Estimating Evapotranspiration using Machine Learning Techniques

Muhammad Adnan
Institute of Manufacturing Information and Systems, Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan City 701, Taiwan

M. Ahsan Latif
Department of Computer Science, University of Agriculture, Faisalabad, Pakistan

Abaid-ur-Rehman
Department of Computer Science, University of Agriculture, Faisalabad, Pakistan

Maria Nazir
Department of Computer Science, COMSAT, Lahore, Pakistan

***Abstract*—The measurement of evapotranspiration is the most important factor in irrigation scheduling. Evapotranspiration means loss of water from the surface of plant and soil. Evaporation parameters are being used in studying water balances, water resource management, and irrigation system design and for estimating plant growth and height as well. Evapotranspiration is measured by different methods by using various parameters. Evapotranspiration varies with the climate change and as the climate has a lot of variation geographically, the pre-developed systems have not used all available meteorological data hence not robust models. In this research work, a model is developed to estimate evapotranspiration with more authentic and accurate reduced meteorological parameters using different machine learning techniques. The study reveals to learn and generalize the relationship among different parameters. The dataset with reduced dimension is modeled through time series neural network giving the regression value R=83%.**

Keywords*—Evapotranspiration; principle component analysis; neural network; irrigation scheduling

I. Introduction

Appropriate irrigation scheduling enhances crop yield and income, resulting from water saving. Therefore, conservation of water resources would positively affect soil and groundwater quality. A large number of new techniques and methodologies are introduced by FAO-56 can be used in irrigation scheduling design. These include precisely estimated crop water requirements and crop evapotranspiration (ET_c) from climatic data.

Evapotranspiration (ET) is a process which includes loss of water from the plant as well as soil surface into the environment. Evapotranspiration assumes a paramount part in the hydrological cycle what's more it will be recognized a significant reason for water disaster around the universe. It depends upon different meteorological variables such as temperature, rainfall, wind speed, etc. The total amount of water through precipitation that soil receives, nearly 62% is lost through the process of evapotranspiration.

Monitoring and modeling of evapotranspiration rates have been the interest of many researchers. Various hydrological processes driving the hydrology of the reclaimed watershed can be simulated as a unique system, which is complicated considering the interrelationships among the various processes. By monitoring and modeling these processes, one can understand the evapotranspiration rate better and adopt more effective strategies in irrigation management and future reclamation designs.

Accurate assessment of evapotranspiration is of vital importance from different points of view, such as reliable quantification of hydrological water balance, hydrological design, water resource planning and management, irrigation system design and management, and crop yield simulation. In this study actual evapotranspiration, as an individual hydrological process is of interest to be modeled, estimated, and analyzed. The realization of the evapotranspiration process, which is obtained through an understanding of the temporal variations of AET (actual evapotranspiration) time series and the meteorological variables influencing the AET, can be considered as a step forward in the global aim of better understanding and management of irrigation scheduling. Water management has been repeatedly emphasizing on scientific irrigation scheduling.

The usability of ANNs (MLP and RBFN) and ε-SVR artificial intelligence methods is in the estimation of evaporation. In the development of ANN models, four different ANN algorithms GDX, LVM, SCG, and RBP were used in the MLP method. They considered different environmental factors (temperature, relative humidity, wind speed, and precipitation), which mainly affects the process of evaporation, as an input in their study. The pan evaporation values were used as an output. As a result of the evaluation of all obtained model performances, it was observed that all ANN models were more effective than ε-SVR and empirical Meyer and Romanenko methods [1].

Kisi O worked for estimation of evapotranspiration on monthly basis. The accuracy of LSSVM, MARS and M5Tree

models were compared with each other in estimating ET_0 by using air temperature, solar radiation, relative humidity and wind speed as inputs. Cross-validation method was used for each applied models by dividing data into four subsets. Different parameters were tried for each LSSVM model. Those parameters were considered which gave the minimum RMSE in the testing period [2].

The abilities of square-support vector regression (LS-SVR): fuzzy logic, ANN and ANFIS techniques are used to improve the accuracy of ET_0 estimation. In this study, the Gamma Test (GT) was used to estimate the noise variance among input to apply Machine Learning for best prediction. This model gave the nearest value as compared with the actual value. Regression was the best way to find the relationship between input and output on the basis of this study [3].

The irrigation management system works on the basis of short-term temperature and rainfall data. The old statistical model works on the basis of the monthly mean of ET_0. This model fails due to the rapid changing in the weather. In this study, they developed the numerical weather prediction model that worked more efficiently as compared to the old statistical model. This model showed the acceptable results on the observed scale [4].

The hybrid model (BD) consists of back propagation neural network and dynamic factor to estimate the pan evaporation. In this study, researchers tried their best to minimize the errors. But this model could not work well in all conditions. Under those circumstances, it did not prove a robust and dynamic model thereby; its results are not close to reality. The hybrid model gave significant results in generalization and estimation of ET_0 [5].

The hydrological cycle, the ET_0 is one of the main factors that depend upon the climate. In a study, the multi-layer perceptron network machine learning technique was used. A number of network model structures gave different results for ET_0. Machine learning algorithms used the limited meteorological data in this model to predict ET_0 [6].

The Pan Evaporation is one of the most famous methods to measure ET_0 but accuracy is not 100% in this method. By using the data of the sunshine, wind speed, relative humidity and temperature they developed an ANN Model for the prediction of ET_0. They used three-year data for training and one-year data for testing and validation of the model. The model consists of feedforward multilayer network with sigmoid as an activation function [7].

Several ANNs-based ET_0 (evapotranspiration) models, correspond to the best ranking conventional ET_0 estimation methods. They compared the results with FAO-56 PM ET_0 estimation model. The ANN models were consistent with the non-ideal condition of data availability and predicted ET_0 values with better closeness to the FAO-56 PM ET_0 than the conventional methods [8].

Abhishek Agarwal presented a progressive calculation method, which can be helpful for reducing the size of the hyperspectral information to constitutional dimensionality. In the progressive PCA, the data is divided into different parts, PCA was applied to each part independently and the outcomes were joined. The results of classification and lessened information via PCA were compared. The outcomes demonstrated that decreased information got by various level of PCA can contrast positively with the outcomes got from unique information. PCA gave comparative data content when contrasted with conventional PCA. The trials performed in his study utilized the maximum near normal PCA system [9].

The reliability of RBF-ANNs to estimate ET_0 uses three-calibrated temperature-based approach. Reference wheat crop evapotranspiration to estimate the utility of ANNs models were examined and it was found reasonable to predict ET_0. The ANN model proved effective in terms of accuracy by using minimum parameters for the estimation of ET_0 [10].

A methodology used was in view of Principal Component Analysis (PCA) for lessening the information size while saving the greater part of the data. PCA changed the information by separating accurately independent segments. This methodology offered a potentially useful procedure of tending to the issue of discarding and testing of ICs (integrated circuits) with an expansive number of test and estimation values. Lessening the information to an isolated measurement additionally encouraged simple representation and helped in major judgment [11].

In this study, the computational models are developed to estimate the ET_0. Computational models can deal with the complex system of ET_0 estimation and may also be used to determine the dependent variables. The contribution of the meteorological variables like maximum temperature, minimum temperature, average temperature, sun radiation, humidity, rainfall and wind speed to the ET_0 temporal variations is also of interest and examined using machine learning. We applied the Principle component analysis to reduce the data dimension and also to predict actual evapotranspiration. PCA is a technique which limits the total number of statistically independent parameters, to only those, which have more contribution towards the final output. The PCA technique is being used successfully for data dimension reduction procedures in the fields of agriculture and engineering. For developing the model, the ANNs were used. This piece of research helps us to reduce the computational time as well as the cost needed for the estimation of evapotranspiration.

II. Material and Methods

The climate data was observed in agriculture meteorological cell of the University of Agriculture Faisalabad. The observed location coordinates are 73.06^0 E, 31.25^0 N and an altitude of 184 meter above sea level. This area has general cropping pattern. The weather conditions of Faisalabad are semi-arid and it faces hot summer with maximum temperature of 50^0 C and a minimum temperature of -2^0 C in winter. The average maximum temperature of summer is 39^0C and minimum is 27^0C while in winter the maximum temperature is 17^0C and 6^0C is the minimum temperature. The average rainfall of the year is about 400-450mm. The half of the rainfall occurs in July and August. The data set consists of 4142 samples of following variables. In the proposed model, seven variables, i.e., maximum temperature, minimum temperature, average temperature,

wind speed, rainfall, solar radiation and relative humidity were used as inputs while ET_0 was used as the target variable.

PCA transforms the dataset into a new coordinate system. It places the variable of maximum variance at the first coordinate and the second maximum variable, with regard to variance, on the second coordinate system and so on. We applied principal component analysis in our experiment. We used MATLAB toolbox for that purpose. The following steps are involved in the process.

A. Standardize

The conversion of information into unit scale is a prerequisite for the ideal execution of many machine learning algorithms. So first of all, we convert all the elements in the dataset on a unit scale.

B. Calculate Covariance

Covariance (is a type of value used in statistics) describes the linear relationship between the two variables. More the covariance among two variables, the more closely their values follow the same trends over a range of data points. If the two variables are inclined to increase, it is positive covariance whereas the covariance will be negative for the case when one variable increases and other decreases. We measured the covariance between the meteorological variables and evapotranspiration rate. The formula for computing the covariance of the variables X and Y is

$$\sum n_i = 1(X_i - \overline{x})(Y_i - \overline{y})n - 1$$

$\overline{x}$ and $\overline{y}$ describe the means of X and Y respectively. In this way, we measured how evapotranspiration depends upon the metrological variables that were used as an input. The covariance among the variable are shown in Table 1.

TABLE I. COVARIANCE

Variables	Maxtem	Mintem	Avgtem	RH	RF	Radiation	WS	ET_0
Maxtem	65.8878	63.2234	64.6206	-80.3004	1.1395	11.7539	5.9663	14.9874
Mintem	63.2234	82.7533	73.0631	-52.4303	3.8843	8.3020	8.1418	14.4334
Avgtem	64.6206	73.0631	69.0977	-66.3082	2.5063	10.0406	7.0652	14.7291
RH	-80.3004	-52.4303	-66.3082	284.2012	19.1781	-23.9676	-5.8038	-23.7240
RF	1.1395	3.8843	2.5063	19.1781	31.7212	-2.8076	2.2386	0.4137
Radiation	11.7539	8.3020	10.0406	-23.9676	-2.8076	11.1336	0.3580	2.7802
WS	5.9663	8.1418	7.0652	-5.8038	2.2386	0.3580	6.9758	2.3811
ET_0	14.9874	14.4334	14.7291	-23.7240	0.1437	2.7802	2.3811	5.1610

C. Selecting Principal Components

PCA technique is commonly used for the reduction of dataset dimensions with the least loss of information where the whole dataset is projected on a new subspace. This method of projection is useful in order to reduce the computational costs and the error of parameter estimation. However, those eigenvectors best define the directions of the new axis, since they have all the same unit length.

The eigenvectors are dropped which have less useful information for the development of that lower-dimensional subspace. In this step, PCA reduced the data dimension on the basis of dependency. In our study, PCA reduced the seven meteorological variables to five variables on the basis of their importance. The eigenvalues can be found by the following relation:

$$\sum v = \lambda v$$

Where, Σ, v and λ represent the covariance matrix, eigenvector and eigenvalue respectively.

To solve for the eigenvalues, we use the determinant of the matrix to get a quadratic equation. The eigenvector with the largest eigenvalue is the direction of the greatest variation, the one with the the second largest eigenvalue is the (orthogonal) direction with the next highest variation and so on.

We have to decide which eigenvector(s) we need to drop to develop our lower-dimensional subspace. For that purpose, we examined the related eigenvalues of the eigenvectors. Approximately speaking, the eigenvectors with the least eigenvalues contains the minimum information about the distribution of the data, and the individuals would be those we need to drop. The basic methodology is to rank those eigenvalues from highest to lowest for the selection of top eigenvectors. Thus PCA marks the significant variables out of the large dataset leading to a reduced dataset.

D. Transforming the Samples into the New Subspace

In the last step, we used dimensional matrix W. It was computed to transform our samples on the new subspace via the equation given below. The new variables were used for measuring of evapotranspiration.

$$Y = W^T \times X$$

The new variables termed as the principal components are uncorrelated with each other and can be represented as a linear combination of the original variables. The process places the largest variance of the variables at the first position as the first principle component and the second largest variance of variables at the second position and so on in this similar fashion. In general, mostly the first few components are enough to provide the maximum information. Similarly, in our case, PCA gave five new transformed variables that we used for estimation of evapotranspiration rate. After reducing the dimension of data, we applied the time series neural (NAR) network modeling. Tan-sigmoid is the default transfer function in the hidden layer and the output layer has the linear transfer function. In NAR, there is only one series involved. The future values of a time series y(t) are predicted only from the past values of that series. This form of prediction is called nonlinear autoregressive, or NAR, and can be written as follows:

$$y(t) = f(y(t-1), \ldots\ldots, y(t-d))$$

The Dividend function is used for division of data for training, validation, and testing. Dividend separates the overall data into 70 percent for training, 15 percent for testing and 15 percent for validation. The model used the *trainlm* function for training. It gave faster results as compared to the other available functions [12].

III. Result and Discussions

A. Dimension Reduction

We applied the principal component analysis on the input variables (maximum temperature, minimum temperature, average temperature, rainfall, wind speed, relative humidity, solar radiation and evapotranspiration).

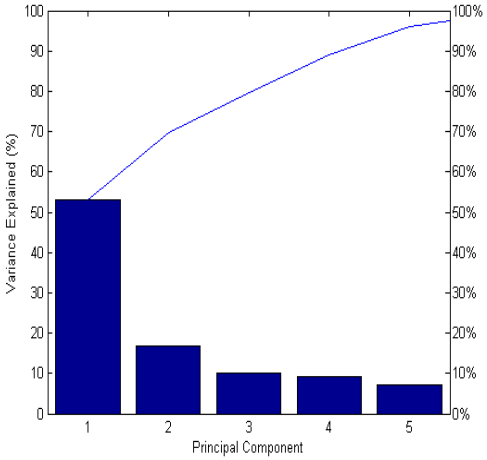

Fig. 1. Principal components and the respective variances.

Here Fig. 1 shows five principle components along x-axis and variance along the y-axis. The first component in the graph has more than 50% of the total variation in the dataset. That means that first component has higher and significant impact in the data set. Higher variation of the first component depicts the dependency of the evapotranspiration. Moreover, first five components showed total 95% of the variation in the dataset. So the other three variables have been discarded in the principal component analysis. In that case, the dimension of data sets has been reduced and falls to five elements.

Now, to evaluate the components values from dataset PCA generated a 3-dimensional graph. Fig. 2 shows PC_1 along x-axis and PC_2 along the y-axis. The dependency of variables can find out if its coefficient value is definable. In Fig. 3, the principal components along with the respective coefficient values are shown. It is obvious that the *Maxtemp* has a higher coefficient value among all the other components which is 0.48. That shows the significant contribution of the variable *Maxtemp* in the first principal component and this ultimately points towards its main role in defining the evapotranspiration. The other variables in the first component reflect their behavior from their respective coefficient values. For the 2nd principal component, the variable *is* has a value of 0.41 which is higher among all the other variables for the 2nd component.

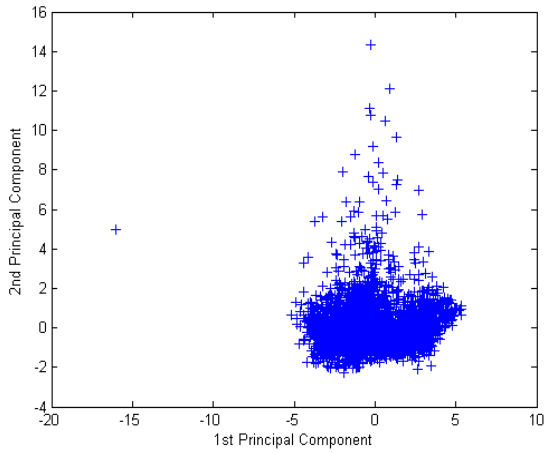

Fig. 2. He scaled data projected onto the first two principal components.

All the eight variables are symbolized in this bi-plot by a vector, and the direction and the length of the vector specify how each variable contributed to the dependent and independent variable in the plot (Fig. 3). The labeled diagram clearly defines the variable importance. The graph shows that the variables along x-axis have large data dispersion. This provides the identification of the major variation in those specific parameters. The analysis reveals evapotranspiration primarily depends upon the average temperature, minimum temperature, maximum temperature, rain fall and wind speed.

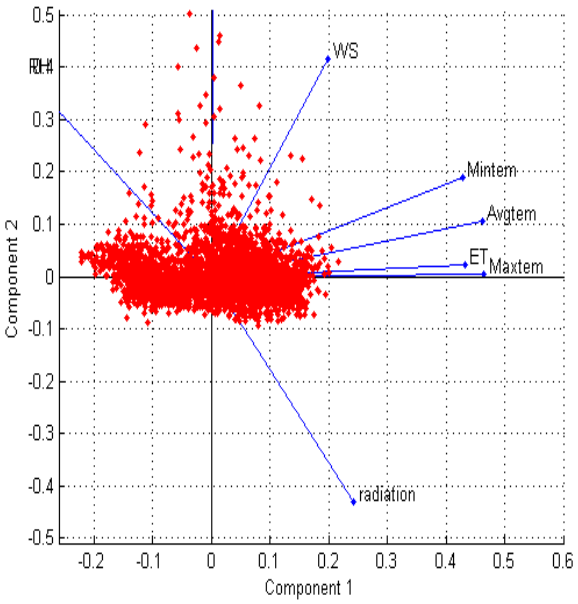

Fig. 3. PCA coefficients and PC scores.

B. Time Series Neural Network with reduced variables

Fig. 4 shows the architecture of the time series NN. We take five inputs; hidden layer activation function is log sigmoid, the delay is 2 and 10 neurons is used in the hidden layer. The activation function for the output layer is linear.

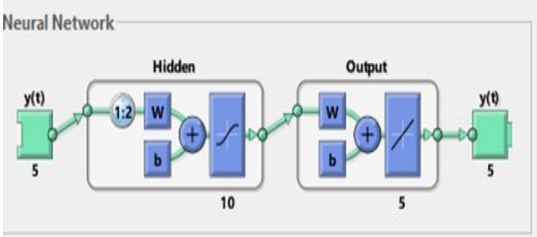

Fig. 4. Architecture of time series.

NNA after developing and training the model, the estimated regression between evapotranspiration and the reduced set of inputs is found as R=0.83426. The regression values show good fitting as given in Fig. 5.

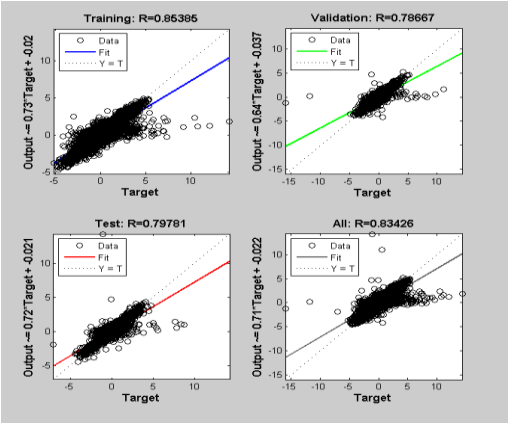

Fig. 5. Time series regression plot.

The model performance was best for the validation value 0.59693 at epoch number 9 which is shown below in Fig. 6. The blue, green and red lines indicate the performance of the model against the training data, validation data and the test data, respectively whereas the dotted line indicates the best situation.

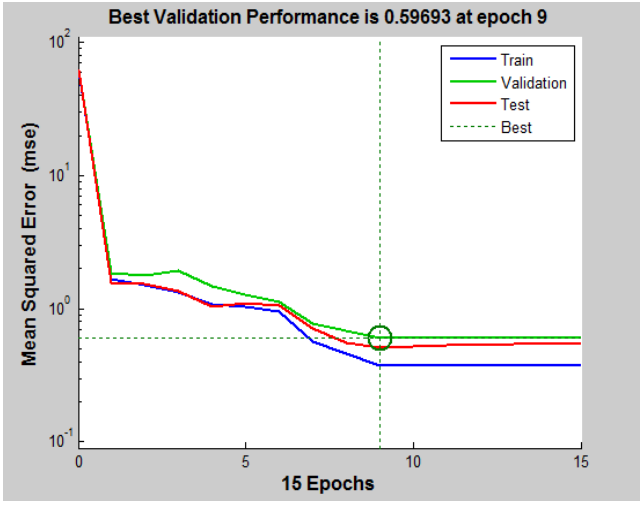

Fig. 6. Time series performance plot.

It has been observed that by applying PCA, we got the required results with greater accuracy. In this way, we reduced the computational time and power by using reduced and new variables provided by the PCA. The reduced variables generated almost the same results as we got considering all the variables, to measure evapotranspiration.

The regression line equation can be expressed according to our model as:

$$Y = m_1x_1 + m_2x_2 + m_3x_3 + m_4x_4 + m_5x_5 + b + \in$$

x_1, x_2, x_3, x_4 and x_5 are the main factors on which evapotranspiration rate depends basically. By using these variables, we estimate the evapotranspiration rate easily. The regression plot in Fig. 5 show that how much data are close to the line. Every point lies below or above the line with little distance called residuals. The residuals define the difference between the actual value and its value with respect to the regression line. It provides the useful information about the data. If there will be an association pattern underlying those data, it will appear in the residuals. Data that provides a good regression line has residuals that are haphazardly distributed on a residual plot.

IV. Conclusion

The measurement of evapotranspiration is the most critical and an important part of irrigation scheduling. It has observed that by using PCA, the new reduced variables gave the regression value of R=0.83426 in time series neural network. PCA is an effective method in reducing the data dimension and without loss of important information. Time series neural network provided better results as compared to other available methods. In this way, we can save computational time and cost. We can also measure evapotranspiration with greater accuracy. The time series neural network model predicted the evapotranspiration with an accuracy of 83% which is considerably higher than the other models.

References

[1] Tezel, Gulay, and Meral Buyukyildiz (2015). Monthly evaporation forecasting using artificial neural networks and support vector machines. *Theoretical and Applied Climatology* : 1-12.

[2] Kisi, Ozgur (2015). Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology* 528 : 312-320.

[3] Goyal, Manish Kumar, et al. (2014).Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. *Expert Systems with Applications* 41.11: 5267-5276.

[4] Perera, Kushan C., et al. (2014).Forecasting daily reference evapotranspiration for Australia using numerical weather prediction outputs. *Agricultural and Forest Meteorology* 194: 50-63.

[5] C,.Chang, F. J., W. Sun, and C. H. Chung (2013). Dynamic factor analysis and artificial neural network for estimating pan evaporation at multiple stations in northern Taiwan. *Hydrological Sciences Journal* 58.4: 813-825.

[6] Khoshhal, J., and M. Mokarram. (2012). Model for Prediction of Evapotranspiration Using MLP Neural Network. *Internetional Journal of Environmental Sciences* 3.3: 1000-1009.

[7] Ariapour, A., and Mojtaba Nassaji Zavareh (2011).Estimation of Daily Evaporation Using of Artificial Neural Networks (Case Study; Borujerd Meteorological Station). *Journal of Rangeland Science* 1.2 (2011).

[8] KuKumar, M., et al. (2008).Comparative study of conventional and artificial neural network-based ETo estimation models. *Irrigation Science* 26.6: 531-545.

[9] Abhishek Agarwal (2007). Efficient Hierarchical-PCA Dimension Reduction for Hyperspectral imagery.2007 IEEE International Symposium on Signal Processing and Information Technology.

[10] Trajkovic, Slavisa (2005). Temperature-based approaches for estimating reference evapotranspiration. *Journal of irrigation and drainage engineering* 131.4: 316-323.

[11] [Ashish S. Banthia, Anura P. Jayasumana, and Yashwant K. Malaiya (2005).Data size reduction for clustering-based binning of ICs using principal component analysis (PCA). *Current and Defect Based Testing, 2005. DBT 2005. Proceedings.2005 IEEE International Workshop on*. IEEE, 2005.

[12] [Hernández, Sergio, Luis Morales, and Philip Sallis (2011).Estimation of reference evapotranspiration using limited climatic data and Bayesian model averaging. *Computer Modeling and Simulation (EMS), Fifth UK Sim European Symposium on*. IEEE, 2011.