Journal of Hydrology 508 (2014) 1–11

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Generalizability of Gene Expression Programming-based approaches for estimating daily reference evapotranspiration in coastal stations of Iran

Jalal Shiri [a,1,*], Ali Ashraf Sadraddini [a], Amir Hossein Nazemi [a], Ozgur Kisi [b], Gorka Landeras [c], Ahmad Fakheri Fard [a], Pau Marti [d]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[b] Civil Engineering Department, Architecture and Engineering Faculty, Canik Basari University, Samsun, Turkey
[c] NEIKER, AB, Basque Country Research Institute for Agricultural Development, Alava, Basque Country, Spain
[d] Departamento de Ingeniería Rural y Agroalimentaria, Universidad Politécnica de Valencia, c/Vera s/n, 46022 Valencia, Spain

ARTICLE INFO

Article history:
Received 23 July 2013
Received in revised form 21 October 2013
Accepted 22 October 2013
Available online 31 October 2013
This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Evapotranspiration
Exogenous data
Gene Expression Programming
Local training

SUMMARY

When dealing with climatic variables, the performance assessment of many Artificial Intelligence (AI) and/or data mining applications is based on a single data set assignment of the training and test sets. Further, it is very usual that this assignment is defined according to a local and temporary criterion, i.e. the models are trained and tested using data of the same station. Based on this procedure, the performance of the models outside the training location cannot be inferred. The present work evaluates the performance of Gene Expression Programming (GEP) based models for estimating reference evapotranspiration (ET_0) according to temporal and spatial criteria and data set scanning procedures in coastal environments of Iran. The accuracy differences between the local and the external performance depend on the specific climatic trends of the test stations, as well as on the input combination used to feed the models. When relying on a suitable input selection, externally trained models might be a valid alternative to locally trained ones, which would be a crucial advantage in places where only limited climatic variables are available. K-fold testing is a good choice to prevent partially valid conclusions derived from model assessments based on a simple data set assignment. Further, calibration of the GEP model may not be needed, if enough climatic data are available at other stations for external model application. The performance of the GEP model fluctuates chronologically and spatially. A suitable assessment of the model should consider a complete local and/or external scan of the data set used.

1. Introduction

Evapotranspiration (ET) can be quantified directly by relatively high cost aerodynamic as well as irradiative Bowen ratio methods or by utilization of lysimeters, based on a water balance in a controlled crop area (Allen et al., 1998). The term reference ET (ET_0) was introduced because the interdependence of the factors affecting the ET makes the study of the evaporative demand of the atmosphere difficult. In this way, the Penman–Monteith equation (FAO56-PM) has been adopted as a reference equation for estimating ET_0 and calibrating other equations (Allen et al., 1998). However, the need for large number of climatic variables (e.g. air temperature, relative humidity, solar radiation and wind speed) is a major disadvantage of the FAO56-PM model. Therefore, the development and validation of models relying on fewer climatic data is of critical importance for the regions where the measured climatic data are limited. In the last decades, the application of Artificial Intelligence (AI) techniques (e.g. Genetic Programming) for modeling agro-hydrologic parameters (e.g. ET) has been viable. Numerous studies have demonstrated that AI-based ET_0 estimation models are superior to traditional empirical and semi empirical ET_0 estimation models (e.g. Kisi et al., 2012c; Pour Ali Baba et al., 2013; Rahimi Khoob, 2008; Shiri and Kisi, 2011b; Shiri et al., 2012a, 2013a,b).

Genetic Programming (GP) was first proposed by Koza (1992) and is particularly suitable where: (a) the interrelationships among relevant variables are poorly understood; (b) finding the optimum solution is hard; (c) conventional mathematical analysis does not, or cannot, provide analytical solutions; (d) an approximate solution is acceptable; (e) small improvements in the performance are routinely measured (or easily measurable) and highly valued; and (f) there is a large amount of data, in computer readable form, that requires examination, classification, and integration (Banzhaf et al., 1998).

* Corresponding author. Tel.: +98 4113340081.
E-mail address: j_shiri2005@yahoo.com (J. Shiri).
[1] Ph.D. Student.

0022-1694/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2013.10.034

GEP (Gene Expression Programming) is comparable to GP but involves computer programs of different sizes and shapes encoded in linear chromosomes of fixed lengths. The most important advantages of GEP are (Ferreira, 2001): (i) the chromosomes are simple entities: linear, compact, relatively small, easy to manipulate genetically (replicate, mutate, recombine, etc.); (ii) the expression trees are exclusively the expression of their respective chromosomes; they are entities upon which selection acts, and according to fitness, they are selected to reproduce with modification.

Notable applications of GP (i.e. GEP) in modeling water resources systems have been reported in the literature, including e.g. predicting velocity in compound channels (Harris et al., 2003); determination of chezy resistance factor (Giustolisi, 2004); determining the unit hydrograph of the urban basins (Rabunal et al., 2007); modeling flow and water quality variables in watersheds (Preis and Otsfeld, 2008); predicting groundwater table fluctuations (Shiri and Kisi, 2011a; Shiri et al., 2013c); river flow prediction (Shiri et al., 2012b); modeling daily precipitation (Kisi and Shiri, 2011); modeling river suspended sediment load (Kisi and Shiri, 2012; Kisi et al., 2012a); modeling daily lake level fluctuations (Kisi et al., 2012b); estimating daily incoming solar radiation (Landeras et al., 2012), modeling daily dewpoint temperature (Shiri et al., 2013d), and modeling rainfall-runoff procedure (e.g. Aytek and Alp, 2008; Kisi et al., 2013). Nonetheless, some few studies have been reported in literature including GP application for modeling evaporation/evapotranspiration. Parasuraman et al. (2007) applied GP for modeling the dynamics of ET. Guven et al. (2008) used GEP for modeling ET_0 in USA. Guven and Kisi (2010) investigated linear genetic programming (LGP) and ANN applications to model daily pan evaporation. Izadifar and Elshorbagy (2010) compared ANN, GEP and statistical models for estimating hourly actual ET. Kisi and Guven (2010) used linear genetic programming for modeling ET. Shiri and Kisi (2011b) compared GEP, ANFIS and ANNs to estimate daily pan evaporation values using recorded and estimated weather variables. Shiri et al. (2012a) applied GEP for modeling daily reference evapotranspiration with a local (individual station) as well as pooled (the whole region) approaches.

Commonly, many AI and GP based applications consider only a single data set assignment, as well as, exclusively, a temporary and local management of the data sets, i.e. models are trained and tested using data of the same station. Apart from not performing a suitable and complete performance assessment of the local patterns, another important limitation of this approach is that the generalizability of the developed models is not assessed outside the training station. This is decisive to evaluate the real usefulness of many published procedures, especially those presenting an accurate performance of locally trained models relying on limited inputs. Although requiring few inputs for their application, those models might only be useful in the training stations, unless the external generalizability is also validated, which is not the case in most applications, as mentioned. If these models are only accurate in the training stations, their real applicability is limited to local emergency cases, like breakdowns in the data acquisition system. A new user would not be able to apply that model in a different station, because the external performance was not evaluated, and would require a suitable set of patterns, including the targets, for training a new local model relying on that limited combination of inputs. In most cases, calculated FAO56-PM ET_0 targets are used, due to the usual absence of experimental ones. So, enough inputs would be required for a new user to calculate first the needed targets according to FAO56-PM. Hence, the studies enhancing the usefulness of models relying of limited inputs fail often in the evaluation of their performance and might provide misleading conclusions about their real applicability. Only few studies have tried to assess the external performance of ET_0 models (Kisi, 2007; Kisi et al., 2012c; Martí et al., 2010, 2011; Rahimi Khoob, 2008; Shiri et al., 2011, 2013a,b). Nevertheless, these studies considered only a single data set assignment. Shiri et al. (2013e) performed for the first time an external assessment of the generalizability of GEP based models for estimating pan evaporation based on k-fold testing. The current study aims at applying a similar approach to estimate ET_0 in a different climatic scenario, namely several coastal locations of Iran.

2. Methodology

2.1. Studied region and used data

Eight coastal weather stations from Iran were considered in this study. The geographical positions of the studied weather stations are shown in Fig. 1. The used dataset comprises daily values of maximum air temperature (T_{max}), minimum air temperature (T_{min}), mean air temperature (T_{mean}), wind speed (W_S), relative humidity (R_H) and solar radiation (R_S) between the 1st of January 2000 and the 31st December 2008. Table 1 sums up the average and standard deviation values of the used weather data in the studied stations. In the present study the aridity index (I_A) (UNEP, 1992), a numerical indicator of the degree of dryness of the climate, and the Currey continentality index (CI^{CU}) were applied. These indicators were selected for their simplicity.

$$I_A = \frac{P}{ET_P} \tag{1}$$

$$CI^{CU} = \frac{M_i - m_i}{1 + \frac{\theta}{3}} \tag{2}$$

where ET_P is annual potential evapotranspiration (mm); P (mm) is the average annual precipitation (UNEP, 1992), M_i is the maximum monthly average temperature (°C); m_i is the minimum monthly average temperature (°C); and θ is the latitude (°). To be consistent ET_P and P are expressed in the same units. Potential ET ($ET_P = ET_0$) [which is supposed to be the same as reference evapotranspiration] was calculated using the standard FAO56 Penman Monteith method (Allen et al., 1998). According to the "World Atlas of Desertification" (UNEP, 1992, 1997), dry lands have an aridity index of less than 0.65 and precipitation of less than 600 mm per year. Fig. 2 represents the I_A and CI^{CU} values of the studied stations.

2.2. Gene Expression Programming (GEP)

In the present work the GeneXpro program was used for modeling daily ET_0. The application of the GEP procedure involves the following steps. In the first step the fitness function must be defined. Based on the results obtained by Shiri et al. (2012a), applying the root mean square error (*RMSE*) fitness function produces the most accurate results in modeling ET_0 values. The second step consists of choosing the set of terminals T and the set of functions F to create the chromosomes. In the present study, the terminal set includes the weather variables. The choice of the appropriate function is not so obvious and depends on the experience and intuition of the user. The appropriate functions for modeling ET_0 were selected based on Shiri et al. (2012a) [i.e. $\{+, -, \times, \div, \sqrt[3]{}, \sqrt{}, \ln, e^x, x^2, x^3, \sin x, \cos x, Arctgx\}$]. The length of head was $h = 8$, while three genes per chromosome were employed, which are commonly used values in literature (e.g. Ferreira, 2001). The fourth step is the choice of the linking function. The linking function must be chosen as "addition" or "multiplication" for algebraic sub trees (Ferreira, 2001). In general, the choice of the linking function depends on the problem and there is not any basic rule to identify which of these functions is more suitable. Here, addition linking functions were applied according to Shiri et al. (2012a). The fifth and final step is to choose