Theoretical and Applied Climatology (2020) 142:1481–1492
https://doi.org/10.1007/s00704-020-03380-4

ORIGINAL PAPER

Performance evaluation of numerical and machine learning methods in estimating reference evapotranspiration in a Brazilian agricultural frontier

Diego Bispo dos Santos Farias[1] · Daniel Althoff[1] · Lineu Neiva Rodrigues[1,2] · Roberto Filgueiras[1]

Received: 25 March 2020 / Accepted: 13 September 2020 / Published online: 24 September 2020

Abstract
The reference evapotranspiration (ET_0) estimates is important for water resources and irrigation management. The Penman-Monteith equation is known for its accuracy but requires a high number of climatic parameters that are not always available. Thus, this study aimed to evaluate the performance of machine learning techniques (cubist regression, artificial neural network with Bayesian regularization, support vector machine with linear kernel function) and stepwise multiple linear regression method to estimate daily ET_0 with limited weather data in a Brazilian agricultural frontier (MATOPIBA). Climatic data from 2000 to 2016 obtained from 23 weather stations were used. Five data scenarios were evaluated: (i) all variables, (ii) radiation and temperature, (iii) temperature and relative humidity, (iv) wind speed and temperature, and (v) temperature. The results showed that the machine learning methods are robust in estimating ET_0, even in the absence of some variables. Among the methods evaluated using only temperature data, the cubist regression showed better performance. When estimating water demand for soybean and maize crops using only temperature, the cubist regression and calibrated Hargreaves-Samani equation showed the smallest errors.

1 Introduction

The intensification of agriculture, that is, increasing production per unit of planted area combined with the reduction of environmental impacts, is the most appropriate strategy to increase food production in a sustainable manner (Pradhan et al. 2015). The intensification of agriculture, in turn, will increasingly depend on irrigation, which is the main user of water resources in Brazil and worldwide (ANA 2017; FAO 2015).

Increasing the irrigated area may intensify conflicts over the use of water, especially in hydrographic basins where there is already a compromised water availability. In order to have water security in those basins, it is important that water is used in a sustainable manner. Therefore, it is necessary to improve the management and efficiency of use of irrigation water (Fishman et al. 2015)

Obtaining reliable estimates of crop evapotranspiration (ETc) is essential for the development of irrigation management strategies. In addition, these estimates for remote rural areas, with little information, which are prevalent in Brazil, are of special interest for water resources management.

The Penman-Monteith model (FAO-56) has been used as standard for estimating reference evapotranspiration (ET_0), which serves as a basis for estimating ETc using empirical/semi-empirical methods (Allen et al. 1998; López-Urrea et al. 2006; Stöckle et al. 2004). The application of the Penman-Monteith model, however, has been hampered by the set of information necessary for its execution (Doorenbos and Pruitt 1977; Allen et al. 1998). The lack of input data required by this model has hindered its application several regions of Brazil. For instance, Althoff et al. (2019) highlight a large variation of weather station density among different biomes in Brazil. Thus, it is important to evaluate other techniques that allow estimations in conditions of limited data.

In recent years, alternative methods such as machine learning have been studied to estimate ET_0 (Ferreira and da Cunha

✉ Diego Bispo dos Santos Farias
diego.farias@ufv.br

[1] Department of Agricultural Engineering, Federal University of Viçosa (UFV), Av. Peter Henry Rolfs, s.n, Viçosa, Minas Gerais 36570-900, Brazil

[2] Brazilian Agricultural Research Corporation, Embrapa Cerrados, BR-020, Km 18, Planaltina, DF 73310-970, Brazil

2020; Wu and Fan 2019; Keshtegar et al. 2018; Wen et al. 2015). These methods aim to estimate ET_0 based on techniques and methods that require a small number of variables and, consequently, are less costly. Althoff et al. (2018) evaluated models to estimate the ET_0 in the mesoregions of Northwest of Minas and Triângulo Mineiro/Alto Paranaíba, in the Minas Gerais State of Brazil, and concluded that machine learning methods perform well in ET_0 prediction even when limited weather input data is used. Ferreira et al. (2019) evaluated machine learning algorithms for ET_0 estimation across the entire Brazilian territory. The authors only used temperature and relative humidity data and obtained results close to those estimated by the Penman-Monteith model.

Many studies have been carried out comparing the reference evapotranspiration calculated from heuristic methodologies with the reference evapotranspiration calculated by the Penman-Monteith model (Shiri et al. 2014; Kisi and Alizamir 2018; Wu and Fan 2019; Seifi and Riahi 2018). However, there is a lack of studies evaluating the impact of the results of the ET_0 simulations on the water demand of the crop, which is an important analysis for the decision-making of which method is most appropriate to be used.

Considering that the evapotranspiration estimation is important for irrigation management in agricultural areas, the objective of the present study was to (i) evaluate the performance of machine learning techniques in estimating ET_0 in the MATOPIBA region, the latest agricultural frontier in the Brazil, and (ii) assess the impact of ET_0 estimates on water demand for maize and soybean crops, two crops of great interest for the MATOPIBA region.

2 Materials and methods

2.1 Study area and data set

The study region, MATOPIBA, includes a range of areas in the states of Maranhão, Tocantins, Piauí, and Bahia and is one of the largest grain producers in Brazil (Silva et al. 2018). Most of the agricultural production in this region is in the Cerrado biome, which contains about 78% and 64% of all the center pivots and all of the irrigated area in Brazil, respectively (Althoff and Rodrigues 2019; Sparovek et al. 2014). The MATOPIBA territory comprises 324,326 agricultural establishments (de Miranda et al. 2014), which make it complex in terms of water resources management.

To evaluate the models, 17 years (2000–2016) of daily weather data from 23 weather stations was used (Fig. 1). The following climatic data were used: average air temperature (T_{mean}, °C), maximum (T_{max}, °C) and minimum (T_{min}, °C) temperatures, relative humidity (RH, %), wind speed (WS, m s^{-1}), and sunshine duration (hours), which was converted to solar radiation (SR, MJ m^{-2} day^{-1}) using the methodology presented by Allen et al. (1998). The data were obtained from the Meteorological Database for Teaching and Research (BDMEP), made available by the National Institute of Meteorology (INMET) of Brazil.

The INMET's standard conventional weather stations are from the manufacturer R Fuess. The equipment's sensitivity for temperature, relative humidity, and wind speed readings are 0.2 °C, 5%, and 0.1 m s^{-1}, respectively. Solar radiation was estimated from the number of hours of sunshine, following the FAO-56 methodology. Days with missing data were discarded for modeling.

2.2 Reference evapotranspiration

Table 1 shows the equations used to calculate the reference evapotranspiration. Extraterrestrial radiation was calculated based on the methodology presented by Allen et al. (1998) and used in models that did not use solar radiation. For the purpose of evaluating the performance of the equations, the Penman-Monteith method was used as a reference, hereinafter referred to as standard reference evapotranspiration ($ET_{0\text{-}PM}$). The equations evaluated in this study (Hargreaves and Samani 1985; Makkink 1957; Priestley and Taylor 1972), presentes in Table 1, had their empirical coefficients calibrated for the study region. For this, the Levenberg-Marquardt algorithm (Moré 1978) was used, which minimizes the sum of the residual squares. To evaluate the performance of machine learning models in estimating ET_0, different combinations of climatic variables were used (Table 2).

2.3 Models developed for the estimation of ET_0

2.3.1 Stepwise regression

Multiple linear regression was obtained using the stepwise (SW) method. The SW provides a linear equation where only significant independent variables are present (Abraham et al. 2017). For this, the independent variables were added and removed one by one from the regression set. At each stage, the performance of the model was evaluated to make sure which variables had a minimum level of significance ($\alpha <$ 5%). The final equation was obtained when no variable available to be added or no variable could be discarded without loss of performance (Wang et al. 2011).

2.3.2 Machine learning models

The machine learning models used were cubist regression (CB), the artificial neural network (NN) with Bayesian regularization, and the support vector (SV) machine with linear kernel function. To develop the models, the language and environment for statistical computing R (R Core Team 2018) and the

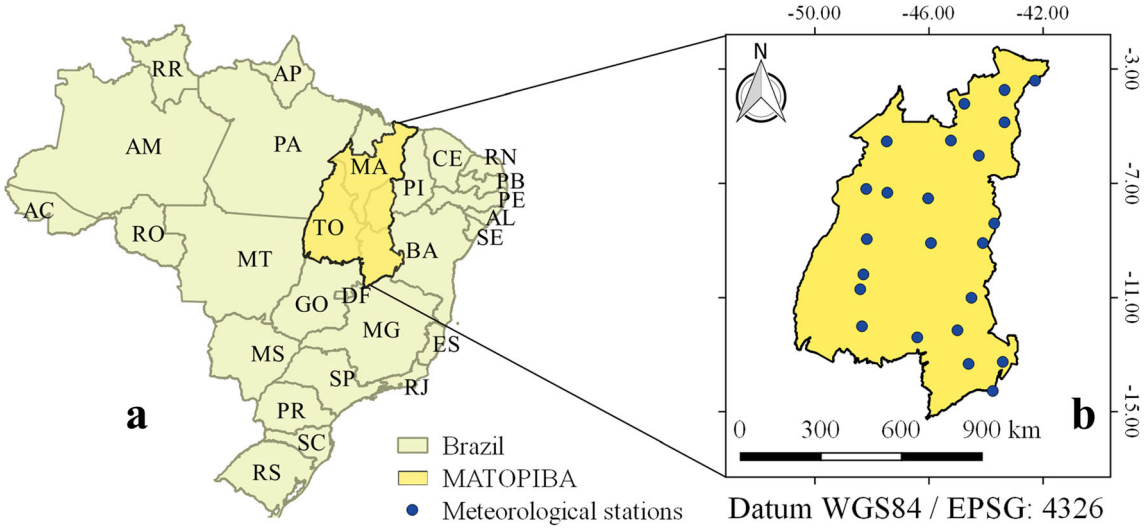

Fig. 1 **a** Location of the MATOPIBA study region in relation to Brazil. **b** Spatial distribution of the weather stations used in the study

libraries Cubist (Kuhn and Quinlan 2018), brnn (Pérez-Rodriguez and Gianola 2013), and kernlab (Karatzoglou et al. 2004) were used. The parameters of the models were adjusted using the caret library (Kuhn et al. 2018), aiming to minimize the root of the mean square error (RMSE).

2.4 Development and validation of models

For the training and testing of the models, the data set was randomly divided into a training set (with 70% of the data) and a test set (with 30% of the data). The training set was used to calibrate ET_0 equations and to model ET_0 with heuristic models. The prediction of the test set was used to evaluate the performance of the equations and models.

To assess the impact of reference evapotranspiration estimates on the demand for maize and soybean crops, the consecutive planting of these crops was simulated for three years (2013, 2014, and 2015) for the municipality of Barreiras, state of Bahia, considering maize planted on April 2015 and soybean on November 2015.

The total evapotranspiration of the crop cycles was compared using the equations of PM, Hargreaves-Samani, and calibrated Hargreaves-Samani (HS_cal), and the machine learning models were developed for the simplest data set (CML). The Hargreaves-Samani, Hargreaves-Samani calibrated, and machine learning methods considering the ML5 were selected to assess whether the models developed from the smallest number of predictor variables presented satisfactory performance.

Crop coefficient data (kc) and duration of the crop cycle were obtained from FAO Bulletin 56 (Allen et al. 1998). The duration of the crop cycle was equal to 140 days and 120 days for maize and soybeans, respectively. The kc values were equal to 0.30, 1.20, and 0.35 for the maize crop and were equal to 0.40, 1.15, and 0.50 for the soybean crop, for the initial, mid-season, and late season phases, respectively. As the frequency of irrigation varied from 2 to 3 days, the water stress coefficient of the crop was considered equal to 1.

2.5 Model performance metrics

The mean bias error (MBE), the mean absolute error (MAE), the RMSE, and the coefficient of determination (R^2) were used as statistical metrics to assess the performance of the

Table 1 Summary of ET_0 equations used

Models of ET_0	Weather data	Equations
Priestley-Taylor (PT)	T_{max}, T_{mean}, T_{min}, SR	$ET0 = \alpha \frac{\Delta}{\Delta+\gamma} \frac{(R_n-G)}{\lambda}$
Makkink (MK)	T_{max}, T_{mean}, T_{min}, SR	$ET0 = \eta \frac{\Delta}{\Delta+\gamma} \frac{SR}{\lambda} - \sigma$
Hargreaves-Samani (HS)	T_{max}, T_{mean}, T_{min}	$ET0 = \delta (T_{max} - T_{min})^{\tau} (T_{mean} + \omega) Ra$
Penman-Monteith (PM)		$ET0{-}PM = 0.408\,(R_n{-}G)\,\frac{+\gamma \frac{900}{T_{avg}} WS\,(e_s - e_a)}{\Delta+\gamma\,(1+0.34\,WS)}$

$\alpha = 1.26$ = empirical constant of Priestley-Taylor; $\eta = 0.61$ and $\sigma = 0.12$ = empirical constants of Makkink; $\delta = 0.0023$, $\tau = 0.5$, and $\omega = 17.8$ = empirical constants of Hargreaves-Samani; λ = latent heat of vaporization (2.45 MJ kg^{-1}); SR = solar radiation, MJ m^{-2} day^{-1}; Ra = extraterrestrial radiation, MJ m^{-2} day^{-1}; $ET_{0\text{-}PM}$ = reference evapotranspiration (PM), mm day^{-1}; R_n = radiation balance, MJ m^{-2} day^{-1}; G = ground heat flow, MJ m^{-2} day^{-1}; WS = wind speed at 2 m of altitude, m s^{-1}; e_s = vapor saturation pressure, kPa; e_a = actual vapor pressure, kPa; Δ = slope of the vapor saturation pressure curve, kPa $°C^{-1}$; γ = psychrometric constant, kPa $°C^{-1}$; T_{mean} = average temperature; T_{max} = maximum temperature; T_{min} = minimum temperature

Table 2 Summary of input settings used to implement machine learning (ML) models

	ML1	ML2	ML3	ML4	ML5
T_{mean}	•	•	•	•	•
T_{max}	•	•	•	•	•
T_{min}	•	•	•	•	•
RH	•		•		
WS	•			•	
SR	•	•			

ML1 = models with all variables; ML2 = models with radiation and temperature; ML3 = models with temperature and relative humidity; ML4 = models with wind speed and temperature; ML5 = models with temperature; T_{mean} = average temperature; T_{max} = maximum temperature; T_{min} = minimum temperature; RH = relative humidity (%); WS = wind speed (m s^{-1}); SR = solar radiation (MJ m^{-2} day^{-1})

equations and models used in the estimation of ET_0, according to Eqs. 1 to 4

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (3)$$

$$R^2 = \frac{\left(\sum_{i=1}^{n}\sum\left(P_i - \overline{P_i}\right)\left(O_i - \overline{O_i}\right)\right)^2}{\sum_{i=1}^{n}\left(P_i - \overline{P_i}\right)^2 \sum_{i=1}^{n}\left(O_i - \overline{O_i}\right)^2} \quad (4)$$

where O_i is the observed reference evapotranspiration data of order i and P_i is the modeled reference evapotranspiration data of order i.

3 Results

3.1 Standard reference evapotranspiration and daily meteorological data for the study area

Table 3 presents the statistics of the standard reference evapotranspiration values and daily meteorological data for the study area. It is noted that the $ET_{0\text{-}PM}$ daily rate ranged from 1.5 to 10.1 mm, with an average of 4.6 mm. The SR presented a daily average of 19.7 MJ m^{-2}, ranging from 6.8 to 33.7 MJ m^{-2}. The maximum temperature showed a maximum value equal to 44.7 °C, with an average of 33.5 °C and a minimum of 21.5 °C. For the minimum temperature, the value varied from 6.3 to 30.8 °C, with an average equal to 21.3 °C. The average temperature data set showed a maximum value of 35.4 °C, an average of 27.5 °C, and a minimum of 17.7 °C. It was observed that SR (0.846) and T_{max} (0.704) were the variables that showed the highest correlation with $ET_{0\text{-}PM}$.

3.2 Calibration of equations for calculating reference evapotranspiration

After calibration, the value of the constant α in the Priestley-Taylor equation changed from 1.26 to 1.195. The η and σ values of the Makkink equation changed from 0.61 and 0.12 to 0.738 and 0.049, respectively. The values of the constants in the Hargreaves-Samani equation (δ, τ, ω) were altered from 0.0023, 0.5, and 17.8 to 0.0026, 0.633, and 2.63, respectively. From now on, the calibrated equations were called PT_cal, MK_cal, and HS_cal, respectively.

3.3 Evaluation of model performance

Table 4 presents the statistical metrics for assessing the performance of the different models used, stratified by a group of variables established in the methodology. A better performance of the models that had the temperature and the solar radiation (ML2) as input data was observed.

In the temperature-radiation group ML2 (Table 4), the MAE and the RMSE obtained by the heuristic methods were lower than the MK_cal by 20.8% and 47.7% for NN2, by 22.9% and 10.9% for CB2 and SV2, and by 12.5% and 6.3% for SW2, respectively. For the PT_cal method, there was also a reduction in MAE and RMSE, when compared to heuristic methodologies by 32.1% and 30.5% for NN2, by 33.9% and 30.5% for CB2, by 35.7% and 29.3% for SV2, and by 25.0% and 26.8% for SW2.Within the temperature group (ML5), comparing the HS_cal method with the heuristic methodologies, an improvement in performance was observed, also being observed a reduction in the same MAE and RMSE metrics by 10.3% and 7.9% for NN5, by 11.8% and 9% for CB5, by 11.8% and 7.9% for SV5, and by 1.5% and 2.3% for SW5.

The scatterplot between $ET_{0\text{-}PM}$ and estimated ET_0 show R^2 values above 0.7 for the machine models developed in the temperature-radiation group, and a similar behavior for the MK_cal and MK methods in this group (Fig. 2a–f). However, when analyzing the MBE values for the test set (Table 4), negative values are observed, indicating a tendency of the CB2, SV2, and MK methods to underestimate ET_0. The NN2 and SW2 methods presented MBE values equal to 0.0, showing that there was no overestimation or underestimation, whereas the MK_cal method presented an overestimation of 10% of ET_0. Analyzing the MBE values of the PT and PT_cal method, an overestimation of 25% and an underestimation of 19% are observed in ET_0 compared to $ET_{0\text{-}PM}$; in addition,

Table 3 Standard reference evapotranspiration ($ET_{0\text{-}PM}$) and daily meteorological data for the study area

	Maximum	Medium	Minimum	SD_x	Skew	Kurt	PCC	SCC
T_{max}	44.7	33.5	21.5	2.70	− 0.11	0.26	0.704	0.699
T_{mean}	35.5	27.5	17.7	1.98	− 0.20	0.38	0.032	0.023
T_{min}	30.8	21.4	6.3	2.50	− 1.05	1.59	0.501	0.502
RH	100.0	67.8	5.0	15.45	− 0.34	− 0.76	− 0.646	− 0.647
WS	10.0	1.4	0.0	0.98	1.47	4.21	0.556	0.507
SR	33.7	19.7	6.8	4.43	− 0.51	− 0.24	0.846	0.881
$ET_{0\text{-}PM}$	10.1	4.6	1.5	1.26	0.47	0.56	1.00	1.00

SD_x = standard deviation; Skew = asymmetry; Kurt = kurtosis, PCC = Pearson's correlation coefficient in relation to evapotranspiration; SCC = Spearman rank correlation coefficient; T_{mean} = average temperature; T_{max} = maximum temperature; T_{min} = minimum temperature; RH = relative humidity (%); WS = wind speed (m s^{-1}); SR = solar radiation (MJ m^{-2} day^{-1}); $ET_{0\text{-}PM}$ = reference evapotranspiration (PM), mm day^{-1}

these methods were those that showed the least correlation with $ET_{0\text{-}PM}$, with R^2 values below 0.6 (Fig. 2g, h).

Figure 3 shows the dispersion graphs obtained for the test set of the temperature group. Observations of the NN5, CB5, and SV5 present R^2 approximations of 0.6 (Fig. 3a–c) while the SW5, HS, and HS_cal methods obtained a value of approximately 0.5 when compared with the $ET_{0\text{-}PM}$ method (Fig. 3d–f). The MBE values (Table 4) showed that there is a tendency for the CB5, SV5, and HS_cal methods to underestimate the ET_0 by 6%, 7%, and 2%, respectively. The NN5 and SW5 methods presented MBE values equal to 0.0, showing that there was no overestimation or underestimation, whereas the HS method showed a 64% overestimation in ET_0, according to its MBE value.

The monthly ET_0 estimates obtained by the heuristic methods (Fig. 4), in the group with all the data (ML1), show an underestimation by the methods SV1, NN1, CB1, and SW1 in the months from January to March, with an average of underestimation of 1.7%, 1.2%, 1.0%, and 1.5%, and from September to December, with an average underestimation of 1.38%, 1.16%, 1.0%, and 2.1%, respectively, with the highest underestimation occurring in January in all methods. In the months of April to August, these methods overestimated ET_0 by an average of 1.6%, 1.8%, 1.5%, and 2.9%, respectively, with the highest overestimation in June in all methods.

In the temperature-radiation group (ML2) on a monthly scale, an underestimation by the methods of SV2 and CB2 was observed in the months of January and February, with an average of 0.8% and 1.2%, and in the months of June to December, with an average of 3.39% and 3.7%, respectively, with the highest underestimation in September. The NN2 method underestimated ET_0 in the months of July to December with an average of 1.9%, and the SW2 method underestimated ET_0 in the months from August to December, with an average underestimation of 3.4%, with the highest underestimation occurring in the month of September in these methods. There was an overestimation by the SV2 and CB2 methods in the months of March to May with an average overestimation of 1.9% and 1.4%, with the highest overestimation occurring in April. The NN2 method overestimated ET_0 in the months from January to June by an average of 2.5%, and the SW2 method overestimated ET_0 in the months from January to July by an average of 3.2%, with the highest overestimation occurring in May in both methods.

In the temperature-relative humidity (ML3) group on a monthly scale, an underestimation by the methods of SV3, CB3, and NN3 was observed in the months of January to March, with an average of 2.0%, 2.5%, and 2.6%, and in the months of June to August, with an average of 1.6%, 1.4%, and 1.2% respectively, with the highest underestimation in February for the SV3 and CB3 methods and in March for the NN3 method. The SW3 method underestimated ET_0 in the months of January to June, with an average of 0.9%, and in the months of August and September, with an average underestimation of 0.8%, with the highest underestimation occurring in the month of April. There was an overestimation by the methods of SV3, CB3, and NN3 in the months of September to December, with an average of overestimation of 1.7%, 1.7%, and 1.4%, and in the months of April and May, with an average of 1.4%, 2.1%, and 1.9%, with a greater overestimation occurring in April (SV3 and NN3) and December (CB3). The SW3 method overestimated ET_0 in July, October, November, and December by an average of 1.2%, with the greatest overestimation occurring in November.

In the temperature-wind speed group (ML4) on a monthly scale, an underestimation by the methods of SV4 and CB4 was observed in the months of February, March, August, and September with an average of 1.2% and 1.3%, with the highest underestimation occurring in March and an average overestimate of 1.7% in the months of January, April to July, and October to December, with the highest overestimation occurring in December (SV4) and May (CB4). The NN4

Table 4 MBE, MAE, RMSE, and R^2 for ET_0 equations and heuristic models during the testing phase

	MBE (mm day^{-1})	MAE (mm day^{-1})	RMSE (mm day^{-1})	R^2
Group with all variables (ML1)				
NN1	0.00	0.08	0.10	0.99
CB1	0.00	0.07	0.10	0.99
SV1	− 0.01	0.09	0.11	0.99
SW1	0.00	0.24	0.33	0.93
Temperature-radiation group (ML2)				
NN2	0.00	0.38	0.57	0.79
CB2	− 0.11	0.37	0.58	0.79
SV2	− 0.09	0.36	0.58	0.79
SW2	0.00	0.42	0.60	0.74
MK	− 0.87	0.88	1.09	0.75
MK_cal	0.10	0.48	0.64	0.75
PT	0.25	0.69	0.84	0.59
PT_cal	− 0.19	0.56	0.82	0.59
Temperature-relative humidity group (ML3)				
NN3	− 0.01	0.56	0.74	0.65
CB3	0.00	0.53	0.72	0.67
SV3	0.01	0.54	0.74	0.65
SW3	− 0.01	0.61	0.79	0.61
Temperature-wind speed group (ML4)				
NN4	0.00	0.45	0.61	0.77
CB4	0.03	0.42	0.58	0.78
SV4	0.03	0.43	0.60	0.77
SW4	0.00	0.52	0.68	0.71
Temperature group (ML5)				
NN5	0.00	0.61	0.82	0.58
CB5	− 0.06	0.60	0.81	0.58
SV5	− 0.07	0.60	0.82	0.58
SW5	0.00	0.67	0.87	0.52
HS	0.64	0.92	1.12	0.46
HS_cal	− 0.02	0.68	0.89	0.50

NN1, NN2, NN3, NN4, NN5, CB1, CB2, CB3, CB4, CB5, SV1, SV2, SV3, SV4, SV5, SW1, SW2, SW3, SW4, and SW5 are Bayesian regularized neural networks, cubist regression, support vector machine with linear kernel function, and stepwise models with all variables (xx1), radiation and temperature (xx2), temperature and relative humidity (xx3), wind speed and temperature (xx4), and temperature (xx5), respectively. MK = Makkink equation; MK_cal = calibrated Makkink equation; PT = Priestley-Taylor equation; PT_cal = calibrated Priestley-Taylor equation; HS = Hargreaves-Samani equation; HS_cal = calibrated Hargreaves-Samani equation

method underestimated ET_0 from February to April and from August to October with an average of 1.5%, with the highest underestimation in March, and overestimated ET_0 in the months of January, May to July, and November to December by an average of 1.7%, with the highest overestimation in December. The SW4 method underestimated the ET_0 in the months of April, August, September, and October, with an average underestimation of 1.3%, with the highest underestimation in the month of August, and overestimated an average of 0.8% in the months from January to March, from May to July, and from November to December, with the highest overestimation in December.

In the temperature group (ML5) on a monthly scale, an underestimation by the methods of SV5, CB5, and NN5 was observed in the months of January to March and in the months of June to October, with an average of 2.9%, 2.7%, and 2.3%, respectively, with a higher underestimation occurring in August, also showing an overestimation of 2.0%, 2.0%, and 3.5% in the months of April, May, November, and December, with the highest overestimation in April in the three methods. The SW5 method underestimated ET_0 from August to October with an average of 3.4%, with the highest underestimation in August, and overestimated ET_0 by

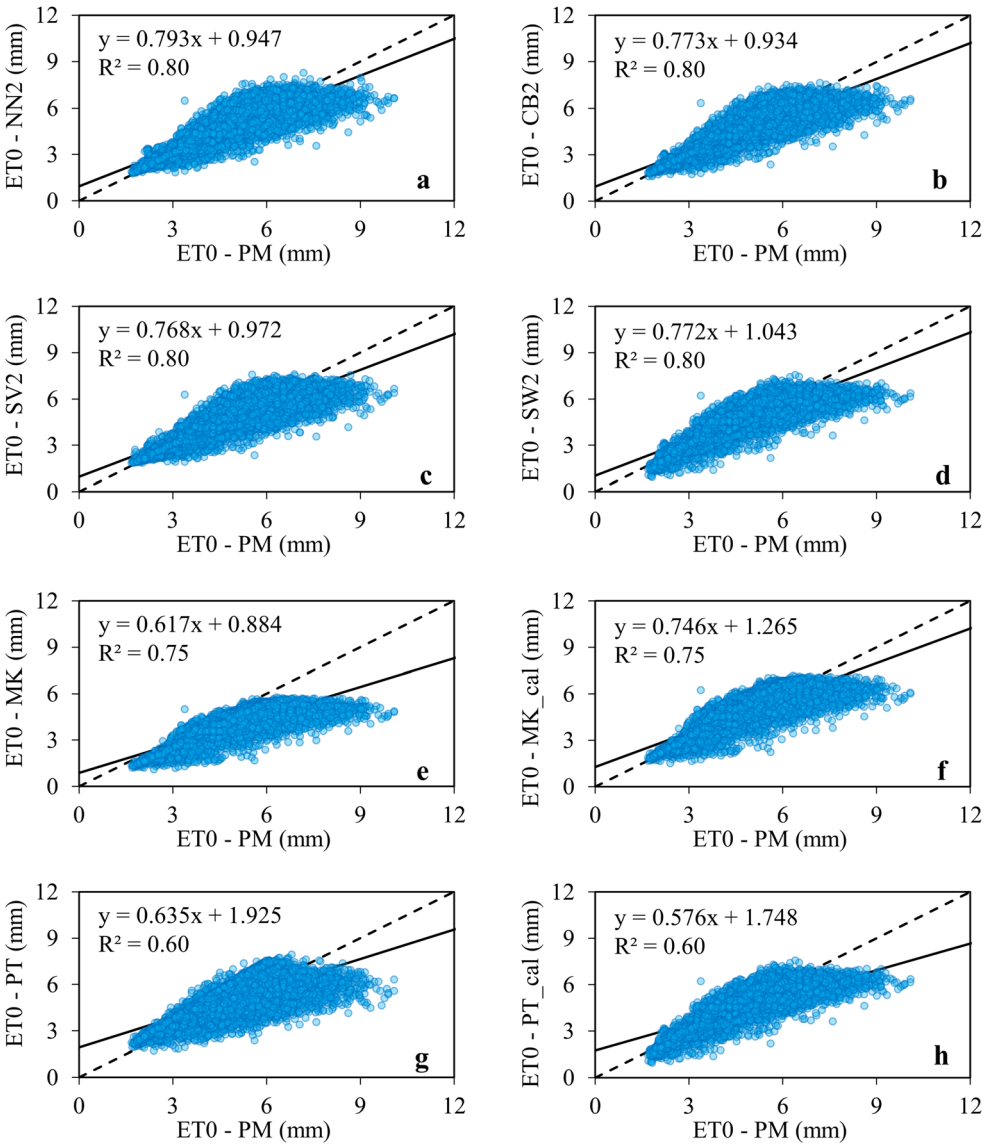

Fig. 2 Comparison of daily reference evapotranspiration, for the radiation and temperature group in the test phase, calculated by the heuristic methods (**a** Bayesian regularization (NN2), **b** cubist regression (CB2), **c** support vector machine with linear kernel function (SV2), **d** stepwise (SW2)) and the **e** Hargreaves-Samani, **f** calibrated Hargreaves-Samani, **g** Priestley-Taylor, **h** calibrated Priestley-Taylor, and Penman-Monteith methods

1.5% in the months from January to June and from November to December.

The MK method underestimated ET_0 values in all months (average of 18.5%). When the model was calibrated, however, there was an improvement in performance, which started to underestimate the ET_0 in the months of August to November by an average of 3.7% and to overestimate the ET_0 by an average of 6.3% in the months of January to July and December. The PT method overestimated ET_0 values by 15.9%, 17.8%, 18.6%, 16.3%, and 9.9%, respectively, from January to May. After calibration, the overestimation decreased to 5.3%, 7.0%, 7.7%, and 5.6%, in the months from January to April, with no further overestimation in the month of May. The HS method overestimated ET_0 values in all months, with the highest overestimation occurring in February (24.3%) and the lowest in August (3.4%); however, when there was calibration, the overestimated value in February was reduced to 5.3% and the estimated value of August started to be underestimated by 7.8%. Figure 4 makes it clear that the machine learning techniques show lower bias monthly when compared to the equations, even when these are calibrated.

3.4 Simulation of water demand for crops

Table 5 presents the results of simulations of water demand for maize and soybean crops using the daily reference evapotranspiration data calculated by the heuristic and Penman-Monteith methods.

An overestimation of ETc of 21.7% (110 mm), 18.1% (93 mm), and 17.8% (90 mm) was observed using the HS method when compared with the PM method in maize plantations in the 2013, 2014, and 2015, respectively. After calibrating the method, it is possible to obtain an improvement in ETc. In this situation, the method overestimated ETc by 4.4% (22 mm) in 2013 and 0.3% (2

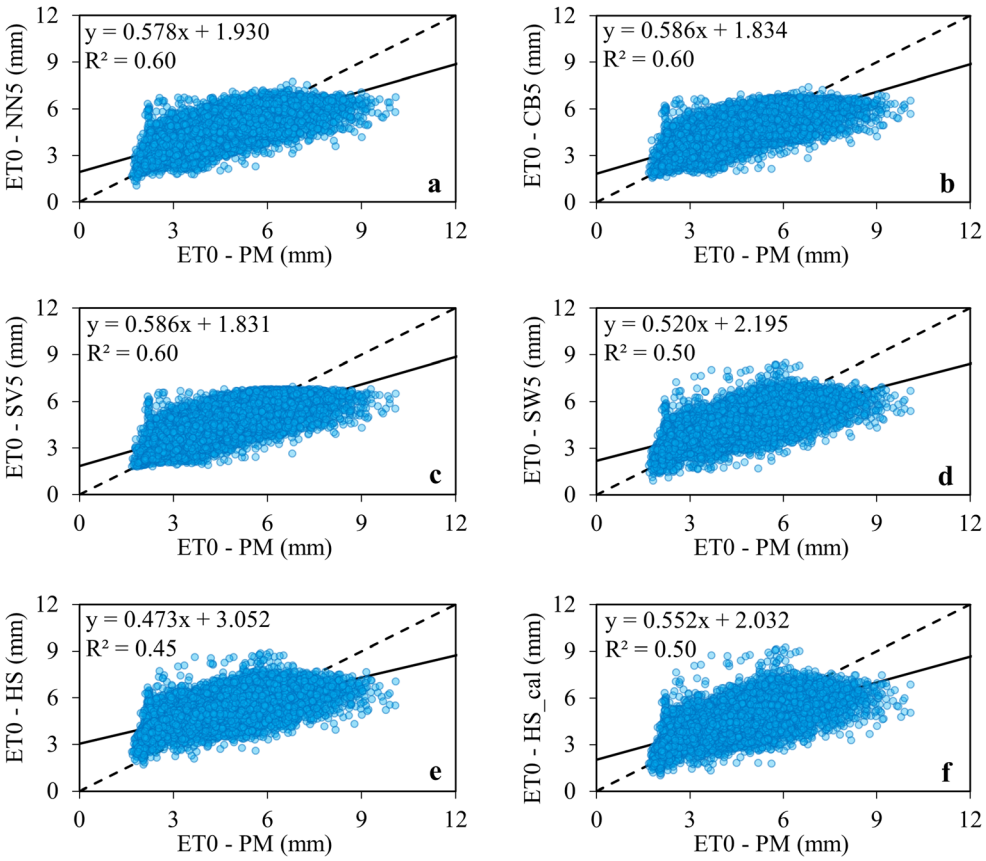

Fig. 3 Comparison of the daily reference evapotranspiration, for the temperature group in the test phase, calculated by the heuristic methods (**a** Bayesian regularization (NN5), **b** cubist regression (CB5), **c** support vector machine with linear kernel function (SV5), **d** stepwise (SW5)) and the **e** Hargreaves-Samani, **f** calibrated Hargreaves-Samani, and Penman-Monteith methods

mm) in 2014. In 2015, the method underestimated ETc by 0.1% (0.5 mm). Similar behavior was observed in soybean crops, where the HS method overestimated ETc by 21.3% (107 mm), 20.4% (109 mm), and 20% (110 mm) in the 2013, 2014, and 2015 plantations, respectively. After calibrating the method, there was an improvement in the ETc estimates. In this situation, the method overestimated ETc only by 1.3% (7 mm), 2.6% (14 mm), and 4.1% (22 mm), for the years 2013, 2014, and 2015, respectively.

The method that used only temperature data (CB5) overestimated ETc for the maize crop in the 2013, 2014, and 2015 seasons by 9.1% (46 mm), 7.2% (37 mm), and 7.0% (36 mm), respectively, and for soy, there was an underestimation of ETc in the plantations of 2013 and 2014, by 3.7% (18 mm) and 2.2% (12 mm), and an overestimate of 0.2% (1 mm) in 2015, when compared with the PM method.

Figure 5 shows the behavior of the ETc estimated by each method in the simulated period. It is observed that the ETc values estimated by HS are higher daily before calibration, and after calibration, these valuesare reduced considerably, getting closer to the values estimated by the PM method.

4 Discussion

Analyzing the correlation between the meteorological variables of the study area (Table 3), as expected, a greater correlation between $ET_{0\text{-}PM}$ and SR is observed, justifying the fact that solar radiation is the climatic variable that most influences the reference evapotranspiration (Allen et al. 1998), corroborating with other results obtained in similar studies (Gurski et al. 2018; Hupet and Vanclooster 2001).

The overestimations observed in the ET_0 values obtained from the Priestley-Taylor equation may have occurred due to the fact that this equation was developed for saturated surface conditions, a condition not found in the locations where the meteorological stations used in this study are installed (Cavalcante Junior et al. 2011). Fernandes et al. (2012) and da Silva Farias et al. (2020) calculating ET_0 by the Priestley-Taylor method for the regions of Campos dos Goytacazes and Pará, respectively, also observed a tendency of this equation to underestimate ET_0. According to Caporusso and Rolim (2015), better performance of this method is observed during the rainy seasons and lower performances in the dry seasons. The overestimation observed in the Hargreaves-Samani method may be due to the high temperatures that occur in the study area (Aguilar and Polo 2011). Several studies have shown a tendency to overestimate ET_0 by the Hargreaves-Samani method (Ferreira et al. 2018; Palaretti et al. 2014; Tabari 2010; Martinez and Thepadia 2009). The underestimation of the Makkink equation may be related to local climatic conditions, which was also verified in other studies carried out in

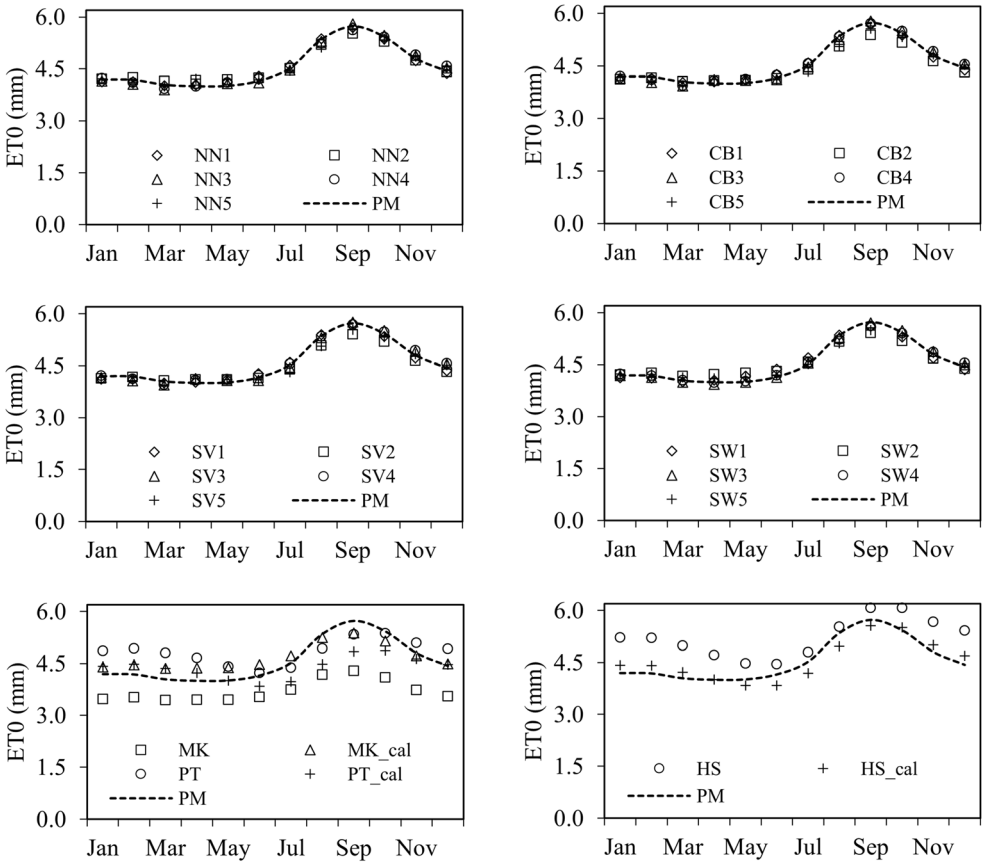

Fig. 4 Average monthly reference evapotranspiration calculated by the heuristic and Penman-Monteith methods for the phase test. NN1, NN2, NN3, NN4, NN5, CB1, CB2, CB3, CB4, CB5, SV1, SV2, SV3, SV4, SV5, SW1, SW2, SW3, SW4, and SW5 are Bayesian regularization, cubist regression, support vector machine with linear kernel function, and stepwise models with all variables, radiation and temperature, temperature and relative humidity, wind speed and temperature, and temperature, respectively. MK Makkink equation; MK_cal calibrated Makkink equation; PT Priestley-Taylor equation; PT_cal calibrated Priestley-Taylor equation; HS Hargreaves-Samani equation; HS_cal calibrated Hargreaves-Samani equation; PM Penman-Monteith equation

dry and humid conditions (Fernandes et al. 2012; Pilau et al. 2012; Lacerda and Turco 2015).

The calibrated Priestley-Taylor, Hargreaves-Samani, and Makkink equations, as expected, performed better. The Priestley-Taylor equation variable (α) after calibration had its value reduced by 5.43%. The original variables of the Hargreaves-Samani equation (δ, τ, ω) had their values changed to 0.0023, 0.633, and 2.63. For the Makkink equation, the η coefficient had its value increased by 21% and the σ coefficient had its value decreased by 59.2%. The improvement in the performance of these equations observed after their calibration indicates the empirical character of the equations and the need for local calibrations.

The comparative analysis of the results indicated that the models NN, CB, and SV, in general, presented better performance than the SW. The SW method uses interactions to generate models from an adjusted multiple linear regression, that is, it does not have as much complexity when compared with other methods. The NN, CB, and SV methods are machine learning methods, that is, they are more robust and provide the model with greater generalization capacity for new data sets (Torres et al. 2019). Hassan et al. (2017) studying solar radiation estimation models in five different regions proved that machine learning models are more accurate than standard models.

Within the simulations performed with missing data, it is observed that the groups with temperature-radiation (ML2) and temperature-wind speed (ML4) have close R^2 values (NN2 = 0.80, CB2 = 0.80, SV2 = 0.80, and SW2 = 0.75 for ML2, and NN4 = 0.80, CB4 = 0.80, SV4 = 0.80, and SW4 = 0.70 for ML4), this method being indicated as an alternative

Table 5 Crop evapotranspiration for soybean and maize crops for the simulation period

	Crop evapotranspiration (mm)			
	PM	HS	HS_cal	CB5
Maize				
2013	506	616 (+ 21.7%)	528 (+ 4.4%)	552 (+ 9.1%)
2014	511	604 (+ 18.1%)	513 (+ 0.3%)	548 (+ 7.2%)
2015	506	596 (+ 17.8%)	505.5 (− 0.1%)	542 (+ 7.0%)
Soybean				
2013	499	606 (+ 21.3%)	506 (+ 1.3%)	481 (− 3.7%)
2014	535	644 (+ 20.4%)	549 (+ 2.6%)	523 (− 2.2%)
2015	552	662 (+ 20.0%)	574 (+ 4.1%)	553 (+ 0.2%)

HS = Hargreaves-Samani equation; HS_cal = calibrated Hargreaves-Samani equation; CB5 = cubist regression with temperature; PM = Penman-Monteith equation

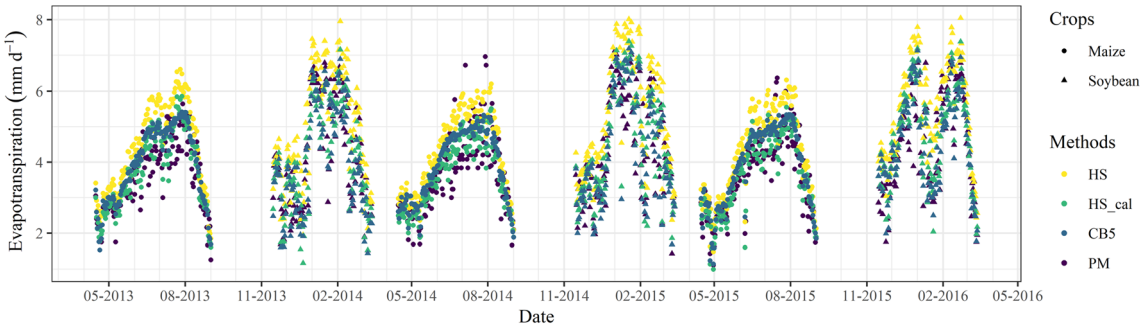

Fig. 5 Behavior of the evapotranspiration values of the corn and soybean crop for the simulation period of the methods. HS Hargreaves-Samani equation; HS_cal calibrated Hargreaves-Samani equation; CB5 cubist regression with temperature; PM Penman-Monteith equation

when there is no solar radiation data, which, in general, are more difficult to obtain.

Previous studies (Torres et al. 2011; Tabari et al. 2012; Antonopoulos and Antonopoulos 2017) have shown that the more input parameters the model has, the better the accuracy of the method's prediction tends to be. However, in different climates, the contribution of meteorological variables is different, as seen in Table 3 of the present study.

Note that when using machine learning techniques, ET_0 prediction becomes viable, even in situations where there is a lack of any variable. This is due to the fact that there is a high capacity for generalization of the model, making the lack of variables not a problem for the prediction of ET_0 (Ferreira et al. 2019; Zhang and Yan 2014; Zscheischler et al. 2012). The use of machine learning is promising in terms of accuracy, stability, and computational effectiveness in predicting daily ET_0. Thus, these techniques gain importance in studies of irrigation management and management of water resources in regions with a lack of climatic data.

In the present study, the machine learning method with the least number of data (CB5) presented an RMSE value of 0.8, corroborating with another study that used the same amount of data using the same input with neural networks and support vector machines and presented average RMSE of 0.8 for Brazil (Ferreira et al. 2019), showing that the performance achieved for the MATOPIBA region is adequate.

The greatest demands for the soybean crop were observed in the estimates made by the PM, HS, HS_cal, and CB5 methods, during the 2014 and 2015 plantations; for maize, the greatest water demands were observed in the estimates made by the PM method referring to the 2014 planting and for the HS, HS_cal, and CB5 methods in the 2013 planting (Table 5). Although there are harvests in which greater water demand is observed, the difference between them is relatively low.

The irrigated areas in the MATOPIBA region are relatively large areas. The central pivot is the main irrigation system in the region, with an average irrigated area of 80 ha. Small errors in the water depth calculation can have a big impact in terms of the volume of water withdrawn. For example, when using the HS equation to calculate the ETc for the maize crop planted in 2013, it is noted that, compared to the demand calculated by the $ET_{0\text{-}PM}$ equation, about 1100 m^3 of additional water would be used per cultivated hectare. When comparing HS_cal and CB5, which are methods that have the same input data, this value would drop to 220 m^3 and 460 m^3, respectively. However, although HS_cal had a total water demand closer to the reference (PM), its performance was worse (higher RMSE and R^2), which indicates greater variability in relation to the reference.

5 Conclusions

Machine learning methods were robust in predicting ET_0, even when there is no variable, showing superior performance when compared to other alternative methods established in the literature. However, the greater the number of input data for the models, the better the results, especially when using solar radiation or wind speed.

Among the machine learning methods, the cubist regression method in the temperature group performed better, with the least number of variables that provided reference evapotranspiration results closer to the standard Penman-Monteith method, and, when compared to Hargreaves-Samani calibrated equation that has the same climatological variables, obtained better statistical metrics. When the simulation of water demand for soybeans and maize is observed, it is noted that the cubist regression method in the temperature group performed better when compared to the Hargreaves-Samani method.

The cubist regression and support vector machine methods were, for all combinations of input variables, the methods with the highest determination coefficients and the best results for MBE, MAE, and RMSE. The smallest errors in estimating water demand for soybean and maize crops were obtained by the calibrated equation of Hargreaves-Samani and cubist regression methods in the temperature group, obtaining greater precision in estimating crop evapotranspiration with the use of few input variables.

Acknowledgments The authors would like to thank the National Institute of Meteorology (INMET) for providing the climatic data used in the present study.

Authors' contributions All authors contributed to the conception and design of the study. Material preparation, data collection, and analysis were carried out by Diego Bispo dos Santos Farias, Daniel Althoff, Lineu Neiva Rodrigues, and Roberto Filgueiras. The first draft of the manuscript was written by Diego Bispo dos Santos Farias, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, Finance Code 001

Data availability Not applicable

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Code availability Not applicable

References

Abraham S, Raisee M, Ghorbaniasl G, Contino F, Lacor C (2017) A robust and efficient stepwise regression method for building sparse polynomial chaos expansions. J Comput Phys 332:461–474

Agência Nacional de Águas (ANA) (2017) Atlas irrigação: uso da água na agricultura irrigada. ANA, Brasília 85 p

Aguilar C, Polo MJ (2011) Generating reference evapotranspiration surfaces from the Hargreaves equation at watershed scale. Hydrol Earth Syst Sci 15(8):2495–2508

Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspiration-guidelines for computing crop water requirements-FAO irrigation and drainage paper 56. FAO, Rome, 300(9), D05109

Althoff D, Rodrigues LN (2019) The expansion of center-pivot irrigation in the Cerrado biome. IRRIGA 1(1):56–61

Althoff D, Bazame HC, Filgueiras R, Dias SHB (2018) Heuristic methods applied in reference evapotranspiration modeling. Ciência Agrotecnol 42(3):314–324

Althoff D, Filgueiras R, Dias SHB, Rodrigues LN (2019) Impact of sum-of-hourly and daily timesteps in the computations of reference evapotranspiration across the Brazilian territory. Agric Water Manag 226:105785

Antonopoulos VZ, Antonopoulos AV (2017) Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. Comput Electron Agric 132:86–96

Caporusso NB, Rolim G d S (2015) Reference evapotranspiration models using different time scales in the Jaboticabal region of São Paulo, Brazil. Acta Scientiarum. Agronomy 37(1):1–9

Cavalcante Junior EG, Oliveira AD, Almeida BM, Sobrinho JE (2011) Métodos de estimativa da evapotranspiração de referência para as condições do semiárido Nordestino. Semina Ciências Agrárias 32(supl.1):1699–1708

da Silva Farias VD, Costa DLP, de Novoa Pinto JV, de Souza PJOP, de Souza EB, Ortega-Farias S (2020) Calibration of reference evapotranspiration models in Pará. Acta Sci Agron 42:e42475–e42475

de Miranda EE, Magalhães LA, de Carvalho CA (2014) Proposta de Delimitação Territorial do MATOPIBA. Embrapa Territorial-Outras publicações técnicas (INFOTECA-E)

Doorenbos J, Pruitt WO (1977) Guidelines for predicting crop water requirements. Rome: FAO, 179 p. (Irrigation and Drainage Paper, 24)

FAO (2015) Towards a water and food secure future: critical perspectives for policy-makers. Food and Agriculture Organization of the United Nations, Rome, and World Water Council, Marseille. 61 pp

Fernandes LC, Paiva CM, Rotunno Filho OC (2012) Evaluation of six empirical evapotranspiration equations - case study: Campos dos Goytacazes/RJ. Rev Bras Meteorol 27(3):272–280

Ferreira LB, da Cunha FF (2020) New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agric Water Manag 234:106113

Ferreira LB, Cunha FF, Duarte AB, Sediyama GC, Cecon PR (2018) Calibration methods for the Hargreaves-Samani equation. Ciência Agrotecnol 42(1):104–114

Ferreira LB, da Cunha FF, de Oliveira RA, Fernandes Filho EI (2019) Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – a new approach. J Hydrol 572:556–570

Fishman R, Devineni N, Raman S (2015) Can improved agricultural water use efficiency save India's groundwater? Environ Res Lett 10(8):084022

Gurski BC, Jerszurki D, Souza JLMD (2018) Alternative methods of reference evapotranspiration for Brazilian climate types. Rev Bras Meteorol 33(3):567–578

Hargreaves GH, Samani ZA (1985) Reference crop evapotranspiration from temperature. Appl Eng Agric 1(2):96–99

Hassan MA, Khalil A, Kaseb S, Kassem MA (2017) Exploring the potential of tree-based ensemble methods in solar radiation modeling. Appl Energy 203:897–916

Hupet F, Vanclooster M (2001) Effect of the sampling frequency of meteorological variables on the estimation of the reference evapotranspiration. J Hydrol 243(3–4):192–204

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab-an S4 package for kernel methods in R. J Stat Softw 11(9):1–20

Keshtegar B, Kisi O, Arab HG, Zounemat-Kermani M (2018) Subset modeling basis ANFIS for prediction of the reference evapotranspiration. Water Resour Manag 32(3):1101–1116

Kisi O, Alizamir M (2018) Modelling reference evapotranspiration using a new wavelet conjunction heuristic method: wavelet extreme learning machine vs wavelet neural networks. Agric For Meteorol 263:41–48

Kuhn M, Quinlan R (2018) Cubist: rule-and instance-based regression modeling. R package version 0.2. 2

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Team RC (2018) caret: classification and regression training. R package version v6. 0.82. 2017

Lacerda ZC, Turco JE (2015) Estimation methods of reference evapotranspiration (ETo) for Uberlândia-MG. Engenharia Agrícola 35(1):27–38

López-Urrea R, de Santa Olalla FM, Fabeiro C, Moratalla A (2006) Testing evapotranspiration equations using lysimeter observations in a semiarid climate. Agric Water Manag 85:15–26

Makkink GF (1957) Testing the Penman formula by means of lysimeters. J Inst Water Eng 11:277–288

Martinez CJ, Thepadia M (2009) Estimating reference evapotranspiration with minimum data in Florida. J Irrig Drain Eng 136(7):494–501

Moré JJ (1978) The Levenberg-Marquardt algorithm: implementation and theory. In: Numerical analysis. Springer, Berlin, pp 105–116

Palaretti LF, Mantovani EC, Sediyama GC (2014) Análise da sensibilidade dos componentes da equação de Hargreaves-Samani

para a região de Bebedouro-SP. Rev Bras Meteorologia 29(2):299–306

Pérez-Rodriguez P, Gianola D (2013) brnn: brnn (Bayesian regularization for feed-forward neural networks). R package version 0.3. R Found. Stat. Comput., Vienna

Pilau FG, Battisti R, Somavilla L, Righi EZ (2012) Desempenho de métodos de estimativa da evapotranspiração de referência nas localidades de Frederico Westphalen e Palmeira das Missões, RS. Ciência Rural 42(2):283–290

Pradhan P, Fischer G, van Velthuizen H, Reusser DE, Kropp JP (2015) Closing yield gaps: how sustainable can we be? PLoS One 10(6): e0129487

Priestley CHB, Taylor RJ (1972) On the assessment of surface heat flux and evaporation using large-scale parameters. Mon Weather Rev 100(2):81–92

R Core Team (2018). R version 3.5. 0. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Seifi A, Riahi H (2020) Estimating daily reference evapotranspiration using hybrid gamma test-least square support vector machine, gamma test-ANN, and gamma test-ANFIS models in an arid area of Iran. Journal of Water and Climate Change, 11(1):217–240

Shiri J, Nazemi AH, Sadraddini AA, Landeras G, Kisi O, Fard AF, Marti P (2014) Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Comput Electron Agric 108:230–241

Silva VPR, Maciel GF, Braga CC, Júnior S, Souza EP, Almeida RSR, Silva MT, Holanda RM (2018) Calibration and validation of the AquaCrop model for the soybean crop grown under different levels of irrigation in the Motopiba region, Brazil. Ciência Rural, 48(1), e20161118. Epub December 21, 2017.https://doi.org/10.1590/0103-8478cr20161118

Sparovek G, Maule RF, Barretto AGOP, Dourado Neto D, Martins SP (2014) Análise territorial para o desenvolvimento da agricultura irrigada no Brasil. MI/FEALQ, Piracicaba

Stöckle CO, Kjelgaard J, Bellocchi G (2004) Evaluation of estimated weather data for calculating Penman-Monteith reference crop evapotranspiration. Irrig Sci 23:39–46

Tabari H (2010) Evaluation of reference crop evapotranspiration equations in various climates. Water Resour Manag 24(10):2311–2337

Tabari H, Nikbakht J, Talaee PH (2012) Identification of Trend in Reference Evapotranspiration Series with Serial Dependence in Iran. Water Resources Management 26(8):2219–2232

Torres AF, Walker WR, McKee M (2011) Forecasting daily potential evapotranspiration using machine learning and limited climatic data. Agric Water Manag 98(4):553–562

Torres R, Ohashi O, Pessin G (2019) A machine-learning approach to distinguish passengers and drivers reading while driving. Sensors 19(14):3174

Wang Y, Liu B, Su B, Zhai J, Gemmer M (2011) Trends of calculated and simulated actual evaporation in the Yangtze River basin. J Clim 24(16):4494–4507

Wen X, Si J, He Z, Wu J, Shao H, Yu H (2015) Support-vector-machine-based models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions. Water Resour Manag 29(9):3195–3209

Wu L, Fan J (2019) Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration. PloS one, v. 14, n. 5, p. e0217520

Zhang X, Yan X (2014) Temporal change of climate zones in China in the context of climate warming. Theor Appl Climatol 115(1–2):167–175

Zscheischler J, Mahecha MD, Harmeling S (2012) Climate classifications: the value of unsupervised clustering. Procedia Computer Science 9:897–906