somewhat similar to the MLR models

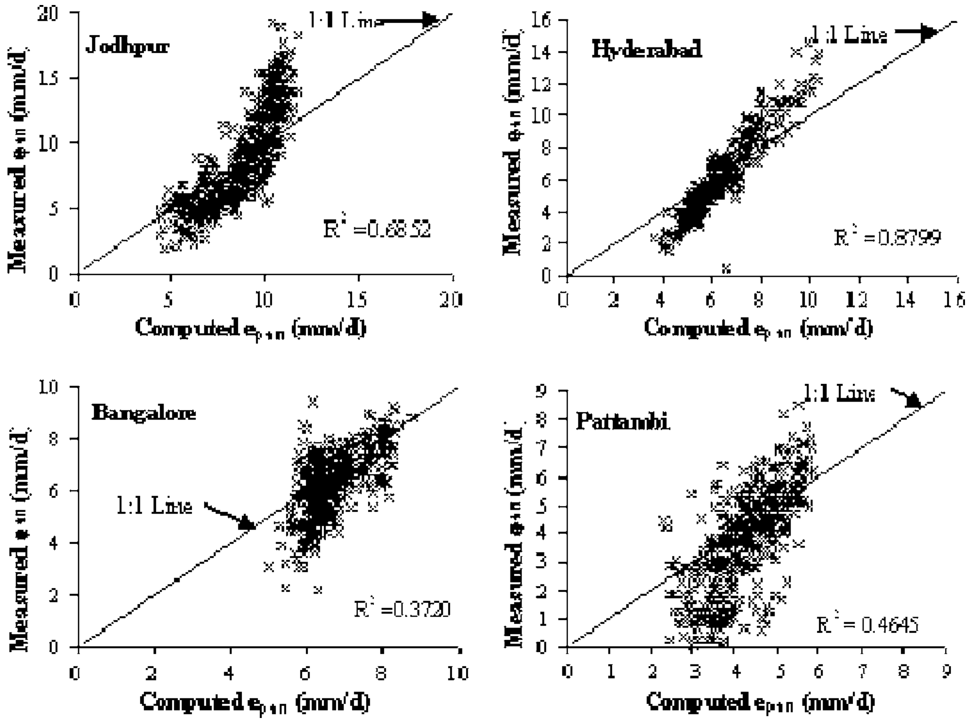

Fig. 2. Comparison of observed daily pan evaporation with those computed using PCR models

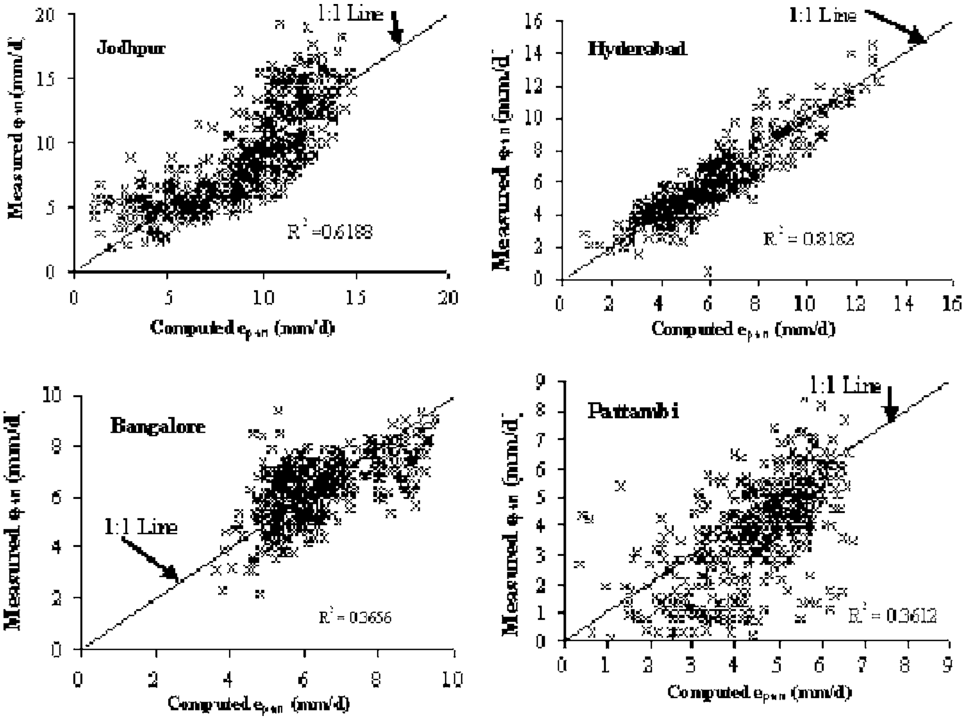

Fig. 3. Comparison of observed daily pan evaporation with those computed using PLS regression models

(Table 8). Considering the R^2 statistic, the best predictions were obtained at the semiarid (Hyderabad site), predictions were reasonably good at the Jodhpur (arid) site, and model performances at the other two sites were poor. SEE was highest at the arid site and moderate at the remaining three sites. A similar pattern was evident with regard to σ.

Interestingly, the climate-dependent predictive capabilities of the regression models developed using PLS followed the same pattern as the other two approaches (Table 8). Predictions seemed best at the semiarid location both in terms of R^2 and SEE values; whereas at the arid site, R^2 was reasonably good but SEE was the highest among all PLS models developed.

In regard to the relative comparison between the performances of MLR, PCR, and PLS regression models, overall results shown in Table 8 do not indicate substantial differences in predictive capabilities. All three approaches appear to provide the best predictions at the semiarid Hyderabad site, and moderately good predictions at the arid site. The performances of all three methods are poor at the subhumid (Bangalore) and humid (Pattambi) sites.

Considering performances across all stations, mean statistics shown in Table 8 indicate that the MLR method yielded the highest R^2 and lowest SEE values in comparison to the other two methods. However, σ was highest for this method, indicating larger variabilities in estimates of the response variable. The PCR and PLS methods yielded almost similar values of SEE, but the former method clearly appeared to outperform the latter in terms of both R^2 and σ values.

However, even though predictive capabilities appear more or less similar, it is very important to note the differences in input data (predictor variables) required by the final models derived from the three approaches. For instance, the number of predictor variables included in the final regression models derived by the stepwise MLR approach varies from as high as six at the Hyderabad site to as small as three at the subhumid Bangalore site (Table 5). In contrast, it can be seen from Table 6 that the PCR models for all four sites involve all the predictor variables indicating the highly data-intensive nature of these models. The PLS models appear to be the most parsimonious in terms of input data requirements, since the number of predictor variables involved in the final models is three for the arid and semiarid sites, and only two for the other two sites (Table 7).

Therefore, it appears that among all the multivariate regression approaches used in this study to develop models for predicting pan evaporation from climatic variables, the PLS approach provides the most optimal models in terms of the number of predictor variables needed to produce predictions that are comparable to those obtained from the MLR and PCR approaches. While the MLR models are not as parsimonious as the PLS models, they are less data-intensive than the PCR models. However, the power of the PCR models may lie in their capability as explanatory tools rather than as predictive tools.

Conclusions

In spite of certain inherent limitations when applied to datasets comprising significant degree of correlation between the predictor variables (multicollinearity), the multiple least-squares regression (MLR) approach has found wide applications in the development of empirical models for estimation of evaporation/evapotranspiration rates from climatic observations and also in calculating several other climate-dependent parameters. In order to circumvent problems associated with multicollinearity, two other multivariate regression techniques: principal components regression (PCR) and partial least squares (PLS) regression have been developed and are widely used in social sciences and chemistry. However, few attempts have been made to apply them in evapotranspiration studies and compare their predictive capabilities relative to the MLR approach. Therefore, in the present study, an attempt was made to explore the applicability of the MLR, PCR, and PLS approaches in the development of regression models for predicting daily pan evaporation depths (e_{pan}) from climate variables ($T_{\max}$, $T_{\min}$, $\mathrm{RH}_{\max}$, $\mathrm{RH}_{\min}$, u_2, and n/N). The objective was to develop regression models by the three approaches for datasets obtained from four distinct climate regimes in India and to evaluate the relative prediction accuracies of the developed models.

Separate regression models were developed using stepwise MLR, PCR, and PLS approaches using a part of the available historical daily climate records at the four Indian sites representing the major climatic regimes: Jodhpur arid, Hyderabad semiarid, Bangalore sub-humid, and Pattambi humid. These models were subsequently validated for their prediction accuracies (quantified in terms of SEE, R^2, and σ statistics between estimated and observed e_{pan} values) using the remaining climate data that was not used in the calibration exercise. Results indicated that the performances of the regression model developed using a particular approach varied from one climate to another. However, the same pattern was exhibited by all the approaches. That is, regression models developed from MLR, PCR, and PLS approaches were most accurate at the semiarid Hyderabad site (SEE between $0.7941-1.3029$ mm/d), reasonably good at the arid Jodhpur site (SEE between $1.3708-2.4207$ mm/d), and poor at the subhumid Bangalore site (SEE between $1.0542-1.2067$ mm/d), and humid Pattambi site (SEE between $1.4320-1.5842$ mm/d). At a given site, more or less similar prediction accuracies were obtained by all three approaches, and it was difficult to identify the best approach based on performance statistics. However, the final forms of the regression models developed by the three approaches differed substantially from one another. In all cases, the models derived using PLS contained the smallest number of predictor variables, between two to three out of a possible maximum of six predictor variables. The MLR approach yielded models with three to six predictor variables and PCR models included all six predictor variables. This implies that the PLS regression models are the most parsimonious in terms of input data required for estimating e_{pan} from climate variables, and yet yield predictions that are almost as accurate as the more data-intensive MLR and PCR models. While accepting that our conclusions are specific to the datasets analyzed, the findings of this study highlight the need for more extensive testing of the advantages offered by the PCR and PLS regression approaches, relative to the popular MLR approach.

Notation

The following symbols are used in this paper:

e_{pan} = pan evaporation (mm/d);
ET_0 = reference crop ET (mm/d);
N = maximum possible duration of sunshine (hours);
N_d = number of data points;
n = actual duration of sunshine (hours);
PCA = principal component analysis;
PCR = principal components regression;
PLS = partial least-squares regression;
PRCC = partial rank correlation coefficient;
R^2 = coefficient of determination of the linear fit;
$\mathrm{RH}_{\max}$ = maximum relative humidity (%);
$\mathrm{RH}_{\min}$ = minimum relative humidity (%);
$T_{\max}$ = maximum air temperatures (°C);
$T_{\min}$ = minimum air temperatures (°C);
u_2 = 24 h wind speed (m/s) at 2 m height;
ζ_k = set of orthogonal factors used in PCR;
λ = eigenvalue for the particular factor; and
σ = standard deviation.

References

Abdi, H. (2003). "Partial least squares regression (PLS-regression)." *Encyclopedia for research methods for the social sciences*, M. Lewis-Beck, A. Bryman, and T. Futing, eds., Sage, Thousand Oaks, Calif.

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). "Crop evapotranspiration—Guidelines for computing crop water requirements." *Food and Agricultural Organization of the United Nations (FAO) irrigation and drain. paper No. 56*, Rome.

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

Doorenbos, J., and Pruitt, W. O. (1977). "Guidelines for predicting crop water requirements." *FAO Irrigation and drain. paper No. 24*, Rome.

Draper, N. R., and Smith, H. (1981). *Applied regression analysis*, 2nd Ed., Wiley, New York.

Fekedulegn, B. D., Colbert, J. J., Hicks, R. R., Jr., and Schuckers, M. E. (2002). "Coping with multicollinearity: An example on application of principal component regression in dendroecology." *United States Department of Agriculture Forest Service*, Newtown Square, Pa ⟨www.fs.fed.us/ne⟩.

Geladi, P., and Kowalski, B. (1986). "Partial least-squares regression: A tutorial." *Anal. Chim. Acta*, 185, 1–17.

Haan, C. T. (1995). *Statistical methods in hydrology*, Affiliated East–West Press Pvt. Ltd., New Delhi, India.

Hargreaves, G. H., and Allen, R. G. (2003). "History and evaluation of Hargreaves evapotranspiration equation." *J. Irrig. Drain. Eng.*, 129(1), 53–63.

Huth, R. (2002). "Sensitivity of climate change estimates using statistical downscaling to the method and predictors." *Proc., 13th Symp. on Global Change and Climate Variations*, Amer. Soc. of Meteorology, Orlando, Fla.

Irmak, S., Irmak, A., Allen, R. G., and Jones, J. W. (2003a). "Solar and net radiation-based equations to estimate reference evapotranspiration in humid climates." *J. Irrig. Drain. Eng.*, 129(5), 336–347.

Irmak, S., Irmak, A., Jones, J. W., Howell, T. A., Jacobs, J. M., Allen, R. G., and Hoogenboom, G. (2003b). "Predicting daily net radiation using minimum climatological data." *J. Irrig. Drain. Eng.*, 129(4), 256–269.

Kotsopoulos, S., and Babajimopoulos, C. (1997). "Analytical estimation of modified Penman equation parameters." *J. Irrig. Drain. Eng.*, 123(4), 253–256.

McCuen, R. H., and Snyder, W. M. (1986). *Hydrologic modeling: Statistical methods and applications*, Prentice-Hall, Englewood Cliffs, N.J.

Nandagiri, L., and Kovoor, G. M. (2006). "Performance evaluation of reference evapotranspiration equations across a range of Indian climates." *J. Irrig. Drain. Eng.*, 132(3), 238–249.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear regression models*, 3rd Ed., Irwin Book Team, Chicago.

Newbold, P., Carlson, W. L., and Thorne, B. M. (2003). *Statistics for business and economics*, 5th Ed., Pearson Education Inc., Upper Saddle River, N.J.

Snyder, R. L., Orang, M., Matyac, S., and Grismer, M. E. (2005). "Simplified estimation of reference evapotranspiration from pan evaporation data in California." *J. Irrig. Drain. Eng.*, 131(3), 249–253.

Subrahmanyam, V. P. (1983). "Some aspects of water balance in the tropical monsoon climates of India." *Proc., Hamburg Symp.*, I.A.H.S. Publication No. 140, 325–331.