Empirical and learning machine approaches to estimating reference evapotranspiration based on temperature data

Matheus Mendes Reis⁎, Ariovaldo José da Silva, Jurandir Zullo Junior, Leonardo David Tuffi Santos, Alcinei Místico Azevedo, Érika Manuela Gonçalves Lopes

School of Agricultural Engineering, University of Campinas, Avenida Cândido Rondon, 501, Barão Geraldo, 13083-875 Campinas, SP, Brazil

Institute of Agrarian Sciences, Federal University of Minas Gerais, Avenida Universitária, 1000, Bairro Universitário, 39404-547 Montes Claros, MG, Brazil

ARTICLE INFO

Keywords:
Soft computing
Artificial neural networks
Multiple linear regression
Extreme learning machines
Artificial intelligence
Meteorological data

ABSTRACT

The precise estimation of reference evapotranspiration (ET₀) is crucial for the planning and management of water resources and agricultural production. In this study, the applicability of the Hargreaves Samani (HS), artificial neural network (ANN), multiple linear regression (MLR) and extreme learning machine (ELM) models were evaluated to estimate ET₀ based on temperature data from the Verde Grande River basin, southeastern Brazil. These models were evaluated in two scenarios: local and pooled. In the local scenario, training, calibration and validation of the models were performed separately at each station. In the pooled scenario, meteorological data from all stations were grouped for training and calibration and then separately tested at each station. The ET₀ values estimated by the Penman-Monteith model (FAO-56 PM) were considered the target data. All the developed models were evaluated by cluster analysis and the following performance indices: relative root mean square error (RRMSE), Pearson correlation coefficient (r) and Nash-Sutcliffe coefficient (NS). In both scenarios evaluated, local and pooled, the results revealed the superiority of the artificial intelligence methods (ANN and ELM) and the MLR model compared to the original and adjusted HS models. In the local scenario, the ANN (with r of 0.751, NS of 0.687 and RRMSE of 0.112), ELM (with r of 0.747, NS of 0.672 and RRMSE of 0.116) and MLR (with r of 0.743, NS of 0.665 and RRMSE of 0.068) models presented the best performance, in addition to being grouped in the same cluster. Similar to the observations from the local scenario, the ANN (with r of 0.718, NS of 0.555 and RRMSE of 0.165), ELM (with r of 0.724, NS of 0.601 and RRMSE of 0.151) and MLR (with r of 0.731, NS of 0.550 and RRMSE of 0.091) models presented the best performance in the pooled scenario and were grouped in the same cluster. The locally trained models presented higher precision than the models generated in the pooled scenario; however, the models generated in the pooled scenario could be used to estimate ET₀ in cases of unavailability of local meteorological data. Although the MLR, ANN and ELM models, based on temperature data, are appropriate alternatives to accurately estimate ET₀ in the Verde Grande River basin, southeastern Brazil, the MLR model presents the advantage of the use of explicit algebraic equations, facilitating its application.

1. Introduction

The reference evapotranspiration (ET₀), introduced by the Food and Agriculture Organization of the United Nations (FAO) as a methodology for computing crop evapotranspiration (Doorenbos and Pruitt, 1977), is an essential component in irrigation planning, river basin hydrology, and hydrological balance studies (Antonopoulos and Antonopoulos, 2017; Traore et al., 2010). In addition, ET₀ is a key element in executing effective water management practices and optimizing their use in agricultural production areas (Smith, 2000). In the case of arid and semi-arid regions, understanding ET₀ is even more important for efficient irrigation planning (Huo et al., 2012).

ET₀ can be determined by lysimeters (Anapalli et al., 2016; Xu et al., 2018), the energy balance (Yan et al., 2017), scintillometers (Valayamkunnath et al., 2018), or by using empirical equations based on meteorological data (Antonopoulos and Antonopoulos, 2017). The Penman-Monteith (PM) method is universally recommended by the FAO as the only precise equation to calculate ET₀ (Allen et al., 1998). The PM model incorporates thermodynamic and aerodynamic aspects, and it has been shown to be relatively accurate in humid and arid conditions. However, the use of PM requires detailed information on wind speed, aerodynamic resistance, and ground cover, which may not always be available.
regions (Smith et al., 1991; Yin et al., 2008). However, the greatest disadvantage of the PM method is the need for several types of climatic data that are not always available (Cobaner et al., 2017; Fan et al., 2018; Feng et al., 2017; Gocic et al., 2015).

In the Brazilian semi-arid region, the distribution and density of the meteorological stations are inadequate or insufficient, limiting the use of the PM method in irrigation management in this region, which has the largest public irrigated perimeter in Latin America and other important irrigation perimeters.

The Hargreaves Samani (HS) method is an alternative to the PM method. In regions with low availability of meteorological data, such as the semi-arid region, where the Verde Grande River basin is located, the low spatial density of meteorological stations in the region hampers the accurate estimation of ET0. The development of improved methods to estimate the amount of water required by the crops is essential to improve the accuracy of the irrigation level and to increase the efficiency of water use.[1]

The Verde Grande River basin is one of the main irrigated fruit growing regions in Brazil and stands out because it has the largest irrigated perimeter in Latin America; however, the low availability of meteorological stations in the region hampers the accurate estimation of ET0. The development of improved methods to estimate the amount of water required by the crops is essential to improve the accuracy of the irrigation level and to increase the efficiency of water use (Antonopoulos and Antonopoulos, 2017), especially in areas with water scarcity problems, such as the semi-arid region, where the Verde Grande River basin is located. The low spatial density of meteorological stations in the Verde Grande River basin and its poor distribution make it impossible to apply the PM equation in irrigation management in regions with low availability of meteorological data.

Table 1

<table>
<thead>
<tr>
<th>Station (longitude; latitude and altitude)</th>
<th>Station code</th>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esponosa (42.8°W, 14.9°S and 569.6 m)</td>
<td>1</td>
<td>Pr (mm)</td>
<td>0.0</td>
<td>126.7</td>
<td>1.9</td>
<td>7.7</td>
<td>411.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RH (%)</td>
<td>22.5</td>
<td>97.7</td>
<td>63.4</td>
<td>13.7</td>
<td>21.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmin (°C)</td>
<td>9.6</td>
<td>27.7</td>
<td>20.0</td>
<td>2.6</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmax (°C)</td>
<td>20.2</td>
<td>40.0</td>
<td>31.3</td>
<td>2.8</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>U2 (m s−1)</td>
<td>0.0</td>
<td>6.4</td>
<td>2.2</td>
<td>1.1</td>
<td>49.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rs (MJ m−2)</td>
<td>7.5</td>
<td>76.6</td>
<td>21.9</td>
<td>4.9</td>
<td>22.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ET0 (mm d−1)</td>
<td>1.5</td>
<td>11.5</td>
<td>4.9</td>
<td>1.3</td>
<td>26.3</td>
</tr>
<tr>
<td>Januá (43.3°W, 15.8°S and 516.0 m)</td>
<td>2</td>
<td>Pr (mm)</td>
<td>0.0</td>
<td>157.0</td>
<td>2.0</td>
<td>8.1</td>
<td>402.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RH (%)</td>
<td>21.0</td>
<td>97.8</td>
<td>59.1</td>
<td>13.9</td>
<td>23.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmin (°C)</td>
<td>7.8</td>
<td>29.0</td>
<td>19.3</td>
<td>2.5</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmax (°C)</td>
<td>20.2</td>
<td>41.8</td>
<td>32.1</td>
<td>2.9</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>U2 (m s−1)</td>
<td>0.0</td>
<td>6.5</td>
<td>1.1</td>
<td>1.0</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rs (MJ m−2)</td>
<td>7.3</td>
<td>36.5</td>
<td>21.4</td>
<td>4.5</td>
<td>20.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ET0 (mm d−1)</td>
<td>1.6</td>
<td>10.0</td>
<td>4.3</td>
<td>1.1</td>
<td>24.6</td>
</tr>
<tr>
<td>Juramento (43.7°W, 16.8°S and 650.0 m)</td>
<td>3</td>
<td>Pr (mm)</td>
<td>0.0</td>
<td>127.0</td>
<td>2.4</td>
<td>8.5</td>
<td>348.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RH (%)</td>
<td>24.0</td>
<td>99.8</td>
<td>69.6</td>
<td>12.5</td>
<td>17.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmin (°C)</td>
<td>4.2</td>
<td>25.3</td>
<td>16.8</td>
<td>3.2</td>
<td>19.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmax (°C)</td>
<td>19.7</td>
<td>40.0</td>
<td>29.9</td>
<td>2.8</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>U2 (m s−1)</td>
<td>0.0</td>
<td>6.5</td>
<td>1.1</td>
<td>0.5</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rs (MJ m−2)</td>
<td>7.0</td>
<td>29.9</td>
<td>20.6</td>
<td>4.6</td>
<td>22.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ET0 (mm d−1)</td>
<td>1.3</td>
<td>7.8</td>
<td>3.9</td>
<td>1.0</td>
<td>25.0</td>
</tr>
<tr>
<td>Montes Claros (43.8 °W, 16.7 °S and 646.3 m)</td>
<td>4</td>
<td>Pr (mm)</td>
<td>0.0</td>
<td>145.0</td>
<td>2.7</td>
<td>9.3</td>
<td>344.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RH (%)</td>
<td>24.0</td>
<td>98.3</td>
<td>62.7</td>
<td>14.6</td>
<td>23.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmin (°C)</td>
<td>5.8</td>
<td>28.0</td>
<td>18.0</td>
<td>3.0</td>
<td>16.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmax (°C)</td>
<td>20.0</td>
<td>40.3</td>
<td>30.2</td>
<td>2.9</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>U2 (m s−1)</td>
<td>0.0</td>
<td>4.0</td>
<td>1.4</td>
<td>0.5</td>
<td>33.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rs (MJ m−2)</td>
<td>7.1</td>
<td>30.1</td>
<td>20.9</td>
<td>5.0</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ET0 (mm d−1)</td>
<td>1.4</td>
<td>8.5</td>
<td>4.2</td>
<td>1.1</td>
<td>26.3</td>
</tr>
<tr>
<td>Monte Azul (42.8 °W; 15.1 °S and 603.6 m)</td>
<td>5</td>
<td>Pr (mm)</td>
<td>0.0</td>
<td>140.1</td>
<td>2.1</td>
<td>7.8</td>
<td>381.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RH (%)</td>
<td>21.0</td>
<td>98.8</td>
<td>58.6</td>
<td>14.3</td>
<td>24.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmin (°C)</td>
<td>11.3</td>
<td>28.0</td>
<td>20.4</td>
<td>2.2</td>
<td>11.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tmax (°C)</td>
<td>20.7</td>
<td>41.1</td>
<td>31.3</td>
<td>2.9</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>U2 (m s−1)</td>
<td>0.0</td>
<td>6.7</td>
<td>1.7</td>
<td>1.1</td>
<td>60.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rs (MJ m−2)</td>
<td>6.4</td>
<td>29.7</td>
<td>21.4</td>
<td>4.6</td>
<td>21.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ET0 (mm d−1)</td>
<td>1.4</td>
<td>9.4</td>
<td>4.8</td>
<td>1.2</td>
<td>24.9</td>
</tr>
</tbody>
</table>

Pr – precipitation; RH – relative humidity; Tmin – minimum temperature; Tmax – maximum temperature; U2 – wind speed; Rs – solar radiation; ET0 – reference evapotranspiration by Penman-Monteith (Allen et al., 1998); Min – minimum; Max – maximum; SD – standard deviation; CV (%) – coefficient of variation.
(ET₀) only with air temperature data as the input, thus allowing the accurate determination of ET₀ in locations without available relative air humidity, solar radiation and wind speed data. To achieve this purpose, Hargreaves Samani (HS), adjusted HS, artificial neural network “multilayer perceptron” (ANN), multiple linear regression (MLR) and extreme learning machine (ELM) were compared to the Penman-Monteith method (FAO 56 – PM). These models were evaluated in two scenarios: local and pooled. In the local scenario, all models were trained, calibred and validated separately at each station. In the pooled scenario, the meteorological data from all stations were grouped for training and calibration of the models and then tested separately at each station. The generation of models by a combined approach allows the estimation of ET₀ in cases of unavailability of local meteorological data, that is, in regions that are distant from meteorological stations, but with the availability of temperature data.

2. Material and methods

2.1. Study area and data set

The Verde Grande River basin has an area of approximately 31,410 km² and a population of 741,500 inhabitants. The Verde Grande River basin stands out in the world scenario for housing the largest irrigated perimeter in Latin America, with the irrigation projects Jaiiba, Gorutuba, Lagoa Grande and Estreito, and an irrigated area of approximately 742 km². The study area is located in the semi-arid region of Brazil. The climate of the region is classified, according to Köppen, as Aw, warm tropical with a dry winter. The meteorological stations and statistical properties of the climatic variables are shown in Table 1.

The daily meteorological variables (maximum (Tmax) and minimum (Tmin) air temperature at a height of 2 m, mean relative humidity (RH), wind speed at a height of 10 m (U₁₀) and sunshine duration) were obtained from the five meteorological stations of the Instituto Nacional de Meteorologia (INMET) located within the Verde Grande River basin (Fig. 1) between 1996 and 2016. The meteorological data provided by INMET are of satisfactory quality certified by ISO 9001: 2008. (INMET, 2019). The missing data for Tmax and Tmin were reconstituted by the linear interpolation method, and the missing data for RH, U₂ and Rs were estimated according to the methodology proposed by Allen et al. (1998).

2.2. FAO-56. Penman-Monteith model

The FAO-recommended Penman-Monteith equation (FAO 56-PM) (Equation (1)) (Allen et al., 1998) was used to estimate ET₀ data, which were used as the targets for the calibration and evaluation of the HS, adjusted HS, MLR, ANN and, ELM models. This process is an accepted and commonly used practice (Antonopoulos and Antonopoulos, 2017; Didari and Ahmadi, 2019; Dou and Yang, 2018).

$$\text{ET}_0 = \frac{0.408 \Delta (R_n - G) + \frac{900}{273 + T} u_2 (e_s - e_a)}{\Delta + \gamma (1 + 0.34 u_2)}$$  (1)

where ET₀ is the reference evapotranspiration (mm dia⁻¹), Rn is the net radiation (MJ m⁻² day⁻¹), G is the soil heat flux density (MJ m⁻² day⁻¹), T is the mean daily air temperature at a height of 2 m (°C), U₂ is the wind speed at a height of 2 m (m s⁻¹), es is the saturated vapor pressure (kPa), ea is the actual vapor pressure (kPa), Δ is the slope vapor pressure curve (kPa °C⁻¹), and γ is the psychrometric constant (kPa °C⁻¹).

Due to the lack of U₂ and Rs data, these two parameters were estimated (Eqs. (2) and (3)) based on the data for sunshine duration and U₁₀, respectively (Allen et al., 1998).

$$U_2 = U_{10} \frac{4.87}{\ln 67.8z - 5.42}$$  (2)

$$R_s = \left( a_s + b_s \frac{n}{N} \right) R_s$$  (3)
where $U_{10}$ is the wind speed at a height of 10 m ($\text{m s}^{-1}$), $z$ is the height measurement (10 m), $R_s$ is the solar or shortwave radiation (MJ m$^{-2}$ day$^{-1}$), $n$ is the sunshine duration (h), $N$ is the maximum possible sunshine or daylight duration (h), $R_a$ is extraterrestrial radiation (MJ m$^{-2}$ day$^{-1}$), and $a_s$ and $b_s$ are constants with a value of 0.28 and 0.52, respectively, as recommended by FAO-56 (Allen et al., 1998).

### 2.3. Hargreaves Samani model

The Hargreaves Samani (HS) model (Eq. (4)) was initially proposed by Hargreaves and Samani (1985) and requires only air temperature data to estimate $E_T$.

$$E_T = 0.0023 R_a (T_{\text{max}} - T_{\text{min}})^{0.25} (T + 17.8)$$

where $T_{\text{max}}$ and $T_{\text{min}}$ are the maximum and minimum air temperatures ($^\circ\text{C}$), respectively.

The extraterrestrial radiation data ($R_a$) were calculated based on latitude data (Equation (5)).

$$R_a = \frac{24(60)}{\pi} G_{\text{SC}} d_r \left[ \sin(\varphi) \sin(\delta) + \cos(\varphi) \cos(\delta) \cos(\omega_S) \right]$$

where $R_a$ is the extraterrestrial radiation (MJ m$^{-2}$ day$^{-1}$), $G_{\text{SC}}$ is the solar constant (0.0820 MJ m$^{-2}$ min$^{-1}$), $d_r$ is the inverse relative Earth-Sun distance, $\omega_S$ is the sunset hour angle (rad), $\varphi$ is the latitude (rad) and $\delta$ is the solar declination (rad).

The adjusted HS model was obtained by regression (Eq. (6)), which is an accepted and commonly used practice (Droogers and Allen, 2002; Feng et al., 2017b; Shiri et al., 2014).

$$E_T^{\text{HM}} = a + b E_T^{\text{HS}}$$

where $E_T^{\text{PM}}$ is $E_T$ estimated by FAO-56 PM, $E_T^{\text{HS}}$ is $E_T$ estimated by

<table>
<thead>
<tr>
<th>Models</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Z_i = a + b x_i + c y_i + e_i$</td>
</tr>
<tr>
<td>2</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + e_i$</td>
</tr>
<tr>
<td>3</td>
<td>$Z_i = a + b x_i + c y_i^2 + d y_i + e_i$</td>
</tr>
<tr>
<td>4</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + e_i$</td>
</tr>
<tr>
<td>5</td>
<td>$Z_i = a + b x_i + c y_i + d x_i y_i + e_i$</td>
</tr>
<tr>
<td>6</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + f x_i y_i + e_i$</td>
</tr>
<tr>
<td>7</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + f x_i y_i + e_i$</td>
</tr>
<tr>
<td>8</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + f x_i y_i + g x_i y_i + e_i$</td>
</tr>
<tr>
<td>9</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + f x_i y_i + g x_i y_i + h x_i y_i^2 + e_i$</td>
</tr>
<tr>
<td>10</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + f x_i y_i + g x_i y_i + h x_i y_i^2 + e_i$</td>
</tr>
<tr>
<td>11</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + f x_i y_i + g x_i y_i + h x_i y_i^2 + e_i$</td>
</tr>
<tr>
<td>12</td>
<td>$Z_i = a + b x_i + c x_i^2 + d y_i + f x_i y_i + g x_i y_i + h x_i y_i^2 + e_i$</td>
</tr>
</tbody>
</table>

Table 2 Multiple regression models tested to estimate $E_T$ ($Z_i$) as a function of maximum ($x_i$) and minimum ($y_i$) temperatures.

Fig. 2. Workflow of the $E_T$ estimation in this study. MLR – multiple linear regression; ELM – extreme learning machine; HS – Hargreaves Samani; ANN – artificial neural Network “multilayer perceptron”; RRMSE - relative root mean square error; $r$ - Pearson correlation coefficient; NS – Nash-Sutcliffe coefficient.
HS, and a and b are regression coefficients.

2.4. Artificial neural networks

Artificial Neural Networks (ANNs) are mathematical models that are analogous to biological neural networks, which have been applied in many studies to model ET₀ (Dou and Yang, 2018; Traore et al., 2010).

Multilayer “perceptron” (MLP) networks have been used with “backpropagation” learning. For the development of MLP networks, the mlp function of the RSNN package in the R software was used, with the backpropagation algorithm (learnFunc = “Std_Backpropagation”) and a learning rate of 0.1 (learnFuncParams = 0.1). The networks were composed of an input layer with two neurons, corresponding to the number of input variables (Tmax and Tmin). In the output layer, a neuron corresponding to ET₀ was introduced. To define the number of neurons in the hidden layer, a trial and error procedure was used. For this procedure, 1000 networks were tested with a number of neurons in the middle layer ranging from 1 to 10. The most appropriate model for the studied five cities was the one that presented eight neurons in the hidden layer, presenting smaller estimates of the mean square error.

The common sigmoid and linear activation functions were used for the hidden and output layers, respectively. The number of training times was arbitrated as 500.

2.5. Multiple linear regression

To estimate ET₀ from the maximum and minimum temperature, regression was also used. For this procedure, 12 regression models were tested (Table 2), and the quality of the fit was evaluated by estimating the Akaike information criterion (AIC). To adjust the regression models, we used the lm function in software R.

2.6. Extreme learning machine

The ELM learning machine technique was initially proposed by Huang et al. (2006), and it has been applied in several studies on ET₀ estimation (Dou and Yang, 2018; Feng et al., 2017b; Gocic et al., 2016) and in other research areas (Cao et al., 2016; Duan et al., 2016). ELM has an extremely fast learning speed in comparison to other learning machine techniques such as ANN. In addition, the number of neurons in the hidden layer for a specific ELM model does not need to be obtained.
The ELM model consists of an input layer with two neurons (Tmax and Tmin), a single hidden layer with six neurons and the output layer (ET0). For the generation of ELM models, the elmtrain function of the RSNNS package in R software was used. The normalized data for ET0 obtained by the ANN and ELM models were transformed into mm day⁻¹ unit using Eq. (8).

\[ V_{ds} = \frac{V_{max} + (V_{0} - 1)(V_{max} - V_{min})}{V_{max} - V_{min}} \]  

(8)

where \( V_{ds} \) is the denormalized value (dimensionless), \( V_{0} \) is the normalized value, \( V_{min} \) is the minimum value of the sample, and \( V_{max} \) is the maximum value of the sample.

The ANN and ELM models were tested 1000 times, storing their respective mean squared errors (MSE) for the validation data set. The MSE for the validation sample was determined using Eq. (9).

\[ \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)^2 \]  

(9)

where \( n \) is the total number of data tested, and \( P_i \) and \( O_i \) are the ET0 values obtained by the FAO-56 PM method and by the artificial intelligence methods (ANN and ELM), respectively.

The best fit network for the ANN and ELM models was established by the smaller MSE for the validation sample, with the objective to ensure the absence of overfitting.

2.8. Performance criteria

The relative root mean square error (RRMSE) (Eq. (10)), Pearson correlation coefficient \( r \) (Eq. (11)) and Nash-Sutcliffe coefficient (NS) (Eq. (12)) were used to evaluate the performance of the HS, adjusted HS, MLR, and ELM models.

\[ \text{RRMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)^2} \]  

(10)

\[ r = \frac{\sum_{i=1}^{n} (P_i - \bar{P})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{n} (P_i - \bar{P})^2 \sum_{i=1}^{n} (O_i - \bar{O})^2}} \]  

(11)

\[ \text{NS} = 1 - \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} \]  

(12)
Fig. 6. Pearson correlation coefficient (r) (A), Nash-Sutcliffe (NS) coefficient (B) and relative root mean square error (RRMSE) (C) of the adjusted HS, MLR, ANN and ELM models generated with pooled data.

Table 5
Algebraic equations generated by multiple linear regression (MLR) and Hargreaves Samani (HS) models, adjusted to estimate $ET_0$ in the pooled scenario.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Algebraic equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled</td>
<td>$ET_0 = -3.12^{<strong>} + 0.03771^{</strong>} T_{max} + 0.00263^{<strong>} T_{max}^2 + 0.958^{</strong><em>} T_{min} - 0.07869^{</em><strong>} T_{min}^2 - 0.01941^{</strong>*}$</td>
</tr>
<tr>
<td></td>
<td>$T_{max} T_{min} + 0.003502^{<em><strong>} T_{max} (T_{min})^2 - 0.00003909^{</strong></em>} (T_{max})^2 (T_{min})^2$</td>
</tr>
</tbody>
</table>

$r$ – Pearson correlation coefficient; NS – Nash-Sutcliffe coefficient; RRMSE – root mean square error; $ET_0$ – reference evapotranspiration; HS – Hargreaves Samani; MLR – multiple linear regression; ANN – artificial neural networks; ELM – extreme learning machine; PM – Penman-Monteith.

Fig. 7. Monthly variation of $ET_0$ values estimated by PM, HS, adjusted HS, MLR, ANN and ELM models generated with pooled data (A), and under/overestimation values of HS, adjusted HS, MLR, ANN and ELM models in relation to the PM model (B).

Table 6
Global average performance and annual average of $ET_0$ values estimated by adjusted HS, MLR, ANN and ELM models generated with pooled data.

<table>
<thead>
<tr>
<th>Model</th>
<th>r</th>
<th>NS</th>
<th>RRMSE</th>
<th>$ET_0$ (mm)</th>
<th>Under/Overestimation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted HS</td>
<td>0.661</td>
<td>0.305</td>
<td>0.171</td>
<td>1610.1</td>
<td>-1.1</td>
</tr>
<tr>
<td>MLR</td>
<td>0.731</td>
<td>0.550</td>
<td>0.091</td>
<td>1606.7</td>
<td>-1.2</td>
</tr>
<tr>
<td>ANN</td>
<td>0.718</td>
<td>0.555</td>
<td>0.165</td>
<td>1626.4</td>
<td>-0.2</td>
</tr>
<tr>
<td>ELM</td>
<td>0.724</td>
<td>0.601</td>
<td>0.151</td>
<td>1585.1</td>
<td>-2.7</td>
</tr>
<tr>
<td>PM</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>1627.5</td>
<td>/</td>
</tr>
</tbody>
</table>

$r$ – Pearson correlation coefficient; NS – Nash-Sutcliffe coefficient; RRMSE – relative root mean square error; $ET_0$ – reference evapotranspiration; HS – Hargreaves Samani; MLR – multiple linear regression; ANN – artificial neural networks; ELM – extreme learning machine; PM – Penman-Monteith.

Tmax – maximum temperature; Tmin – minimum temperature; R² – coefficient of determination; ns – not significant by Student’s t test; and *, **, *** Significant at levels of 0.05, 0.01 and 0.001 by Student’s t test, respectively.
where n is the total number of data tested, and \( P_i \) and \( O_i \) are the \( ET_0 \) values obtained by the FAO-56 PM method and by the other methods (HS, adjusted HS, MLR, ANN and ELM), respectively. RRMSE is dimensionless and presents the perfect fit with a result of 0. The NS and \( r \) coefficients are dimensionless and have the perfect fit with a result of 1.

2.9. Multivariate analysis

A multivariate analysis of the data set was performed using cluster analysis (CA). The \( ET_0 \) data for the HS, adjusted HS, MLR, ANN, and ELM models were previously standardized to avoid misclassification due to size differences of the data. Subsequently, standardized Euclidean distances as similarity measure were estimated. CA was performed in the data set by the Ward method (Ward, 1963).

2.10. Data processing

Fig. 2 shows the workflow of the \( ET_0 \) estimation in this study.

3. Results and discussion

3.1. Local implementation of the models

The MLR, ANN and ELM models, based on air temperature data, presented similar performance criteria (RRMSE, \( r \) and NS), with \( r \) varying between 0.962 and 0.806 (Fig. 3A), NS ranging from 0.926 to 0.648 (Fig. 3B), and RRMSE ranging from 0.171 to 0.072 (Fig. 3C). The MLR model showed the lowest RRMSE in all cities. In the cities of Espinosa (station 1), Janaúba (station 2), Juramento (station 3) and Montes Claros (station 4), the \( r \) and NS indexes were higher in the ELM model. The MLR, ANN and ELM models, adjusted for the city of Montes Claros (station 4), presented the best performance criteria (RRMSE, \( r \) and NS) compared to the other cities (Fig. 3). The algebraic equations that best estimated \( ET_0 \) for the MLR and adjusted HS models in the local scenario are presented in Table 3.

The ANN and ELM models were more accurate than the original and adjusted HS empirical models, according to RRMSE, \( r \) and NS performance criteria. The original HS model had higher accuracy in the city of Montes Claros (station 4), with \( r \), NS and RRMSE of 0.881, 0.626 and 0.151, respectively. In contrast, the adjusted HS model showed higher precision in the city of Juramento (station 3), with \( r \), NS and RRMSE of 0.875, 0.763 and 0.104, respectively. Regarding the original HS model, the cities of Janaúba (station 2), Juramento (station 3) and Monte Azul (station 5) presented the least accurate results, with NS of −1.162, RRMSE of 0.240 and \( r \) of 0.713, respectively, while the adjusted HS model had lower accuracy in Monte Azul (station 5), with \( r \), NS and RRMSE of 0.710, 0.503 and 0.149, respectively. The use of regression made the \( ET_0 \) estimate more accurate by the adjusted HS model compared to the original HS equation (Fig. 3).

The ANN, ELM and MLR models presented the best performance coefficients (Fig. 3 and Table 3), with \( r \) varying between 0.934 and 0.795 (Fig. 6A). The MLR model had better performance in the cities of Montes Claros (station 4), with \( r \), NS and RRMSE of 0.934, 0.858, and 0.096, respectively. The ELM model performed better in the cities of Janaúba (station 2), Juramento (station 3) and Monte Azul (station 5), with \( r \) of 0.811 and NS of 0.616 (Fig. 6). The adjusted HS model had the worst performance in all cities, with \( r \) ranging from 0.899 to 0.698 (Fig. 6A), NS ranging from 0.706 to 0.070 (Fig. 6B), and RRMSE ranging from 0.234 to 0.188 (Fig. 6C). The algebraic equations that best estimated \( ET_0 \) for the MLR and adjusted HS models, in the pooled data scenario, are presented in Table 5.

Similar to the local scenario, the MLR, ANN and ELM models generated with the pooled temperature data from the five stations showed similar performance, with \( r \) varying between 0.934 and 0.795 (Fig. 6A), NS ranging from 0.858 to 0.415 (Fig. 6B), and RRMSE ranging from 0.223 to 0.094 (Fig. 6C). The MLR model had better performance in the cities of Espinosa (station 1), with \( r \) of 0.900, NS of 0.725 and RRMSE of 0.120), Janaúba (station 2), Juramento (station 3) and Montes Claros (station 4), with \( r \) of 0.934, NS of 0.858 and RRMSE of 0.096), while the ELM model performed better in the cities of Juramento (station 3, with \( r \) of 0.934 and NS of 0.853) and Monte Azul (station 5, with \( r \) of 0.811 and NS of 0.616) (Fig. 6). The adjusted HS model had the worst performance in all cities, with \( r \) ranging from 0.899 to 0.698 (Fig. 6A), NS ranging from 0.706 to 0.070 (Fig. 6B), and RRMSE ranging from 0.234 to 0.188 (Fig. 6C). The algebraic equations that best estimated \( ET_0 \) for the MLR and adjusted HS models, in the pooled data scenario, are presented in Table 5.

Similar to the local scenario, MLR and ELM models underestimated \( ET_0 \) during most of the year, showing monthly mean values that were higher than the PM model (FAO 56 – PM) only in March, May and June, respectively (Fig. 4 and Table 4).
when the MLR model was analyzed, and March, when the ELM model was analyzed. The ANN model underestimated ET₀ during the dry and cold period of the year (May to September), while the adjusted HS model overestimated ET₀ in January, February, November, and December. The adjusted HS, MLR, ANN and ELM models presented underestimated and overestimated values ranging from −8.74 to 5.91%, −5.19 and 1.11%, −6.65 and 4.20%, and −7.73 and 1.64%, respectively (Fig. 7).

Table 6 shows the global average performance of the models generated with pooled data, considering the joint evaluation of the five stations analyzed during the study period. The adjusted HS, MLR, ANN and ELM models underestimated ET₀ by 1.1, 1.2 0.2, and 2.7%, respectively (Table 6). The MLR (with r of 0.731 and RMSE of 0.091) and ELM (with NS of 0.601) models presented the best global performance, although the ANN model provided annual mean ET₀ values closer to the reference model (FAO 56 – PM) (Table 6). The adjusted HS model had the worst global performance, with r of 0.661, NS of 0.305 and RMSE of 0.171 (Table 6).

The MLR, adjusted HS, ANN and ELM models were grouped into two clusters (Fig. 8). Cluster 1 was formed by the adjusted HS, the model that presented the worst performance coefficients (Fig 6 and Table 6). Cluster 2 was formed by the MLR, ANN and ELM models (Fig. 8).

Similar to the observations in the local scenario, the MLR, ANN and ELM models had the best performance coefficients (Fig 6 and Table 6) and were grouped in the same cluster (Fig. 8) in the pooled data scenario. These results indicate the best applicability of ANN and ELM artificial intelligence methods and the MLR model to estimate ET₀, using only temperature data.

The artificial intelligence methods (e.g., ANN and ELM) usually perform better than linear models (e.g., MLR), because they are efficient for modeling phenomena of linear and nonlinear nature, as well as considering possible interactions between variables. The MLR model used in this research presented similar results to the models based on artificial intelligence (ANN and ELM), as observed in other studies (Huo et al., 2012; Ribeiro et al., 2019; Tabari et al., 2012). This can be justified by the fact that the linear, quadratic and interaction effects were considered in some of the MLR models.

The locally trained models were more accurate than the models with pooled temperature data, which is consistent with the data reported by Shiri et al. (2014) in Iran and contrasts with the work of Peng et al. (2017b) in China. However, the generation of models using a combined approach allows the estimation of ET₀ in cases of unavailability of local meteorological data.

The present study confirmed the ability of the ANN, ELM and MLR models to estimate ET₀ only with air temperature data as input. Compared with the ANN model based on MLP, the ELM demonstrated much faster performance when executing the training and testing processes in R software. The MLR models were easier to apply because they provide explicit algebraic equations for the calculation of ET₀ (Table 3 and 5).

Therefore, ANN, ELM and MLR models can be used in irrigation planning and management, thus allowing more precise estimation of crop water requirements through air temperature data; however, the MLR model presents the advantage of the use of explicit algebraic equations, which facilitates their use by farmers and technicians to accurately estimate ET₀.

4. Conclusions

In this study, the applicability of the Hargreaves Samani (HS), adjusted HS, artificial neural network (ANN), multiple linear regression (MLR) and extreme learning machine (ELM) models was evaluated to estimate the reference evapotranspiration (ET₀) based on temperature data from the Verde Grande River basin, southeastern Brazil. This study was carried out in two parts. In the first part (local scenario), all the models were trained, calibrated and validated separately at each station. In the second part (pooled scenario), the meteorological data from all the stations were grouped to train and calibrate the models and then tested separately at each station. To test and validate the models, the Penman-Monteith model (FAO-56 PM) for ET₀ estimation was considered as a reference. The main conclusions can be summarized as follows.

(1) The ANN, MLR and ELM models showed a great ability to estimate ET₀; moreover, they were more accurate than the original and adjusted HS models, both in the local and pooled scenarios.

(2) The ANN, MLR and ELM models had very similar accuracies, but the MLR models were easier to use, as they calculated the ET₀ using explicit algebraic equations.

(3) Although the ANN, MLR and ELM models showed better adjustment in the local scenario, in the pooled scenario, these models allowed the accurate estimation of ET₀ in cases of unavailability of local meteorological data.

(4) The use of the linear regression model to fit the HS equation increased the accuracy of this model to estimate ET₀.

Acknowledgments

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for providing scholarships to the first author and financial support. Espaço da Escrita – Pró-Reitoria de Pesquisa – UNICAMP – for providing the language services.

References


