World Academy of Science, Engineering and Technology
International Journal of Geological and Environmental Engineering
Vol:10, No:6, 2016

Effects of Different Meteorological Variables on Reference Evapotranspiration Modeling: Application of Principal Component Analysis

Akinola Ikudayisi, Josiah Adeyemo

***Abstract*—**The correct estimation of reference evapotranspiration (ET_o) is required for effective irrigation water resources planning and management. However, there are some variables that must be considered while estimating and modeling ET_o. This study therefore determines the multivariate analysis of correlated variables involved in the estimation and modeling of ET_o at Vaalharts irrigation scheme (VIS) in South Africa using Principal Component Analysis (PCA) technique. Weather and meteorological data between 1994 and 2014 were obtained both from South African Weather Service (SAWS) and Agricultural Research Council (ARC) in South Africa for this study. Average monthly data of minimum and maximum temperature (°C), rainfall (mm), relative humidity (%), and wind speed (m/s) were the inputs to the PCA-based model, while ET_o is the output. PCA technique was adopted to extract the most important information from the dataset and also to analyze the relationship between the five variables and ET_o. This is to determine the most significant variables affecting ET_o estimation at VIS. From the model performances, two principal components with a variance of 82.7% were retained after the eigenvector extraction. The results of the two principal components were compared and the model output shows that minimum temperature, maximum temperature and windspeed are the most important variables in ET_o estimation and modeling at VIS. In order words, ET_o increases with temperature and windspeed. Other variables such as rainfall and relative humidity are less important and cannot be used to provide enough information about ET_o estimation at VIS. The outcome of this study has helped to reduce input variable dimensionality from five to the three most significant variables in ET_o modelling at VIS, South Africa.

***Keywords*—**Irrigation, principal component analysis, reference evapotranspiration, Vaalharts.

I. Introduction

WATER is the most crucial natural resource required for human survival, health and sustainable development. Water is also the scarcest natural resource on the earth because only 1% of the global water is available as freshwater [1]. The existence and survival of human being is solely dependent on water especially for domestic, industrial, energy and agricultural use [2]. South Africa falls within the semi – arid region, where the evaporation rate is more than the precipitation rates [3] and therefore, it is crucial to develop tools and models for the accurate estimation of water use and water availability.

Evapotranspiration (ET) has been described as the second most important component in the hydrologic cycle. It replaces the vapor lost to the atmosphere through condensation, thereby aiding the continuity of rainfall within the cycle [3]. ET is a very important component of hydrology, agriculture, meteorology and climatology because it is required for minerals and nutrient transport for plant growth [4]. The estimation of ET in the arid and semi-arid regions are very difficult because there are limited datasets of the variables that makes up ET. In many developing countries around the world, data are limited and scarce. Hence, there is a need to find the correlation between the variables in order to determine the most significant variables affecting the estimation and modeling of ET.

ET rate from a reference surface is called the reference ET and denoted by ET_o [5], [6]. Estimation of ET_o is vital to the sustainability of water resources management practices around the world. The Food and Agriculture Organisation (FAO) of the United Nations approved the Penman-Monteith (PM) equation, which is popularly called FAO-56 method, as one of the most accurate method for estimating ET_o [6]. This method has the capacity to calculate ET_o at different time steps as decided by the user. The FAO-56 method requires climatic variables such as sunshine hour, wind-speed, relative humidity, solar radiation, average temperature as inputs. A major limitation to the successful use of this FAO-56 equation in developing countries like South Africa is non-availability or limited data sets of these required variables. It is therefore important to develop simulation models as an alternative way of estimating ET_o. In the process of developing models for estimating ET_o, it is imperative to determine *a-priori* the correlation and relationship between the variables that makes up ET_o, hence, PCA is adopted in this study.

PCA is a powerful tool that has been widely used for the multivariate analysis of correlated variables [7]. PCA aims at extracting the most important information from the data set. Additionally, it is used to compress the size of the data set by keeping only the important information [8]. PCA rotates the original data space such that the axes of the new coordinate system point into the directions of highest variance of the data. The axes or new variables are termed principal components (PCs) and are ordered by variance. The first principal component (PC1) represents the direction of the highest variance of the data. The second principal component (PC2) accounts for most of the remaining variance under the

A. Ikudayisi (Doctoral Researcher) is with the Department of Civil Engineering & Surveying, Durban University of Technology, South Africa (phone: +27733598326; e-mail: otunbaakinola@yahoo.com).

J. Adeyemo (Professor) is with the Civil Engineering & Surveying Department, Durban University of Technology, South Africa. (e-mail: adeyemoja@gmail.com).

constraint to be orthogonal to the preceding component, PC1 [9].

PCA has been widely used in soil and water research to classify soils and water characteristics and variables [10]. PCA has been adopted by researchers to analyze correlated variables in irrigation schemes around the world. For example, PCA analysis was conducted by Visconti et al. [10] on thirteen chemical properties of soil saturation extracts in an irrigated Mediterranean area. A total of 139 soil samples extracted from 39 sites at Segura River lowland in Spain were analyzed. Three principal components with a variance of 76% were retained after the eigenvector extraction. PCA was adopted by Köksal [11] to analyze the relationship between crop growth level and water use status in an irrigated experimental field located in Turkey. The PCA analysis of smoothed spectral reflectance and first-order derivative spectra was conducted. Two principal components with a variance of about 99.9% were retained.

Biglari and Sutherland [12] presented a study on the use of PCA as a combustion model applied to a non-premixed temporally evolving jet flame with extinction and re-ignition. Jeong et al. [13] applied PCA in a study to determine the characteristics of polyphenolic contents of lettuce leaves grown under different night-time temperatures and cultivation durations up to 20 days using high performance liquid chromatography-tandem mass spectrometry.

The main objectives of this study are therefore: (1) to determine how the five weather parameters affect the estimation of ET_o at VIS, and (2) to identify the most significant variables for the estimation of ET_o at VIS. In other words, to reduce input variable dimensionality using the PCA.

II. Materials and Methods

A. Study Area

This study was carried out with meteorological and weather data collected from SAWS and ARC. This data was extracted from the meteorological stations at VIS. VIS is the largest irrigation scheme in South Africa and the entire world [14]. The scheme is located on a vast land area of about 370 km^2 and majorly used for irrigation. It is located in the Northern Cape Province of South Africa, which is the driest Province in the country. The scheme is supplied with water abstracted from the Vaal River at the Vaal Harts weir about 8 km upstream of Warrenton [15]. Fig. 1 shows the geographical location of the study area. This area is characterized by low, seasonal and irregular rainfall of about 442 mm per year [16].

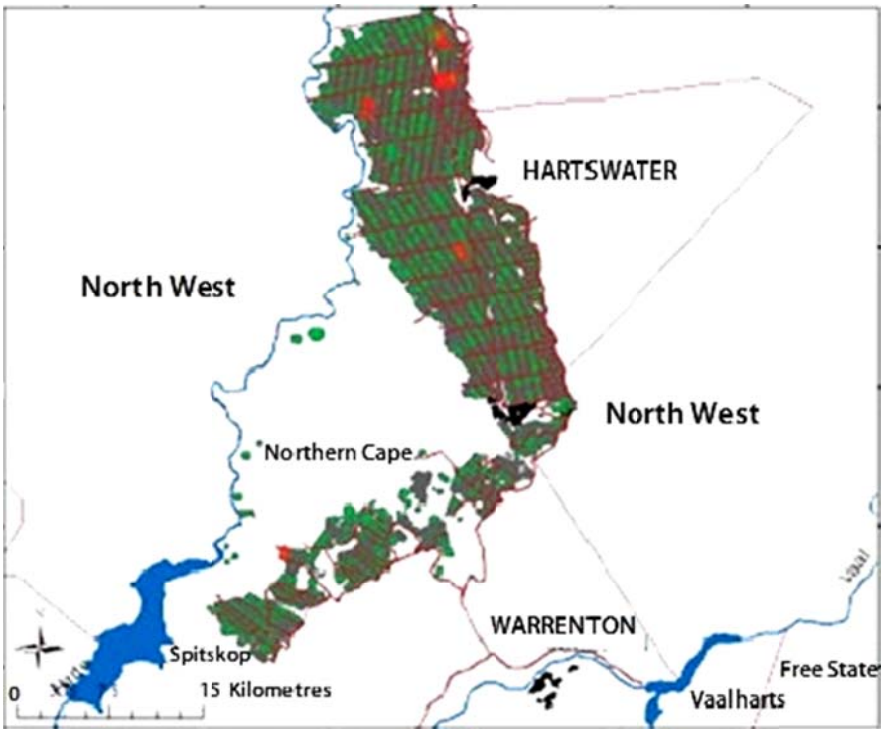

Fig. 1 Map of VIS [17]

B. PCA

Each principal component (PC) in PCA is the linear combination of the variables and gives a maximized variance. Let X be a matrix for n observations by p variables, and the covariance matrix is S. Then for a linear combination of the variables,

$$z_1 = \sum_{i=1}^{p} a1_i x_1 \quad (1)$$

where x_i is the ith variable, $a1_i x_1 i=1,2,...,p$ are linear combination coefficients for z_1, they can be denoted by a column vector a_1, and normalized by $a_1^T a_1=1$. The variance of z_1 will be $a_1^T S a_1$.The vector a_1 is found by maximizing the

variance. And z_1 is called the first principal component.

The second principal component can be found in the same way by maximizing $a_2^T S a_2$ subject to the constraints $a_2^T a_2=1$ and $a_2^T a_1=0$. This gives the second principal component which is orthogonal to the first one. Remaining principal components can be derived in a similar way. In fact coefficients $a_1, a_2, \ldots . a_p$ can be calculated from eigenvectors of the matrix S. Origin uses different methods according to the way of excluding missing values [18]

PCA shows the correlation structure of a data matrix X, approximating it by a matrix product of lower dimension (T × P'), called the principal components (PC), plus a matrix of residuals (E). This can be formulated in (2) below. The term $\left(1 \times \overline{x'}\right)$ represents the variable averages; the second term, the matrix product $\left(T \times P'\right)$, models the structure; and the third term, E, contains the deviations between the original values and the projections.

$$X = \left(1 \times \overline{x'}\right) + \left(T \times P'\right) + E \tag{2}$$

where, T is a matrix of scores that summarizes the X-variables (scores), and P is a matrix of loadings showing the influence of the variables on each score.

The correlation matrix is calculated from (3). After that, the eigenvectors and eigenvalues are estimated, and then the eigenvalues are sorted in descending order [19]. The eigenvector with the highest eigenvalue (PC1) is the most dominant principle component of the data set. The second component (PC2) is computed under the constraint of being orthogonal to PC1 and to have the second largest variance. The functions *pca* and *pcacov* in MATLAB R2009b was used to perform the PCA and to estimate the variable loadings.

$$r_{x,y} = \frac{\sum_{i=1}^{n}\left(x_i - \mu_x\right)\left(y_i - \mu_y\right)}{\left(n-1\right)\sigma_x \sigma_y} \tag{3}$$

where, μ_x and μ_y are the sample means of X and Y; σ_x and σ_y are the sample standard deviations of X and Y.

In implementing PCA using MATLAB, the following processes were used; the size of the data was determined; the mean and standard deviation was calculated; the dataset was standardized by subtracting the sample data from each observation and then dividing by the sample standard deviation; the coefficient of the principal component and their respective variances are done by finding the eigenfunctions of the sample covariance matrix. The MATLAB code executed is as follows:

```
function [signals,PC,V] = pca1(data)
% PCA1: Perform PCA using covariance.
% data - MxN matrix of input data
% (M dimensions, N trials)
% signals - MxN matrix of projected data
% PC - each column is a PC
% V - Mx1 matrix of variances
[M,N] = size(data);
% subtract off the mean for each dimension
mn = mean(data,2);
data = data - repmat(mn,1,N);
% calculate the covariance matrix
covariance = 1 / (N-1) * data * data';
% find the eigenvectors and eigenvalues
[PC, V] = eig(covariance);
% extract diagonal of matrix as vector
V = diag(V);
% sort the variances in decreasing order
[junk, rindices] = sort(-1*V);
V = V(rindices);
PC = PC(:,rindices);
% project the original data set
signals = PC' * data
```

PCA was performed on a correlation matrix of six variables in the system which are: Rainfall, minimum temperature, maximum temperature, relative humidity, wind speed and ET_o. Since the studied variables have different variances and units of measurements, the data set was standardized. This step was done by subtracting off the mean and dividing by the standard deviation. At the end of standardization process, each variable in the dataset is converted into a new variable with zero mean and unit standard deviation

III. Results and Discussions

In this study, PCA was performed on a correlation matrix of six variables in the system; those are: rainfall, minimum temperature, maximum temperature, relative humidity, wind speed and ET_o. Since the studied variables have different variances and units of measurements, the data set was standardized. This step was done by subtracting off the mean and dividing by the standard deviation. At the end of standardization process, each variable in the dataset is converted into a new variable with zero mean and unit standard deviation. The original and standardized variables are displayed in Figs. 2 and 3 respectively.

The correlation between a variable and a PC is known as "loading". Loadings close to ± 1 indicate that the factor strongly affects the measured variable. Components represented by the high loadings can be taken into consideration in evaluating the system. In this study, loadings having an absolute value > 0.40 were considered for grouping.

As listed in Table I, 82.67% of the information (variances) contained in the dataset were retained by the first two principal components (i.e. PC1 and PC2). However, each of the other remaining PCs has an eigenvalue lower than 1; thus only the first two PCs will be used in this study for interpretation.

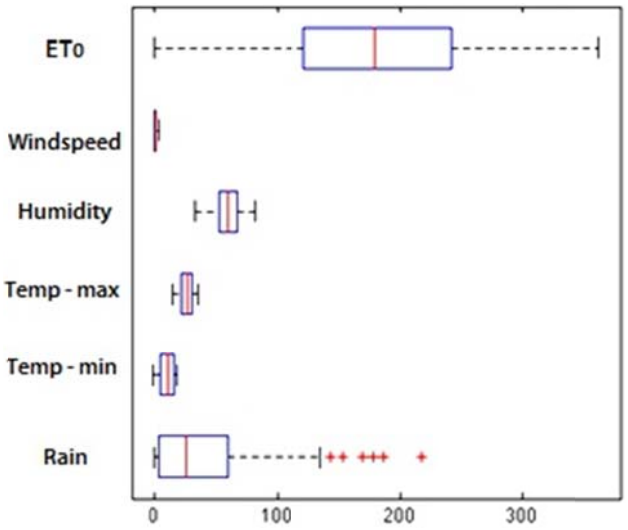

Fig. 2 Original data distribution of the variables

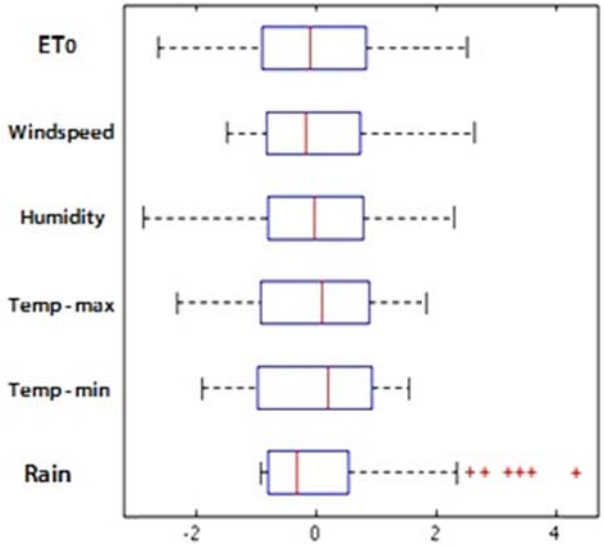

Fig. 3 Data standardization (normalization)

Since PC1 has the highest total variance of 63.53% (Table I), its parameters are the most important in estimating ET_o. The variables "minimum temperature", "maximum temperature", "wind speed" and "ET_o" have high loadings on PC1 with values of 0.47, 0.48, 0.43 and 0.47, respectively. Those high loading variables are more important than other parameters. This indicates that PC1 increases with an increase in minimum temperature, maximum temperature, wind speed and ET_o. Those parameters are on the right side of PC1 (Fig. 4). On the other side, rainfall and relative humidity have no role in explaining the variation in that PC since its absolute loading is lower than 0.4. Using the eigenvectors, the scores on PC1 can be computed as in (4).

$$PC1 = 0.25 \times \text{Rainfall} + 0.47 \times Temp_{min} + 0.48 \times Temp_{max} - 0.29 \times Humidity_{relative} + 0.43 \times \text{wind speed} + 0.47 \times ET_o \quad (4)$$

As listed in Table I, PC2 explains about 19.14% of the total variance, accounting for the next highest variance. It is strongly correlated with rainfall and relative humidity with heavy loadings of 0.70 and 0.61, respectively (Fig. 4). The scores on PC2 were estimated using the eigenvectors as in (5).

$$PC2 = 0.70 \times \text{Rainfall} + 0.28 \times Temp_{min} + 0.05 \times Temp_{max} + 0.61 \times Humidity_{relative} - 0.14 \times \text{wind speed} - 0.20 \times ET_o \quad (5)$$

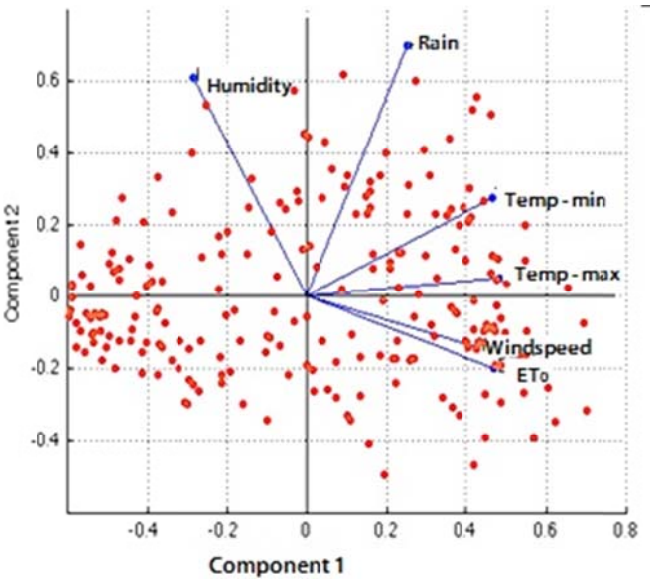

Fig. 4 PCA Loading plot of the dataset

Based on pre-screening using PCA, PC1 classified the measured data according to parameters that mostly affects ET_o.

TABLE I
LOADINGS FOR THE STUDIED VARIABLES

Variables	Loadings	
	PC1	PC2
Rainfall	0.25	0.70
Minimum temperature	0.47	0.28
Maximum temperature	0.48	0.05
Relative humidity	-0.29	0.61
Wind speed	0.43	-0.14
ET_o	0.47	-0.20
Eigenvalues	3.81	1.15
% variance	63.53	19.14
% Cumulative	63.53	82.67

IV. CONCLUSION

This study aims at determining the multivariate analysis of correlated variables involved in the estimation of ET_o at VIS in South Africa using PCA technique. Five different variables which are paramount to the estimation of ET_o were analyzed. From the PCA analysis, it was discovered that temperature and wind speed are the most important variable in the estimation of ET_o. Other variables such as rainfall and relative humidity have less significance on the value of ET_o.

Therefore, it can be concluded that PCA is a powerful tool in reducing input variable dimensionality into the most

significant ones before adequate and further simulation modeling operations takes place in a bit to estimate ET_o at VIS, South Africa.

ACKNOWLEDGMENT

The authors would like to acknowledge South African Weather Service and Agricultural Research Council for providing the dataset used for this study.

REFERENCES

[1] O. Olofintoye, Adeyemo, and F. Otieno, *Evolutionary algorithms and water resources optimisation*. Berlin: Springer Berlin Heidelberg, 2013.
[2] A. K. Mishra and V. P. Singh, "Drought modeling – A review," *Journal of Hydrology,* vol. 403, pp. 157-175, 6/6/ 2011.
[3] A. Ramoelo , N. Majozi , R. Mathieu , N. Jovanovic , A. Nickless , and S. Dzikiti "Validation of Global Evapotranspiration Product (MOD16) using Flux Tower Data in the African Savanna, South Africa," *Remote Sensing,* vol. 6, pp. 7406-7423, 2014.
[4] S. Traore, T. Kerh, and L. A. Gibson, "Modeling Reference Evapotranspiration by Generalized Regression Neural Network in Semiarid Zone of Africa," *WSEAS Transactions on Information Science & Applications,* vol. 6, pp. 991-1000, 2008.
[5] R. G. Allen, L. S. Pereira, D. Raes, and M. Smith, "Crop Evapotranspiration - Guidelines for Computing Crop Water Requirements," Food and Agriculture Organisation09-01-2010 1998.
[6] R. G. Allen, M. E. Jensen, J. L. Wright, and R. D. Burman, "Operational Estimates of Reference Evapo-transpiration," *Agronomy Journal,* vol. 81, pp. 650–662, 1989.
[7] D. Lee and P. Vanrolleghem, "Adaptive Consensus Principal Component Analysis for On-Line Batch Process Monitoring," *Environmental Monitoring and Assessment,* vol. 92, pp. 119-135, 2004.
[8] J. Costa, M. Alves, and E. Ferreira, "Principal component analysis and quantitative image analysis to predict effects of toxics in anaerobic granular sludge," *Bioresource Technology,* vol. 100, pp. 1180-1185, 2009.
[9] J. Lennox and C. Rosen, "Adaptive multiscale principal components analysis for online monitoring of wastewater treatment," *Water Science and Technology,* vol. 45, pp. 227-235, 2002.
[10] F. Visconti, J. M. de Paz, and J. L. Rubio, "Principal component analysis of chemical properties of soil saturation extracts from an irrigated Mediterranean area: Implications for calcite equilibrium in soil solutions," *Geoderma,* vol. 151, pp. 407-416, 7/15/ 2009.
[11] E. S. Köksal, "Hyperspectral reflectance data processing through cluster and principal component analysis for estimating irrigation and yield related indicators," *Agricultural Water Management,* vol. 98, pp. 1317-1328, 5/30/ 2011.
[12] A. Biglari and J. C. Sutherland, "An a-posteriori evaluation of principal component analysis-based models for turbulent combustion simulations," *Combustion and Flame,* vol. 162, pp. 4025-4035, 10// 2015.
[13] S. W. Jeong, G.-S. Kim, W. S. Lee, Y.-H. Kim, N. J. Kang, J. S. Jin, *et al.*, "The effects of different night-time temperatures and cultivation durations on the polyphenolic contents of lettuce: Application of principal component analysis," *Journal of Advanced Research,* vol. 6, pp. 493-499, 5// 2015.
[14] R. G. Ellington, "Quantification of the impact of the impact of irrigtion on the acquifer underlying the Vaalharts irrigation scheme," MSc, Institute of Groundwater Studies, University of Free State, 2003.
[15] O. I. Ojo, "Mapping and Modeling of Irrigation Induced Salinity of Vaal-Harts Irrigation Scheme in South Africa," DTech, Civil Engineering, Tshwane University of Technology, Pretoria, 2013.
[16] VIS, "Vaalharts Irrigation Scheme", ed: Wikimedia Foundation, Inc., 2013.
[17] O. O. Olofintoye, "Real Time Optimal Water Allocation in the Orange River Catchment in South Africa," DTech, Civil Engineering, Durban University of Technology, Durban, 2015.
[18] MATLAB, *PCA Toolbox for use with MATLAB*. USA, 2012.
[19] R. Tantra, C. Oksel, K. N. Robinson, A. Sikora, X. Z. Wang, and T. A. Wilkins, "A method for assessing nanomaterial dispersion quality based on principal component analysis of particle size distribution data," *Particuology,* vol. 22, pp. 30-38, 10// 2015.