Evaluating evapotranspiration using data mining instead of physical-based model in remote sensing

Lamya Neissi1 · Mona Golabi1 · Mohammad Albaji1 · Abd Ali Naseri1

Received: 4 December 2020 / Accepted: 12 October 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2021

Abstract
Precise calculations for determining the water requirements of plants and the extent of evapotranspiration are crucial in determining the volume of water consumed for plant production. In order to estimate evapotranspiration over an extended area, different remote sensing algorithms require numerous climatological variables; however, climatological variable measurements cover only limited areas thus resulting into erroneous calculations over extended areas. The exploiting of both data mining and remote sensing technologies allows for the modeling of the evapotranspiration process. In this research, the physical-based SEBAL evapotranspiration algorithm was remodeled using M5 decision tree equations in GIS. The input variables of the M5 decision tree consisted of Albedo, emissivity, and Normalized Difference Water Index (NDWI) which were defined as absorbed light, transformed light, and plant moisture, respectively. After extracting the best equations in the M5 decision tree model for 8 April 2019, these equations were modeled in GIS using python scripts for 8 April 2019 and 3 April 2020, respectively. The calculated correlation coefficient \( R^2 \), mean absolute error \( MAE \), and root mean squared error \( RMSE \) for 8 April 2019 were 0.92, 0.54, and 0.42, respectively, and for 3 April 2020 were 0.95, 0.31, and 0.23 in order. Moreover, for the further evaluation of the model, a sensitivity analysis and an uncertainty analysis were carried out. The analysis revealed that evapotranspiration is more sensitive to Albedo than the two other model inputs, and when applying data mining techniques instead of SEBAL, the estimation of evapotranspiration has a lower accuracy.

1 Introduction
Nowadays precision agriculture has become a highly important issue and is applied in most agricultural sectors, especially in the irrigation sector. Agricultural activities are now carried out with more precision through the utilization of various methods (in terms of hardware and software) and state-of-the-art technologies which can be best exemplified by the application of various apparatus such as pressurized irrigation systems, drones, and the internet of things for crop monitoring, in addition to software consisting of machine learning, deep learning, and data mining particularly in areas related to decision trees.

Using meteorological data to calculate evapotranspiration allows for the precise and timely scheduling for the irrigation of crops. The literature related to such studies indicate that by applying satellite images and different algorithms, evapotranspiration can be estimated over an extended area, thus allowing for the developing of an accurate irrigation schedule (Jaferian et al. 2019; Song et al. 2018; Goodarzi and Eslamian 2018; Diarraa et al. 2017; Colaizzi et al. 2017; Anderson et al. 2012; Abdolhosseini et al. 2012).

The estimation of the rate of evapotranspiration is a complicated process due to many independent parameters. Evapotranspiration is a dependent parameter which can be affected by many climatological parameters and crop conditions. In order to estimate evapotranspiration, a variety of equations have been developed which are used for different conditions such as the FAO Penman–Monteith equation, and the Blaney-Criddle formula, among others. Ground observations represent the results of one specific point, while great precision is required to generalize the observations over an
extended region; hence, evaporation data varies from one station to another. By applying remote sensing technologies, one may attain satisfactory and superior precision for a specific extended region; moreover, the use of satellite images within the framework of remote sensing allows for ground observations (hard data) to be transformed into soft data. Gibert et al. (2018) point out that among the various methods related to data mining, the M5 decision tree has been successfully applied to estimate the evapotranspiration over an extended area.

The current research is focused on establishing an innovative and highly applicable linear relationship using data mining techniques between independent remote sensing parameters (Albedo, emissivity, and normalized difference water index) and a dependent parameter (evapotranspiration) via a M5 decision tree.

Landsat 8 satellite images along with the SEBAL algorithm were used for estimating evapotranspiration in the study area, and the obtained data was further used for many accurate evapotranspiration estimations. (Mhawej et al. 2020; Elnmer et al. 2019; Kong et al. 2019; Ochege et al. 2019; Gobbo et al. 2019; Kamali and Nazari 2018).

One of the input parameters for estimating the rate of evapotranspiration is land surface temperature (LST); however, the spatial resolution of this band is 100 m. The application of the SEBAL algorithm along with this band length further provides an estimated evapotranspiration image with a 100 m spatial resolution. The aim of the current research is to enhance spatial resolution using the M5 decision tree since input parameters have a 30 m spatial resolution; thus, by applying the obtained equations through the M5 decision tree, an evapotranspiration map with a higher spatial resolution can be obtained.

In the Southwest of Iran, which is extremely dry and arid, an extremely high volume of water is consumed by the sugarcane plantations over an extended area (more than 94,000 ha); hence, by spatially enhancing the image resolution for evaluating evapotranspiration, optimal irrigation programming can be calculated more precisely.

2 Materials and methods

2.1 Study area

This study was conducted at the Amir-Kabir Agro-Industry Sugarcane plantation located in the Southwest of Iran (Fig. 1). The soil texture is clay-loam, and the annual average evapotranspiration for a 20 year period was 3331.812 mm. The total area of the Amir-Kabir agro-industry is over 17,000 ha of which 14,000 ha is cultivated. Each farm is 25 ha in extent, having a low-pressure hydro flume irrigation system along with a subsurface drainage system at 40 m spacing and 1.8 m depth for each drain tile. The total irrigation water consumption is 3,000 mm, and peak irrigation consumption occurs in July.

2.2 Landsat 8 satellite

Landsat 8 OLI satellite images provided the main input data for the remote sensing procedure (http://glovis.usgs.gov). The thermal bands however had a lower resolution as compared to other optic bands. In the case of Landsat 8, a band 10 image represents a thermal band that provides minimum spatial resolution (100 m). Thermal bands are critical for the estimation of the evapotranspiration rate, and Landsat8 provides the most appropriate thermal band for the estimating of agricultural evapotranspiration over an extended region. The 14,000 ha Amir-Kabir agro-industry plantation is extended enough for the calculation of evapotranspiration using remote sensing images.

2.3 Ground measurements

Any estimation of ET requires meteorological data; thus, the meteorological data applied in this study was obtained from the Amir-Kabir agro-industry plantation’s local weather station. The meteorological data used for the calculation of the evapotranspiration rate included max and min temperature, the relative percentage of humidity, wind speed, and hours of sunshine. Ref-et software was used for determining the reference points for the evapotranspiration calculations.

2.4 SEBAL algorithm

The SEBAL algorithm which is known as a one source algorithm for the calculation of evapotranspiration was used for calculating the rate of evapotranspiration in the sugarcane fields of the Amir-Kabir agro-industry plantation in this study. The energy balance equation is the primary equation used in this algorithm. Equation 1 represents the main energy balance.

$$\lambda ET = R_n - G - H$$  \hspace{1cm} (1)

where $\lambda ET$ is the latent heat flux (W/m²), $R_n$ is the net radiation at the surface (W/m²), $G$ is the soil heat flux (W/m²), and $H$ is the sensible heat flux to the air (W/m²) (Bastiaansen et al. 1998).

The net sun radiation ($R_n$) is a balance equation between the incoming and outgoing short and long wave components as shown in Eq. 2 (Bastiaansen et al. 1998):

$$R_n = (1 - \alpha)Rs \downarrow + RL \downarrow - RL \uparrow - (1 - \varepsilon_0)RL \downarrow$$  \hspace{1cm} (2)

where $Rs\downarrow$ is the incoming shortwave radiation (W/m²), $RL\downarrow$ is the incoming long wave radiations emitted by the
atmosphere (W/m²), $RL^\uparrow$ is the outcoming long wave radiations (W/m²), $\alpha$ is the surface Albedo also known as the reflection coefficient, and $\varepsilon_o$ is the broadband surface emissivity.

The incoming shortwave radiation $Rs^\downarrow$ is estimated from the radiation received at the top of the atmosphere and has been calculated as per the equation below: (Allen et al. 2002) (Eq. 3):

$$Rs^\downarrow = Gsc \times \cos \theta \times dr \times \tau_{sw}$$

where $Gsc$ is the solar constant values 1367 (W/m²), $\cos \theta$ is the cosine of the solar Zenith angle, $dr$ is the inverse relative squared distance of the Earth to the Sun, and $\tau_{sw}$ is the atmospheric transparency factor.

The Stephan-Boltzmann equation (Eq. 4) has been used to calculate the incoming long wave radiation emitted by the atmosphere (Allen et al. 2002):

$$RL^\downarrow = \varepsilon_x \times \sigma \times T^4$$

where $\sigma$ is the Stefan-Boltzmann constant 5.67e-8 W/m²K⁴, and $T$ is the surface temperature.
where ε is the atmospheric emissivity (dimensionless), σ is the Stefan-Boltzmann constant \( (5.67 \times 10^{-8} \text{ W/m}^2\text{/K}^4) \), and \( T_a \) is the air temperature in K.

The Stephan-Boltzmann equation has also been applied to estimate the long wave radiation emitted by the surface (Allen et al. 2002) (Eq. 5):

\[
RL \uparrow = \varepsilon_0 \times \sigma \times T_s^4
\]

where \( \varepsilon_0 \) is the broadband surface emissivity (dimensionless), \( \sigma \) is the Stefan-Boltzmann constant \( (5.67 \times 10^{-8} \text{ W/ m}^2\text{/K}^4) \), and \( T_s \) is the surface temperature in K. The amount of the Earth’s surface temperature was calculated through Eq. 6:

\[
T_s = \frac{K_2}{ln(\frac{\varepsilon_0 \times K_1}{\rho_{10}}) + 1}
\]

where \( K_1 \) and \( K_2 \) are the constant coefficients of Landsat 8 image equal to 774.88 and 1321.0789, \( \varepsilon_0 = 1 \) (Allen et al. 2002), and \( \rho_{10} \) is the thermal correction band (band 10) of Landsat 8 image.

The amount of the Albedo coefficient was calculated using Eq. 7:

\[
a = \frac{\alpha_{tot} - \alpha_{path-radiance}}{\tau_{sw}}
\]

where \( \alpha_{path-radiance} \) is the average portion of the incoming solar radiation taking into account all bands that are back-scattered to the satellite before it reaches the earth’s surface and \( \tau_{sw} \) is the atmospheric transmissivity (Allen et al. 2002).

Soil heat flux (G) is the other main parameter for the energy balance equation which represents the heat storage in the equilibrium (Bastiaanssen et al. 1998). G and Rn ratio have a constant value, and by dividing G on Rn, the best obtained ratio is between 0.3 and 0.6. Eq. 8 is used for calculating G (Bastiaanssen et al. 1998):

\[
\frac{G}{R_n} = \frac{T_s \times R_n}{\alpha} \left( \frac{0.0032\alpha + 0.006\alpha^2}{\tau_{sw}^2} \right) \times \left( 1 - 0.978 \text{NDVI}^4 \right)
\]

where \( T_s \) is the surface temperature, while the vegetation index NDVI (normalized difference vegetation index) is used for G calculations using Eq. 9 (Allen et al. 2002).

\[
\text{NDVI} = \frac{\rho_{NIR} - \rho_{R}}{\rho_{NIR} + \rho_{R}}
\]

where near infrared band and the infrared band are used for NDVI calculations and shown as \( \rho_{NIR} \) and \( \rho_{R} \) in Eq. 9, respectively. It was observed that the NDVI of the study area is more than zero due to plant cultivation.

The amount of NDVI changes between −1 and +1. Area with vegetation has NDVI between zero and one; water and clouds usually have NDVI less than zero. If the NDVI value is less than zero, then the area is assumed as water and the \( G/R_n \) ratio is considered to be 0.5. Areas with \( T_s \) less than 4°C and \( \alpha \) greater than 0.45 are considered as snow covered areas and the \( G/R_n \) ratio for these areas is considered to be 0.5 (Allen et al. 2002).

In the SEBAL method, for the estimation of the apparent heat flux (H), two pixels are selected. First, one of these pixels called cold pixels is related to a completely covered area of the irrigated area and plants. The surface temperature of this pixel is close to the air temperature and evapotranspiration equals to reference evapotranspiration. The second pixel, is called the hot pixel, is land without vegetation and dry, so the amount of flowing heat of evaporation in this pixel is assumed to be zero (Sanaei Nezhad et al. 2011).

In SEBAL method, according to Eq. 1 and based on the evapotranspiration values of the two pixels, the apparent heat flux of other pixels is estimated. The value of the heat flux is calculated from Eq. 10 (Allen et al. 2002).

\[
H = \frac{\rho_{air} \times C_{air} \times dT}{r_{ah}}
\]

where \( \rho_{air} \) represents the air density (kg/m3), \( C_{air} \) represents air specific heat (1004 J/kg/K), \( dT \) (K) represents the temperature difference (\( T_1 - T_2 \)) between two heights (\( Z_1 \) and \( Z_2 \)), and \( r_{ah} \) represents the aerodynamic resistance to heat transport (m/s).

Aerodynamic resistance to heat transport is calculated by applying Eq. 11:

\[
r_{ah} = \frac{ln(\frac{Z_2}{Z_1})}{u_s \times k}
\]

\( Z_1 \) and \( Z_2 \) are the two heights; \( k \) is the Von Karman constant (0.41) and \( u_s \) represents friction velocity which is calculated through Eq. 12:

\[
u_s = \frac{k \times u_{200}}{\ln(\frac{200}{Z_{som}})}
\]

\( u_{200} \) (blended wind speed) is the wind speed measured at a height of 200 m at the weather station. \( Z_{som} \) is empirically estimated from the vegetation height, and \( Z_{som} \) is calculated through the leaf area index Eq. 13:

\[
z_{som} = 0.018 \text{LAI}
\]

The LAI using soil-adjusted vegetation index and with Eq. 14 is determined.

\[
\text{LAI} = \frac{\ln(0.69 - \text{SAVI})}{0.59}
\]

\[
\text{SAVI} = \frac{\ln(0.69 - \text{SAVI})}{0.59}
\]
The $\text{LAI}$ (Eq. 14) is calculated using the $\text{SAVI}$ (soil-adjusted vegetation index) (Eq. 15):

$$\text{SAVI} = \frac{(1 + L)(\text{NIR} - R)}{(\text{NIR} + R + L)}$$  \hspace{1cm} (15)

where $L$ is a correction factor for the background soil.

The surface temperature difference which is used in Eq. 10 is a linear relation between $dT$ and $T_s$ which is presented in Eq. 16:

$$dT = aT_s + b$$  \hspace{1cm} (16)

where $dT$ is the near-surface air temperature difference, $T_s$ is the surface temperature, and $a$ and $b$ are empirical coefficients.

In the main equation (Eq. 1), the term latent heat flux for each pixel is calculated by Eq. 17:

$$\text{ET}_{\text{ins}} = \frac{3600 \lambda_{\text{ET}}}{\lambda}$$  \hspace{1cm} (17)

where $\lambda$ is the latent heat of vaporization (J/kg) and can be computed as Eq. 18:

$$\lambda = [0.2501 - 0.00236(T_s - 273)] \times 10^6$$  \hspace{1cm} (18)

The main aim of using the SEBAL algorithm is to obtain $\text{ET}_{24}$ for each pixel using Eq. 19 (Allen et al. 2002):

$$\text{ET}_{24} = \text{ET}_{12} \times \text{ET}_{12-24}$$  \hspace{1cm} (19)

where the $\text{ET}_{24}$ is the total $\text{ET}$ during a 24-h period of the same day.

For this study, two cloud-free satellite images were obtained for 8 April 2019 and 3 April 2020. Actual evapotranspiration ($\text{ET}a$) maps in mm/day are generated by the SEBAL algorithm for each day.

### 2.5 Data mining

Data mining (DM) algorithms are the most fundamental components in data science analysis. Gibert et al. (2018) states that certain DM techniques such as artificial neural networks, clustering, and case-based reasoning or Bayesian networks have been successfully applied in environmental modeling.

The decision tree method applies explanatory variables with higher discriminatory power while considering the response variable and then iteratively subdivides the training sample by building a tree where the internal nodes are associated with the input variables and its corresponding branches as the possible values of the variable (Gibert et al. 2018). The M5 Model Tree (introduced by Quinlan in 1992) has linear regression functions at the leaf nodes, which in themselves create a relationship between input and output variables. The data is then split into subsets, and a decision tree is created. The data in offspring nodes from splitting criterion depends on the treating of the standard deviation of the class values and the calculating of the expected reduction in this error as a result of testing each attribute within that node. The standard deviation reduction (SDR) is calculated as per Eq. 20 (Quinlan 1992):

$$\text{SDR} = sd(T) - \sum_i \frac{|T_i|}{T} \times sd(T_i)$$  \hspace{1cm} (20)

where $T$ is a set of data that reaches the node, $T_i$ is the subset of data that has the $i$th outcome of the potential set, and $sd$ is the standard deviation (Rahimikhoob et al. 2013; Wang and Witten 1997). The data in the offspring nodes are purer due to a less standard deviation in comparison to parent nodes. The M5 tree selects the node that maximizes the expected error reduction after scanning all the possible splits.

Every inner node of a tree has the capability of producing multiple linear regression models using the data associated with that node and all the attributes that are utilized in tests in the sub-tree rooted at that node. The linear regression models are simplified providing that the results have a lower expected error (Etemad-Shahidi and Bonakdar 2009). Figure 2 shows an M5 decision tree structure for two input parameter domains of X1 and X2 with 4 linear models from Y1 to Y4.

### 2.6 Model inputs and output

For estimating the amount of evapotranspiration using the SEBAL algorithm, meteorological data including temperature, humidity, and wind speed, among others are needed. Some of the inputs used in the SEBAL algorithm such as Albedo and emissivity are affected by land surface temperature; thus, in order to minimize the number of variables in data mining, Albedo and emissivity are considered as inputs of the M5 decision tree since these two parameters are easily accessible and display the temperature variances better than most; furthermore, transpiration depends on the moisture of the plant. In data mining calculations, a vegetation index such as the normalized difference water index (NDWI) is used to represent plant moisture; consequently, in the calculations, Albedo and emissivity are defined as absorbed and transformed light to the atmosphere, and NDWI is represented as plant moisture. The basic SEBAL equation was then applied as $\text{ET} = \text{Rn} - \text{G} - \text{H}$ in which the following values replaced the formulas: $\text{ET} = a(\text{Albedo}) - b(\text{emissivity}) - c(\text{NDWI})$ and calculated using constant values by applying the M5 decision tree model. The three inputs of the M5 decision tree are explained hereinafter in greater detail.
2.6.1 Evapotranspiration

One of the crucial parameters for the estimating of evapotranspiration is land surface temperature (LST), a factor upon which radiation and the exchange of energy flux between the earth’s surface and atmosphere depend on. (Weng et al. 2019). Physical models such as the SEBAL algorithm has made it possible to estimate evapotranspiration for large areas. Weng et al. (2019) indicate that among the most important parameters affecting LST are surface biophysical characteristics such as Albedo, greenness, and wetness. The rate of energy distribution is determined through Albedo and emissivity of the surface and atmosphere values. Previous studies show that surface emissivity strongly correlates to vegetation cover (Griend and Owe 1993; Rechid et al. 2009); in addition, vegetation also strongly affects atmospheric properties through evapotranspiration (Gordon et al. 2005). Atmosphere emissivity is in itself determined by atmospheric water vapor pressure (Staley and Jurica 1972; Brutsaert 1975).

2.6.2 Albedo

Zhang et al. (2017) defines Albedo as a dimensionless diffuse reflectivity or reflecting power of a surface and considers it as an important effective parameter for digital climate models and surface energy balance equations. Surface Albedo is computed by correcting the $\alpha_{\text{toa}}$ for atmospheric transmissivity from Eq. 21. This parameter is considered as one of the decision tree inputs for calculating evapotranspiration.

$$\alpha = \frac{\alpha_{\text{toa}} - \alpha_{\text{path,radiance}}}{\tau^2_{\text{sw}}}$$  \hspace{1cm} (21)

where $\alpha_{\text{path, radiance}}$ is the average portion of the incoming solar radiation by considering all bands that is back-scattered to the satellite before it reaches the earth’s surface and $\tau_{\text{sw}}$ is the atmospheric transmissivity (Allen et al. 2002).

2.6.3 Emissivity

The surface emissivity is the ratio of the actual radiation emitted by a surface via a black body with the same surface temperature (Allen et al. 2002). Surface emissivity is an important variable for estimating land surface temperature and determining long wave surface energy balance (Mira et al. 2010). Sobrino et al. (2004) proposed emissivity-based NDVI in three different cases as per Eq. 22:

$$\epsilon = \begin{cases} 0.973 \quad & \text{NDVI} < 0.2 \\ \epsilon_v P_v + \epsilon_s (1 - P_v) + d\epsilon \quad & 0.2 \leq \text{NDVI} \leq 0.5 \\ 0.986 \quad & \text{NDVI} > 0.5 \end{cases}$$  \hspace{1cm} (22)

where $\epsilon_v$ is the vegetation canopy emissivity and $\epsilon_s$ is the bare soil emissivity in this paper $\epsilon_v = 0.986$ and $\epsilon_s = 0.973$. The effects of the geometrical distribution of the natural surfaces are measured as $d\epsilon$ in Eq. 6. $P_v$ is the vegetation proportion obtained according to Carlson and Ripley (1997) as per Eq. 23:

$$P_v = \left[ \frac{\text{NDVI} - \text{NDVI}_s}{\text{NDVI}_v - \text{NDVI}_s} \right]^2$$  \hspace{1cm} (23)

The emissivity of land surfaces can differ significantly due to vegetation, surface moisture, and roughness (Nerry et al. 1988, Salisbury and D’Arî 1992). The minimum value of the NDVI for bare soil over the study region is presented as NDVI$_s$, while NDVI$_v$ is the highest NDVI for a fully vegetated pixel.
2.6.4 Vegetation index

The NDWI spectral index which represents crop moisture is the normalized difference water index. NDWI was used to estimate the equivalent water thickness of the vegetation canopy (Yilmaz et al. 2008). The NDWI considers two infrared bands with a central wavelength near about 0.86 μm (NIR) and a central wavelength of about 1.24 μm (SWIR). The calculation is represented as Eq. 24:

\[
NDWI = \frac{\rho(0.86\mu m) - \rho(1.24\mu m))}{\rho(0.86\mu m) + \rho(1.24\mu m))}
\]  

(24)

The M5 decision tree model uses Albedo, emissivity, and the vegetation index as input, and after performing the data mining process on the said data, it extracts the linear equations. After inserting the linear equations, a highly enhanced special resolution map was obtained as output. Figure 3 shows the flowchart of the M5 decision tree and the SEBAL algorithm.

2.7 Model validation

2.7.1 Statistical analysis

The application of the three inputs (Albedo, emissivity, and NDWI) allows for the evaluation of the M5 decision tree. The accuracy of the M5 decision tree model and the final evapotranspiration map which was integrated with the M5 decision tree was evaluated through the use of statistical parameters such as the correlation coefficient \( R^2 \), root mean squared errors \( RMSE \), and mean absolute errors \( MAE \) (Eqs. 25 to 27):

\[
R^2 = 1 - \frac{\sum (ET_o - ET_p)^2}{\sum (ET_o - \bar{ET})^2}
\]  

(25)

\[
RMSE = \sqrt{\frac{1}{n} \sum (ET_o - ET_p)^2}
\]  

(26)

\[
MAE = \frac{1}{n} \sum \left| ET_o - ET_p \right|
\]  

(27)

where \( N \) is the number of data, \( ET_o \) is the observed evaporation values calculated via the SEBAL algorithm, and \( ET_p \) is the M5 decision tree model for the amount of evapotranspiration.

2.7.2 Uncertainty analysis

Uncertainty analysis was calculated from the percentage of observed data through the application of the 95 PPU (95 percent prediction uncertainty) and the average distance \( d^- \) between the upper and the lower 95 PPU (or the degree of uncertainty) using Eq. 28 (Abbaspour et al. 2007):

\[
d^- = \frac{1}{k} \sum_{i=1}^{k} (X_U - X_L)
\]  

(28)
where \( k \) is the number of observed data, \( X_L \) is the 2.5th, and \( X_U \) is the 97.5th percentiles of the cumulative distribution of every estimated data.

If 100% of the observed data are bracketed by the 95PPU and \( d \) is close to zero, the results will be in an acceptable range of uncertainty analysis; hence, utilizing Eq. 29, the \( d - \text{factor} \) has been calculated.

\[
d - \text{factor} = \frac{d}{\sigma}
\]

where \( \sigma \) is the standard deviation of the measured variable and the obtained is \( d - \text{factor} \) is less than 1 (Abbaspour et al. 2007).

### 2.7.3 Sensitivity analysis

The sensitivity coefficient is a dimensionless index \( (S) \) which is calculated by taking the ratio of the change in output to input into account conditional to the other variables remaining constant. The sensitivity of a dependent variable (evapotranspiration) to a particular independent variable (Albedo, emissivity, and NDWI) can be calculated from the derivative of evapotranspiration with an independent variable that is \( \delta \text{ETA}/\delta X \) (Beven 1979; McCuen 1974). To evaluate the sensitivity of a variable, \( S \) was divided into four classes from least to greatest sensitivity (Table 1). If the range attributed to \( S \) was at a higher level, the independent variable would have a greater impact on the dependent variable.

#### 2.7.3.1 Results

### 2.8 M5 decision tree

Instead of using a physical-based model for the estimating of the amount of evapotranspiration, the M5 decision tree was applied to a data mining model which could increase the spatial resolution of the evapotranspiration map. Using the data mining model, several equations were obtained for estimating the amount of evapotranspiration over a single day (8 April 2019). The extracted equations were then applied for estimating the amount of evapotranspiration during the following year (3 April 2020). The values obtained were then compared with the amount of evapotranspiration calculated by the SEBAL algorithm on that particular day. The results are further discussed in this study.

### Table 1 Level of sensitivity

| Sensitivity Coefficient | Sensitivity levels |
|-------------------------|-------------------|
| 0.00 ≤ |S| < 0.05         | Small            |
| 0.05 ≤ |S| < 0.20         | Medium           |
| 0.20 ≤ |S| < 1.00         | High             |
| |S| ≥ 1.00         | Very high        |

### 2.8.1 Inputs

The input variables of the M5 decision tree consisted of four satellite images which applied the SEBAL algorithm to calculate Albedo, emissivity, normalized difference water index (NDWI), and estimated evapotranspiration.

Figure 4 shows the Albedo maps for 8 April 2019 and 3 April 2020. According to Fig. 4, the calculated Albedo for 8 April 2019 is higher than the Albedo for 3 April 2020 because of the existing drought conditions in 2019 which affect Albedo since it is influenced itself by atmospheric vapor and temperature (Feng and Zou 2019).

Figure 5 shows the obtained emissivity maps for 8 April 2019 and 3 April 2020. Emissivity is the most important input variable for the M5 decision tree because this variable is required for land surface temperature calculations and is affected by the percentage of vegetation.

According to Fig. 5, both maps have approximately the same maximum and minimum range, yet the amount of emissivity varies on the same farm when comparing both years. Emissivity variations depend on the extent and type of cultivation of farms. In April 2019, based on the type of agricultural activities in the area and managerial decisions taken, the western part of the plantation had more cultivated sugarcane (Fig. 5a), and in the next year (Fig. 5b), there was less cultivated sugarcane which most probably affected the amount of emissivity.

### 2.8.2 Output

The values in parentheses under each label in the leaves of the decision tree indicate the number of segments resulting from the corresponding threshold. The second value
indicates the number of times a misclassification occurred (Vieira et al. 2012).

Figure 7 shows the decision tree for the evapotranspiration of 8 April 2019. Twenty different equations were extracted using a correlation coefficient of 0.9429, mean absolute error, and root mean squared error of 0.4749 and 0.6479, respectively.

Through the application of minimum input parameters, specifically Albedo, emissivity, and NDWI, a number of evapotranspiration equations were derived. For the evapotranspiration of 8 April 2019, 20 conditional equations were calculated. According to Fig. 7, the Albedo input variable was located at the top of the decision tree, and the divisions were based on Albedo values, therefore, indicating that Albedo is highly important in the calculating of the amount of evapotranspiration. In the first leaf (first equation), divisions are based on Albedo; as a consequence, as per Fig. 7, the majority of equations become derivatives of Albedo. Taking the geographical location of the study area into account, it can be observed that there is a high amount of received light in the area, and the Albedo input variable is thus defined as absorbed light. The M5 decision tree divisions based on Albedo show that the amount of absorbed light in this area has a major role in the calculation of the evapotranspiration rate in the area.

The NDWI variable is defined as plant moisture which after the Albedo variable plays a major role in evapotranspiration calculations. This particular variable shows that plant moisture in addition to absorbed light is essential in producing the decision tree equations.

The emissivity variable being defined as diffused light had lesser significance in the decision tree derivatives with due regards to the geographical location of the study area.

The obtained equations as a result of applying the M5 decision tree and python scripts in an ArcMap environment are presented as Appendix 1 and 2 respectively at the end of this article.

2.9 Combining M5 and GIS

After deriving the most appropriate equations from the M5 decision tree model, the equations were applied using python scripts for faster and more accurate calculations (available through email by the corresponding author). The equations
were obtained using evapotranspiration data from 8 April 2019 and were applied on input variables from the said year to identify whether the extracted equations have an acceptable performance. Figure 8 shows the evapotranspiration map calculated by the SEBAL algorithm and the M5 decision tree for 8 April 2019 and 3 April 2020, respectively. According to Fig. 8a and c, the evapotranspiration maps for 2019 did not have many differences in the calculated evapotranspiration amounts. Figure 8b shows the calculated evapotranspiration rate using the SEBAL algorithm for 2020, and Fig. 8d shows the calculated evapotranspiration using the M5 decision tree equation extracted from the 2019 SEBAL evapotranspiration map. A comparison of Fig. 8b and d shows that the mined equations from the M5 decision tree of 2019 when applied on the input variables of 3 April 2020 provide acceptable results due to a minimum disparity between the evapotranspiration map calculated by the SEBAL algorithm and that calculated using the M5 decision tree for 2020.

Figure 9 shows the results of the comparison between the SEBAL algorithm and the M5 decision tree for 8 April 2019 and 3 April 2020. Table 2 shows the statistical coefficients for 8 April 2019 and 3 April 2020. According to Fig. 9 and Table 2 when comparing the obtained results for the two different years, it is possible to calculate the amount of evapotranspiration using lesser input variables while focusing on the most important variables due to high correlation coefficients and low errors. As per Table 2, the calculated rate of evapotranspiration for 3 April 2020 shows more improved results than 8 April 2019. This is due to the differences between input variables over the 2 years which cause the derived equations to show better accuracy for 3 April 2020.

2.10 Model evaluation

Using an uncertainty analysis in tandem with a sensitivity analysis, the rate of evapotranspiration obtained using the data mining model was evaluated. The uncertainty analysis of the obtained model was calculated using two criteria that are 95PPU and the d-factor. By increasing the observed data in the 95PPU level while decreasing the average values of the upper and lower bands (less than the standard deviation), while taking into consideration that the bracketed value 95PPU must be in the maximum range, it was observed that

![Emissivity maps of a 8 April 2019 and b 3 April 2020](image-url)
the data (calculated by the SEBAL algorithm) is between the lower and the upper limits of uncertainty defined as 2.5% and 97.5% (XL and XU) levels of the cumulative
Fig. 8  Evapotranspiration map calculated by SEBAL algorithm for a 8 April 2019 and b 3 April 2020 and evapotranspiration map calculated by M5 decision tree for c 8 April 2019 and d 3 April 2020
distribution of the output variables. When 80% of the calculated data (calculated through the data mining model) are within the 95PPU level, it clearly indicates that they are of a higher quality.

Table 2 shows the uncertainty coefficient for 8 April 2019 and 3 April 2020. As observed in the table, the obtained d-factor is less than 1 for both images, and the 95PPU is more than 80%. The calculated uncertainty coefficients show
that the data mining model has a superior quality for estimating the amount of evapotranspiration using less variables and thus can be applied for calculating the amount of evapotranspiration with acceptable certainty.

Table 4 shows that the sensitivity coefficient of the variables vary. The Albedo variable has the greatest sensitivity for calculating evapotranspiration in both images from 8 April 2019 to 3 April 2020. The study area is located in a region which is approximately at sea level and receives a great deal of solar radiation. The net radiation (Rn) is among the major variables for calculating the amount of evapotranspiration using the SEBAL algorithm; moreover, Albedo is capable of representing this variable as one of the data mining model inputs. Moreover, the decision tree obtained for this research (Fig. 7) shows that most of the classifications that were developed derived from the Albedo variable which was incorporated within the nodes of the decision tree which clarifies the major role that Albedo has in deriving the equations due to its high sensitivity. The other two parameters have less sensitivity in the calculation of the evapotranspiration rate for 2020. The NDWI has less sensitivity due to the late start of irrigating the cultivated sugarcane fields as compared with the previous year, i.e., 2019.

### 3 Discussion

Climatological variables which are considered as a physical process can affect the calculations for determining the amount of evapotranspiration. Measuring evapotranspiration over an extended area using remote sensing is a practical procedure and can be done utilizing physical-based algorithms like SEBAL. By transforming a physical-based model to a data mining model, the input variables can be decreased in as such that only the major variables are applied. The utilizing of a M5 decision tree allows for the transforming of a physical-based evapotranspiration algorithm into a data mining model. The three inputs of the decision tree consisted of Albedo, emissivity, and NDWI. The target function was the evapotranspiration value obtained by the SEBAL algorithm. Moreover, the possibility of calculating the amount of evapotranspiration using linear regression equations obtained from the M5 decision tree was considered. In this case, a complicated evapotranspiration process which was dependent on so many variables showed that by applying a simple linear regression, the deriving of equations can be carried out. Another objective of the current research is the applying of the general equation: \( ET = a(\text{Albedo}) - b(\text{emissivity}) - c(\text{NDWI}) \) from which the constant values were obtained using data mining procedures. Based on the basic SEBAL algorithm equation which is represented as the energy balance: \( \lambda ET = Rn - G - H \), evapotranspiration is calculated as absorbed energy (net radiation at the surface) subtracted from transmitted energy (soil heat flux and sensible heat flux to the air), so instead of using energy labels (heat fluxes), light labels were used in the main equation, and evapotranspiration was calculated as the absorbed light (Albedo) minus transmitted light (emissivity) on the basis of the theory that light is another form of energy. Therefore, through the applying of the M5 decision tree, the most acceptable value was obtained for calculating the rate of evapotranspiration using data mining methods.

### 4 Conclusion

The main objective of the current research was to modify a physical-based model via a data mining model since physical-based models need more input variables in comparison to data mining models which use fewer input variables yet obtain acceptable results. Using three input variables that is Albedo, emissivity, and NDWI, mathematical equations were extracted from the M5 decision tree. These three variables were selected from the values obtained by the basic SEBAL algorithm equation: \( ET = Rn - G - H \). The main concept is to calculate evapotranspiration as \( ET = a(\text{Albedo}) - b(\text{emissivity}) - c(\text{NDWI}) \) in which the constant values were calculated from the M5 decision tree model. The extracted equations were then applied on 8 April 2019, and the evapotranspiration calculations are nearly parallel to the SEBAL algorithm evapotranspiration values. Furthermore, the extracted equations from 2019 were applied for 3 April 2020 which when compared to the SEBAL evapotranspiration algorithm paralleled it accurately, making it an acceptable result; however, when applying data mining methods instead of SEBAL, the results obtained show a lower accuracy. In conclusion, evapotranspiration can be calculated over an extended area of the SEBAL algorithm, in tandem with a data mining model, and M5 decision tree utilizes fewer inputs; furthermore, the research has also shown that light labels can be applied instead of energy labels.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00704-021-03822-7.
Evaluating evapotranspiration using data mining instead of physical-based model in remote…

Acknowledgements We are grateful to the Research Council of Shahid Chamran University of Ahvaz for financial support (GN: SCU. WI98.281). Satellite images were supported through USGS earth explorer website. The authors would like to thank the NASA agency for free access images to implementing this research.

Author contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Lamya Neissi, Mona Golabi, Mohammad Albaji, and Abd Ali Naseri. The first draft of the manuscript was written by Lamya Neissi, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. Mohammad Albaji is the corresponding author of this manuscript.

Funding This study was funded by “Shahid Chamran University of Ahvaz.”

Data availability Data and material including of satellite images, evapotranspiration maps, and obtained results will be available as soon as one of the authors receives a request email.

Code availability Python scripts for the M5 decision tree will be available by sending a request email to neissi.lamya@gmail.com.

Declarations

Ethics approval The paper is not currently being considered for publication elsewhere.

Consent to participate Not applicable.

Consent for publication This research was done under the support of “Sugarcane Research and Training Institute.” The institute has no disagreement for publishing this research.

Conflict of interest The authors declare no competing interests.

References

Abbaspour KC, Yang J, Maximov I, Sibor R, Bogner K, Mieleitner J, Zobrist J, Srinivasan R (2007) Modelling hydrology and water quality in the pre-alpine/ alpine Thur watershed using SWAT. J Hydro 333(2–4):413–430

Abdelhosseini M, Eslamian S, Mousavi SF (2012) Effect of climate change on potential evapotranspiration: a case study on Gharrehsou sub-basin. Iran 2(4):362–372

Allen R, Tasumi M, Trezza R (2002) SEBAL (Surface Energy Balance Algorithms for Land)-Advanced Training and User’s Manual-Idaho Implementation. Version 1.0.

Anderson MC, Allen RG, Morse A, Kustas WP (2012) Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. Remote Sens Environ 122:50–65

Bastaiaans W, Menenti M, Feddes R, Holtslag A (1998) A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation J Hyd 212:198–212

Beven K (1979) A sensitivity analysis of the Penman-Monteith actual evapotranspiration estimates. J Hydrol 44(3–4):169–190

Brutsaert W (1975) On a derivable formula for long-wave radiation from clear skies. Water Resour Res 11:742–744

Carlson TN, Ripley DA (1997) On the relation between NDVI, fractional vegetation cover, and leaf area index. Remote Sens Environ 62:241–252

Colaiazzi PD, O’Shaughnessy SA, Evetta SR, Mounce RB (2017) Crop evapotranspiration calculation using infrared thermometers aboard center pivots. Agri Water Manag 187:173–189

Diarraa A, Jarlan L, Er-Rakis C, Le Pageb M, Aouade G, Tavernierb A, Bouleb B, Ezzahard J, Merlioba O, Khabbaas S (2017) Performance of the two-source energy budget (TSEB) model for the monitoring of evapotranspiration over irrigated annual crops in North Africa. Agri Water Manag 193:71–88

Elnmer A, Khadr M, Kanae S, Tawfik A (2019) Mapping daily and seasonally evapotranspiration using remote sensing techniques over the Nile delta. Agri Water Manag 213: 682–692. ISSN 0378–3774, https://doi.org/10.1016/j.agwat.2018.11.009

Etemad-Shahidi A, Bonakdar L (2009) Design of rubble-mound breakwaters using M5 ’ machine learning method. Appl Ocean Res 31(3):197–201

Feng H, Zou B (2019) A greening world enhances the surface-air temperature difference. Sci Tot Environ 658:385–394

Gilbert K, Izquierdo J, Sánchez-Marré M, Hamilton SH, Rodríguez-Roda I, Holmes G (2018) Which method to use? An assessment of data mining methods in Environmental Data Science. Environ Model Software 110:3–27. https://doi.org/10.1016/j.envsoft.2018.09.021

Gobbo S, Lo Presti S, Martello M, Panunzi L, Berti A, Morari F (2019) Integrating SEBAL with in-field crop water status measurement for precision irrigation applications—a case study. Remote Sens 11:2069

Goodarzi M, Eslamian S (2018) Performance evaluation of linear and nonlinear models for the estimation of reference evapotranspiration. Int J Hydro Sci Tech 8(1):1–15

Gordon LJ, Steffen W, Jönsson BF, Folke C, Falkenmark M, Johannessen A (2005) Human modification of global water vapor flows from the land surface. Proc Natl Acad Sci USA 102:7612–7617

Griend AA, Ove M (1993) On the relationship between thermal emissivity and the normalized difference vegetation index for natural surfaces. Int J Remote Sens 16:1119–1131

Jaferian V, Toghaiae D, Pourfattah F, Ali Akbari O, Talebizadehsardari P (2019) Numerical investigation of the effect of water/Al O nanofluid on heat transfer in trapezoidal, sinusoidal and stepped microchannels. INT J NUMER METHOD H. https://doi.org/10.1108/HFF-05-2019-0377

Kamali MI, Nazari R (2018) Determination of maize water requirement using remote sensing data and SEBAL algorithm. Agri Water Manag 209:197–205. ISSN 0378–3774, https://doi.org/10.1016/j.agwat.2018.07.035

Kong J, Hu Y, Yang L, Shan Z, Wang Y (2019) Estimation of evapotranspiration for the blown-sand region in the Ordos basin based on the SEBAL model. Int J Remote Sens 40(5–6):1945–1965. https://doi.org/10.1080/01431161.2018.1508919

McCuen RH (1974) A sensitivity and error analysis of procedures used for estimating evaporation. J Am Water Resour Assoc 10(3):486–497

Mhawej M, Elias G, Nasrallah A, Faour G (2020) Dynamic calibration for better SEBAL ET estimations: validations and recommendations. Agri Water Manag 230: 105955. ISSN 0378–3774, https://doi.org/10.1016/j.agwat.2019.105955.

Mira M, EValor E, Caselles V, Rubio E, Coll C, Galve JM, Niclos R, Sanchez JM, Boluda R (2010) Soil moisture effect on thermal infrared (8–13-μm) emissivity. IEEE Trans Geosci Remote Sens 48(5):2251–60. https://doi.org/10.1109/TGRS.2009.2039143

Nerry F, Laved J, Stoll MP (1988) Emissivity signatures in the thermal IR band for remote sensing: calibration procedure and method of measurements. Appl Optics 27:758–764
Ochege FU, Luo G, Obeta MC, Owusu G, Duulatov E, Cao L, Nsengiyumva JB (2019) Mapping evapotranspiration variability over a complex oasis-desert ecosystem based on automated calibration of Landsat 7 ETM+ data in SEBAL. Gisci Remote Sens 56(8):1305–1332. https://doi.org/10.1080/15481603.2019.1643531

Quinlan JR (1992) Learning with continuous classes. In: Proceedings of Australian Joint Conference on Artificial Intelligence (Singapore: World Scientific Press). 343–348

Rahimikhoob A, Asadi M, Mashal M (2013) A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. Water Resour Manag 27(14):4815–4826

Rechid D, Raddatz TJ, Jacob D (2009) Parameterization of snow-free land surface albedo as a function of vegetation phenology based on MODIS data and applied in climate modelling. Theor Appl Climatol 95:245–255

Salisbury W, D’Arí DM (1992) Emissivity of terrestrial materials in the 8–14 mm atmospheric window. Remote Sens Environ 42:83–106

Sanaei Nezhad SH, Noori S, Hasheminia SM (2011) Estimation of evapotranspiration using satellite image data in Mashhad area. Journal of Water and Soil (Agricultural Sciences and Technology 25(3):540–547

Sobrino JA, Jiménez-Muñoz JC, Paolini L (2004) Land surface temperature retrieval from LANDSAT TM5. Remote Sens Environ 90:434–440

Song L., Liu S., Kustas WP, Nieto H., Sun L., Xu Z., Skaggs TH, Yang Y., Ma M., Xu T., Tang X., Li Q. (2018) Monitoring and validating spatially and temporally continuous daily evaporation and transpiration at river basin scale. Remote Sens Environ 219:72–88

Staley DO, Jurica GM (1972) Effective atmospheric emissivity under clear skies. J Appl Meteorol 11:349–356

Vieira MA, Formaggio AR, Renô CD, Atzberger C, Aguiar DA, Mello MP (2012) Object based image analysis and data mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. Remote Sens Environ 123:553–562

Wang Y., Witten IH (1997) Induction of model trees for predicting continuous classes. In: Proceedings of the Poster Papers of the European Conference on Machine Learning. University of Economics, Faculty of Informatics and Statistics, Prague

Weng Q., Karimi Firozjaei M., Kiavarz M., Alavipanah SK., Hamzeh S. (2019) Normalizing land surface temperature for environmental parameters in mountainous and urban areas of a cold semi-arid climate. Sci Tot Environ 650:515–529

Yilmaz MT., Hunt ER., Jackson TJ. (2008) Remote sensing of vegetation water content from equivalent water thickness using satellite imagery. Remote Sens Environ 112(5):2514–2522

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.