agronomy

MDPI

*Article*

# Machine Learning and Conventional Methods for Reference Evapotranspiration Estimation Using Limited-Climatic-Data Scenarios

**Pietros André Balbino dos Santos [1], Felipe Schwerz [1,*], Luiz Gonsaga de Carvalho [1], Victor Buono da Silva Baptista [2], Diego Bedin Marin [3], Gabriel Araújo e Silva Ferraz [1], Giuseppe Rossi [4], Leonardo Conti [4] and Gianluca Bambi [4]**

[1] Agricultural Engineering Department, Federal University of Lavras, Lavras 37203-202, Brazil; pietros.balbino@gmail.com (P.A.B.d.S.); lgonsaga@ufla.br (L.G.d.C.); gabriel.ferraz@ufla.br (G.A.e.S.F.)
[2] Engineering Department, Federal University of Lavras, Lavras 37203-202, Brazil; victor.buonosb@ufla.br
[3] Agricultural Research Company of Minas Gerais (EPAMIG), Viçosa 36571-000, Brazil; db.marin@hotmail.com
[4] Department of Agriculture, Food, Environment and Forestry, University of Florence, 50121 Florence, Italy; giuseppe.rossi@unifi.it (G.R.); leonardo.conti@unifi.it (L.C.); gianluca.bambi@unifi.it (G.B.)
* Correspondence: felipe.schwerz@ufla.br

**Abstract:** Reference evapotranspiration ($ET_0$) is one important agrometeorological parameter for hydrological studies and climate risk zoning. $ET_0$ calculation by the FAO Penman–Monteith method requires several input data. However, the availability of climate data has been a problem in many places around the world, so the study of scenarios with different combinations of climate data has become essential. The aim of this study was to evaluate the performance of artificial neural network (ANN), random forest (RF), support vector machine (SVM), and multiple linear regression (MLR) approaches to estimate monthly mean $ET_0$ with different input data combinations and scenarios. Three scenarios were evaluated: at the state level, where all climatological stations were used (Scenario I–SI), and at the regional level, where the Minas Gerais state was divided according to the climatic classifications of Thornthwaite (Scenario II–SII) and Köppen (Scenario III–SIII). ANN and RF performed better in $ET_0$ estimation among the models evaluated in the SI, SII, and SIII scenarios with the following data combinations: (i) latitude, longitude, altitude, month, mean, maximum and minimum temperature, and relative humidity and (ii) latitude, longitude, altitude, month, mean temperature, and relative humidity. SVM and MLR models are recommended for all scenarios in situations with limited climatic data where only air temperature and relative humidity data are available. The results and information presented in this study are important for the agricultural chain and water resources in Minas Gerais state.

**Keywords:** artificial neural network; random forest; support vector machine; multiple linear regression; crop water requirements; meteorological data



**Citation:** Santos, P.A.B.d.; Schwerz, F.; Carvalho, L.G.d.; Baptista, V.B.d.S.; Marin, D.B.; Ferraz, G.A.e.S.; Rossi, G.; Conti, L.; Bambi, G. Machine Learning and Conventional Methods for Reference Evapotranspiration Estimation Using Limited-Climatic-Data Scenarios. *Agronomy* **2023**, *13*, 2366. https://doi.org/10.3390/agronomy13092366

Academic Editors: Gniewko Niedbała, Maria do Rosario Cameira and Paula Paredes

Received: 18 July 2023
Revised: 6 September 2023
Accepted: 7 September 2023
Published: 12 September 2023



## 1. Introduction

Evapotranspiration is essential information in agriculture. The agriculture sector is found to be a major water consumer in most countries. The proportion of water withdrawn for agriculture in developing counties is estimated at nearly 81%, while it accounts for 71% of water withdrawal globally. Information on evapotranspiration is important in order to estimate crop water requirements and irrigation water requirements and control several hydrological processes [1–3]. Evapotranspiration (ET) is an agrometeorological parameter that can be measured using a lysimeter or water balance approach. These methods for measuring ET are not always possible to use. The lysimeter and water balance approaches are time-consuming methods and need precisely and carefully planned experiments [4]. Therefore, the use of evapotranspiration estimation methods is very important, and, for that, an adequate meteorological database is necessary to achieve good estimates [5,6].

The concept of evapotranspiration is related to the transfer rate of water from the soil–plant system to the atmosphere. In this study, we focus on the use of reference evapotranspiration ($ET_0$), which is related to the rate of water consumption from a reference crop surface (grass or alfafa). $ET_0$ can be used for a large area, e.g., for climatic classification of a region [7,8], or for small areas, e.g., for obtaining crop water requirements or crop evapotranspiration (ETc) [9–11]. The standard model used today for reference evapotranspiration estimation is the Penman–Monteith evapotranspiration model. This model is considered more realistic physically, but it requires some additional meteorological variables when compared with other methods [8]. This dependence on several meteorological variables combined with the limitations of weather station networks and interruptions and errors in weather databases makes it difficult to measure $ET_0$. Thus, some models are used to estimate $ET_0$. These models seek less dependence on many weather inputs and high predictive power.

Among the models used in the literature, this study focused on the following models: artificial neural network (ANN), random forest (RF), support vector machine (SVM), and multiple linear regression (MLR) models. These models show different levels of predictive capacity for different meteorological variables and in other fields of science [11–14]. ANN, RF, and SVM models can capture complex relationships between input and output data, which makes them powerful models for modeling. These machine-learning models have been successfully used to estimate $ET_0$ with fewer input meteorological data [12,15,16]. Although the inability of MLR to handle non-linear relationships between dependent and independent variables is evident in some studies, MLR has been successfully used to estimate $ET_0$ [13,17].

Considering the models, ANN is a promising and effective tool for non-linear modeling and complex time series. An ANN's architecture is composed of three layers—input, hidden, and output layers—and each layer includes an array of processing elements [6,12,16]. Several papers have shown the excellent predictive capacity of ANN models with different architectures in studies with $ET_0$ [14,15,18]. The RF model is a non-parametric statistical data modeling method that is decision-tree-based. RF is a classification and regression technique that has also been adopted to predict agrometeorological parameters such as $ET_0$ [15,19,20]. RF has been found to be a more efficient predicting tool compared with other tools like ANN [11,21]. SVM is a supervised machine-learning algorithm developed by [22]. SVM is used for regression, classification, pattern recognition, and forecasting. This model has been used in meteorological variable estimation and shown high predictive power [23,24]. MLR aims at explaining the collinearity between a dependent variable and an independent variable by means of a linear combination of independent predictor variables (more than one). This regression technique has been adopted in several fields of science, including climatology, hydrology, and irrigation, with varying performance [17].

There is so much literature on evapotranspiration that in this context it is practically impossible to propose even a partial review. Some remarkable recent contributions are due to [25–32]. This paper focuses on $ET_0$ estimation in the Minas Gerais state, Brazil, using different models. Agriculture has an important role in this region and $ET_0$ estimation on a monthly scale is extremely important for the agricultural chain. Among its main applications are the following: (i) climatic classification of a region—fundamental in the zoning of climatic risk in agricultural regions; (ii) hydrological processes—knowledge of evapotranspiration is fundamental in the hydrological cycle and, consequently, all studies related to hydrology and water resources; (iii) crop water requirements or crop evapotranspiration (ETc)—essential information in planning and implementing irrigation projects (i.e., determining the water demand of a given crop during the months of the year); and (iv) agrometeorological modeling—several models use ET data as an input variable for estimating productivity and other important variables; among other applications. This study also presents a relevant and innovative contribution through evaluation of the evapotranspiration estimates considering different climatic scenarios for the same state; that is, for regions which cover an extremely large area (such as the Minas Gerais state),

there may be a trade-off between generalization capacity and the performance of developed models. Therefore, data partition in the spatial sense aims to achieve the highest efficiency for the evaluated models, thus becomes relevant for the study of different climatic scenarios.

Considering that the presence of gaps or discontinuities in the meteorological data series can delay the state of development, this study proposes to analyze the use of different combinations of input data and climate scenarios for the accurate estimation of $ET_0$, and, especially, with the minimum possible use of input data in these models, this can facilitate the estimation of $ET_0$. The hypothesis of this study is that models based on machine learning are an efficient tool for estimating evapotranspiration, even under conditions of limited climatic data.

$ET_0$ calculated by the FAO Penman–Monteith method requires several input data. This amount of input data makes it difficult to use this method. New technologies can make it easier to obtain $ET_0$ reliably. In this context, the aim of this study was to develop, evaluate, and compare the performance of ANN, RF, SVM, and MLR models in estimating $ET_0$ with four different combinations of input data in three climate scenarios.

## 2. Materials and Methods

### *2.1. Study Area and Data Sources*

The Minas Gerais state is the fourth-largest in Brazil, with a territorial extent of 586,513.993 $km^2$ [33]. The study was performed with the database of Minas Gerais state, Brazil, between the parallels of 14°13′58″ and 22°54′00″, a southern latitude, and the meridians of 39°51′32″ and 51°02′35″ west of Greenwich. Monthly data from 56 climatological stations of the Brazilian National Institute of Meteorology (INMET) were used. Their respective geographical coordinates, altitudes, and climatic classifications have been presented in Table 1.

**Table 1.** Principal climatological stations of the INMET used to estimate $ET_0$.

| ID | Local | Lat/Lon/Alt (°/°/m) | K | Tho |
|---|---|---|---|---|
| 1 | Aimorés | −19.49/−41.07/79.93 | Aw | D |
| 2 | Araçuaí | −16.84/−42.06/317.67 | As | D |
| 3 | Araxá | −19.6/−46.94/1018.28 | Cwb | B2 |
| 4 | Arinos | −15.91/−46.1/523 | Aw | C1 |
| 5 | Bambuí | −20.03/−46/684.43 | Cwa | B2 |
| 6 | Barbacena | −21.23/−43.78/1128.8 | Cwb | B3 |
| 7 | Belo Horizonte | −19.93/−43.95/915.47 | Cwb | B2 |
| 8 | Bocaiúva | −17.1/−43.8/633 | Cwa | C1 |
| 9 | Bom Despacho | −19.72/−45.36/695 | Cwa | B1 |
| 10 | Caparaó | −20.52/−41.9/836.25 | Cwb | B2 |
| 11 | Capinópolis | −18.72/−49.56/608.98 | Aw | C2 |
| 12 | Caratinga | −19.73/−42.13/609.56 | Cwa | C2 |
| 13 | Conceição do Mato Dentro | −19.02/−43.43/663.02 | Cwa | B1 |
| 14 | Coronel Pacheco | −21.54/−43.26/411.03 | Cwa | B2 |
| 15 | Curvelo | −18.74/−44.45/668.26 | Cwa | C1 |
| 16 | Diamantina | −18.23/−43.61/1318.05 | Cwb | B2 |
| 17 | Divinópolis | −20.17/−44.87/787.42 | Cwa | B1 |
| 18 | Espinosa | −14.91/−42.8/565.52 | Cwb | D |
| 19 | Florestal | −19.88/−44.41/753.51 | Cwa | B2 |
| 20 | Formoso | −14.94/−46.23/854.6 | Aw | C2 |
| 21 | Frutal | −20.03/−48.93/547.09 | Aw | C2 |
| 22 | Governador Valadares | −18.84/−41.9/156.54 | Aw | C1 |
| 24 | Itamarandiba | −17.85/−42.85/919.37 | Cwb | C2 |
| 25 | Ituiutaba | −18.95/−49.52/540.09 | Aw | C2 |
| 26 | Jaíba | −15.08/−44.01/453.62 | As | D |
| 26 | Jaíba | −19.49/−42.54/298 | As | C2 |
| 27 | Janaúba | −15.8/−43.29/534.61 | As | D |
| 28 | Januária | −15.44/−44.36/480 | Aw | C1 |

**Table 1.** *Cont.*

| ID | Local | Lat/Lon/Alt (°/°/m) | K | Tho |
|---|---|---|---|---|
| 29 | João Monlevade | −19.82/−43.14/859.84 | Cwb | B2 |
| 30 | João Pinheiro | −17.74/−46.17/759.62 | Aw | C2 |
| 31 | Juiz de Fora | −21.77/−43.36/936.9 | Cwb | B3 |
| 32 | Juramento | −16.77/−43.66/655.59 | Cwb | C1 |
| 33 | Lambari | −21.94/−45.31/884.56 | Cwb | B3 |
| 34 | Lavras | −21.22/−44.97/916.19 | Cwb | B2 |
| 35 | Machado | −21.68/−45.94/892.44 | Cfb | B2 |
| 36 | Maria da Fé | −22.31/−45.37/1281.36 | Cwb | A |
| 37 | Monte azul | −15.16/−42.86/623.22 | As | D |
| 38 | Montes Claros | −16.68/−43.84/645.87 | Cwa | C1 |
| 39 | Paracatu | −17.24/−46.88/711.41 | Aw | C2 |
| 40 | Patos de Minas | −18.52/−46.44/947.68 | Cwa | B1 |
| 41 | Pedra Azul | −16/−41.28/647.97 | As | C1 |
| 42 | Pirapora | −17.34/−44.92/509.52 | Aw | C1 |
| 43 | Poços de Caldas | −21.91/−46.38/1077.08 | Cwb | B3 |
| 44 | Pompéu | −19.22/−45/692.21 | Cwa | C2 |
| 45 | Salinas | −16.15/−42.28/476.07 | As | D |
| 46 | São João Del Rei | −21.3/−44.27/991 | Cwb | B3 |
| 47 | São Lourenço | −22.12/−45.04/930.65 | Cwb | B3 |
| 48 | São Sebastião do Paraíso | −20.9/−47.11/820 | Cwb | B3 |
| 49 | Serra Azul de Minas | −20.02/−44.35/765 | Cwa | B2 |
| 50 | Serra dos Aimorés | −17.79/−40.25/211.92 | Aw | C1 |
| 51 | Sete Lagoas | −19.48/−44.17/753.68 | Cwa | B1 |
| 52 | Teófilo Otoni | −17.86/−41.5/349.11 | Aw | C1 |
| 53 | Uberaba | −19.73/−47.95/753.41 | Cwa | B2 |
| 54 | Uberlândia | −18.91/−48.25/874.6 | Cwa | B2 |
| 55 | Unaí | −16.36/−46.88/595.59 | Aw | C1 |
| 56 | Viçosa | −20.76/−42.86/697.53 | Cwa | B1 |

K—Köppen climatic classification; Tho—Thornthwaite climatic classification; Cwb—Humid subtropical with dry winter and temperate summer; Cwa—Humid subtropical with dry winter and hot summer; Cfb—Humid subtropical with oceanic climate without dry season and with temperate summer; As—Tropical with dry summer; and Aw—Tropical with dry winter; A—super-humid; B4—humid; B3—humid; B2—humid; B1—humid; C2—sub-humid; C3—dry sub-humid; and D—semiarid. Source: the authors.

The input variables that were considered in this study were latitude; longitude; altitude; month; and average monthly data mean, maximum, and minimum air temperatures (Tmean, Tmax, Tmin); relative humidity (RH); atmospheric pressure (P); wind speed (U2); and insolation (n). These data were obtained in climatological stations with at least 10 years of flawless data (no missing or faulty data) from a period between 1989 and 2019 (30 years). This selection criterion led to the inclusion of 56 stations. Due to the removal of inaccurate and inconsistent data, a total of 13,577 data rows (each of these data rows contains all the meteorological variables used in the models) were considered for analysis. Wind speed, measured at a 10 m height, was converted to 2 m [34]. Days with missing or faulty data were removed. Faulty data were identified when Tmin was higher than Tmax or Tmean; Tmean was higher than Tmax; RH was out of the range 0–100%; P was higher than 101.4 kPa; or U2 or n were negative. The output variable was reference evapotranspiration ($ET_0$).

The reasons for using these variables were as follows. Latitude and longitude are the variables related to position. Solar radiation intensity changes as position changes on the terrestrial globe. The altitude variable is regarded as the surface component. It can be stated that the higher the altitude, the lower the temperature. Temperature is the availability of energy in the system, and relative humidity is the difference in gradient; the lower the humidity, the greater the capacity of the environment to absorb humidity. All these factors can influence evapotranspiration.

In general, a more homogeneous region can enhance the accuracy of climatic variable prediction models. According to [12], building models specifically for regions with similar

climatic conditions can increase performance. However, in large areas, there may be a trade-off between generalization capacity and the performance of developed models. Data partition in the spatial sense aims to achieve the highest efficiency for evaluated models; thus, different scenarios were created.

The models were developed in three different scenarios (SI, SII and SIII) in order to achieve the maximum predictive capacity for each model. SI—at state level, the models were trained and tested with data from the 56 climatological stations. The resulting model estimates evapotranspiration in any location within the Minas Gerais state. SII—at regional level, the Minas Gerais state was divided into two regions according to the climatic classification system proposed by Thornthwaite [27]: a region with climate classifications of A, B4, B3, B2, and B1 (Tho1—27 climatological stations) and a region with climate classifications of C2, C1, and D (Tho2—29 climatological stations). The models were trained and tested with data from the climatological stations of each climatic region (Figure 1). SIII—at regional level, the Minas Gerais state was divided into two regions: a region with climate classifications of Cwb, Cwa, and Cfb (K1—35 climatological stations) and a region with climate classifications of Aw and As (K2—21 climatological stations) using the climatic classification system proposed by Köppen. The models were trained and tested with data from the climatological stations of each climatic region (Figure 1).

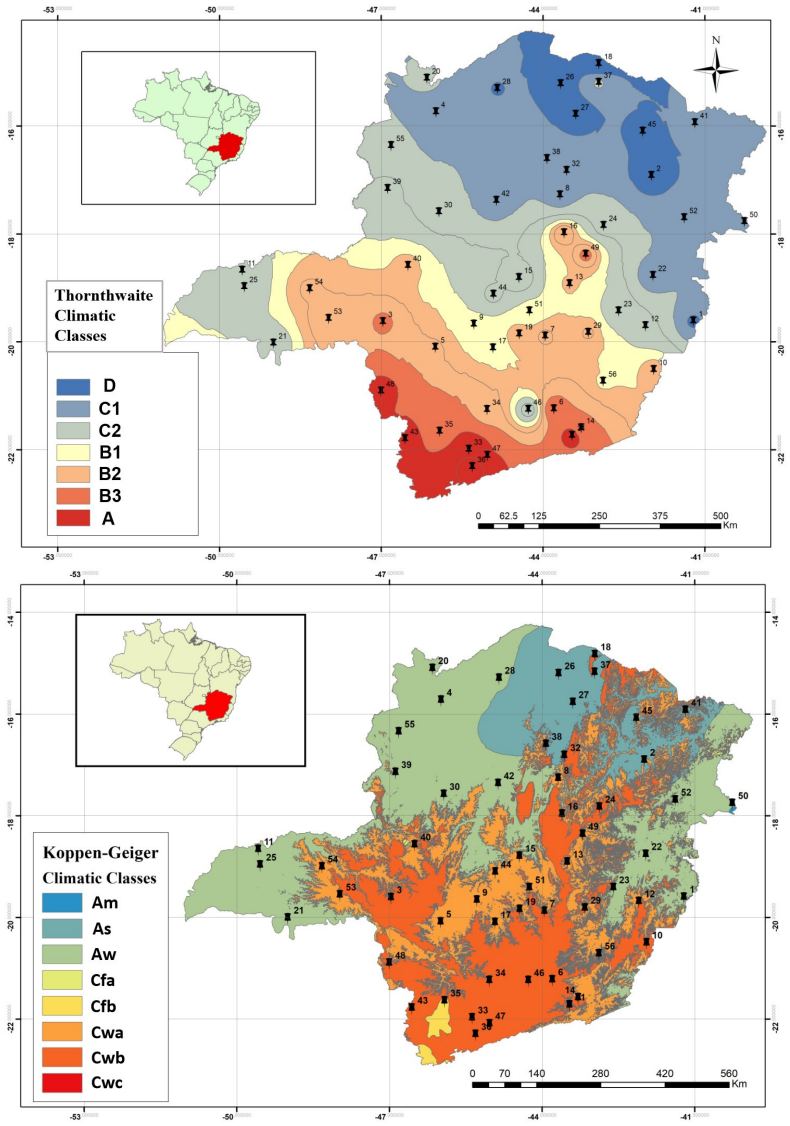


**Figure 1.** Climate classifications for Minas Gerais state according to Thornthwaite [35] (source: the authors, 2023) and Köppen classifications [36]. Black pins represent the meteorological station.

### 2.2. Penman–Monteith FAO Model

The FAO Penman–Monteith equation (FPM) was used to estimate average monthly $ET_0$. This method is described in [34]. It is common practice to use $ET_0$ values estimated by the FPM equation as reference data. The climatological stations used in this study do not provide net solar radiation (Rn) data. The Rn data were obtained from insolation, latitude, day of the year, and other variables. They are calculated using the equations detailed in [26].

Although recommended as a reference method, the equation proposed for FAO has several parameters based on a series of general assumptions about ground cover and vegetation, which means that the FPM equation is a simplification. However, due to the lack of reliable data from lysimeters and the difficulty of handling them, use of the FPM equation is recommended. According to [37], this equation is recommended for estimating $ET_0$ and validating other equations in the absence of experimental measurements; studies that consider FPM targets to train and test models often overlook the implications that arise from this simplification.

### 2.3. Model Development and Statistical Tests

In this study, different input combinations of the average monthly data were used as inputs to estimate $ET_0$. The input data were geographic coordinates, altitude, month, Tmean, Tmax, Tmin, and RH. In the search for better performance, the four input combinations (In: n is the amount of input data) evaluated in this paper were: (I8) latitude, longitude, altitude, month, Tmean, Tmax, Tmin, and RH; (I6) latitude, longitude, altitude, month, Tmean, and RH; (I3) month, Tmean, and RH; and (I2) Tmean and RH. Combinations were employed to investigate the influence of each meteorological variable on estimation and the resulting impact when a variable was removed. Moreover, average temperature (T) and average relative humidity (RH) data were kept consistent across all combinations, as they are responsible for the energy available in the system and the gradient difference, respectively.

The ANN, RF, SVM, and MLR models were trained for each combination. The models were developed using data from each climate scenario. These combinations were compared with each other in each model.

The predictive quality of each model in terms of variation, precision, accuracy, and performance was evaluated by four statistical criteria. The statistical criteria were mean absolute error (MAE), root-mean-square error (RMSE), coefficient of determination (R2), and Pearson's correlation coefficient (r) (equations below). MAE and RMSE indicate how close the predicted values were to the observed value. Thus, the accuracy of each model could be predicted. $R^2$ represents the percentage of the variation in the dependent variable explained by the independent variable. r indicates the degree of dispersion of the data obtained in terms of the mean.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |\mathrm{Pi} - \mathrm{Oi}|}{\mathrm{N}} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} |\mathrm{Pi} - \mathrm{Oi}|^2}{\mathrm{N}}} \quad (2)$$

$$\mathrm{R}^2 = \left[ \frac{\sum_{i=1}^{N} (\mathrm{Pi} - \overline{\mathrm{P}})(\mathrm{Oi} - \overline{\mathrm{O}})}{\sqrt{\left(\sum_{i=1}^{N} (\mathrm{Pi} - \overline{\mathrm{P}})^2\right)\left(\sum_{i=1}^{N} (\mathrm{Oi} - \overline{\mathrm{O}})^2\right)}} \right]^2 \quad (3)$$

$$\mathrm{r} = \frac{\sum_{i=1}^{N} (\mathrm{Pi} - \overline{\mathrm{P}})(\mathrm{Oi} - \overline{\mathrm{O}})}{\sqrt{\left(\sum_{i=1}^{N} (\mathrm{Pi} - \overline{\mathrm{P}})^2\right)\left(\sum_{i=1}^{N} (\mathrm{Oi} - \overline{\mathrm{O}})^2\right)}} \quad (4)$$

where Pi is the predicted value (mm), Oi is the observed value (mm), P is the mean of the predicted values (mm), O is the mean of the observed values (mm), and N is the number of data pairs.

### *2.4. Artificial Neural Networks (ANN)*

ANNs have performance characteristics resembling the biology of the human brain. ANNs, in general, have architectures with connections between nodes (neural networks) and methods to determine the connection weights. In this study, an ANN of the feed-forward multilayer perceptron (MLP) type was used [38]. An MLP is a robust choice due to its ability to handle a variety of problems and learn complex nonlinear functions effectively. With multiple hidden layers, MLPs can capture intricate relationships within the data, providing greater flexibility in modeling [12,16]. The training of this ANN involved two phases. In the first phase, or forward pass, the input sign spreads forward layer by layer. In the second phase, or reverse pass, the sign is backpropagated for correction of the error.

ANN was implemented using the Waikato Environment for Knowledge Analysis (WEKA; version 3.8.2 © 1999–2017) developed by the University of Waikato, Hamilton, New Zealand. The input data consisted of different combinations of the latitude, longitude, altitude, month, Tmean, Tmax, Tmin, and RH for each evaluated location, using $ET_0$ as the output variable.

All adjustments were performed by cross-validation. According to [14], the cross-validation approach enables successful results. The method employed in constructing the models was k-fold cross-validation. This technique uses all available data, which is partitioned into k disjoint subsets roughly equal in size. This partitioning is performed by random sampling of the learning set without replacement. The model is then trained k times, using k-1 subsets for training and the remaining subset for validation and to assess its performance. This procedure is repeated until each of the k subsets has served as the validation set. The average of the performance metrics from all of these interactions is considered the cross-validation performance [39]. This methodology was used due to the limited number of climatological stations within the Minas Gerais state. According to [39], K-fold cross-validation is commonly utilized when the quantity of data is limited, as it helps to maximize the utilization of the available data. In this way, this method makes it possible to work with fewer data and obtain optimal results.

The different ANN configurations and number of folds used in cross-validation are shown in Table 2. In this study, models with one or two hidden layers were utilized (Table 2). Initial tests were performed to determine the best-performing model within each input data combination. Various configurations, including different numbers of neurons in the layers, were tested. The architecture with the highest performance for each data input combination was selected (Figure 2).

**Table 2.** WEKA configuration in the ANN implementation.

| | **ANN** | | | |
|---|---|---|---|---|
| | **I8** | **I6** | **I3** | **I2** |
| Learning rate | 0.3 | 0.3 | 0.3 | 0.3 |
| Momentum | 0.2 | 0.2 | 0.2 | 0.2 |
| Number of training epochs | 1000 | 1000 | 1000 | 1000 |
| Number of input data | 8 | 6 | 3 | 2 |
| Number of hidden layers | 2 | 2 | 1 | 1 |
| Number of neurons into the hidden layer | *7.7* | *7.7* | *7.7* | 7 |
| Number of folds in cross-validation | 18 | 18 | 8 | 8 |

In italics: WEKA default values.

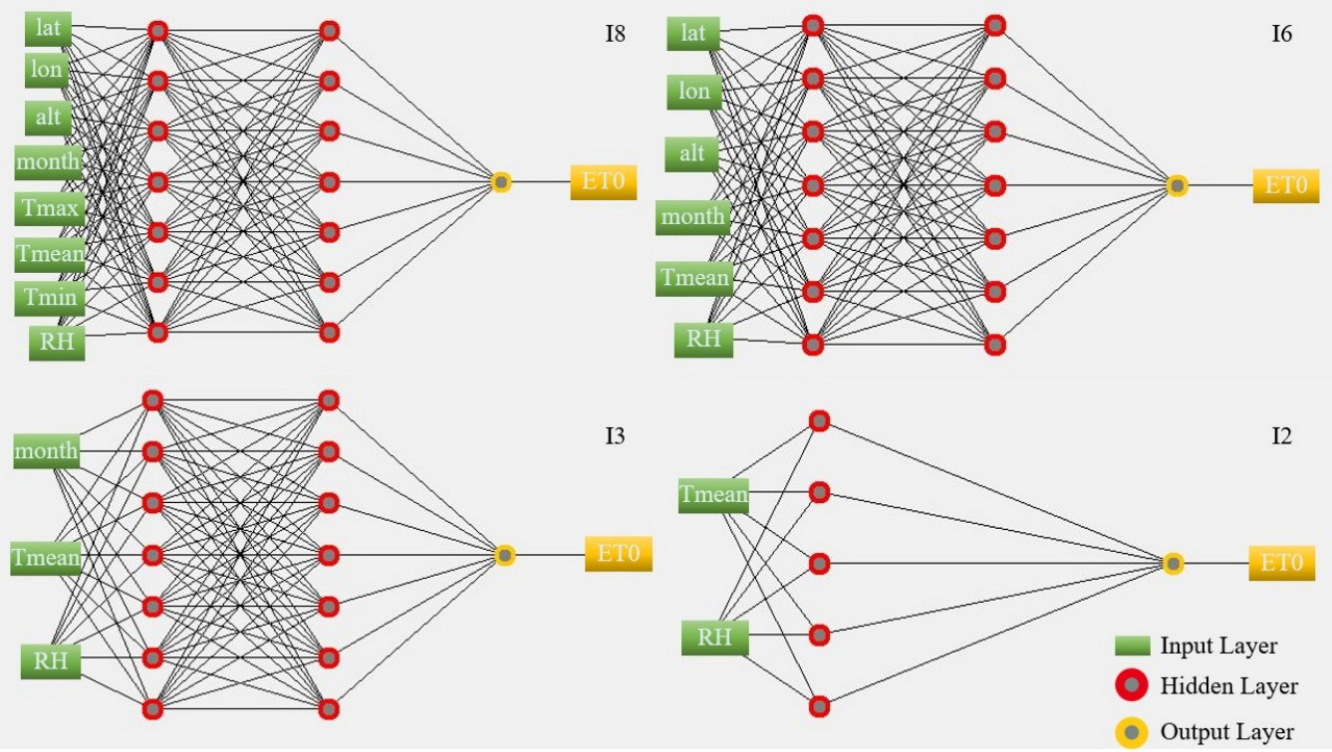


**Figure 2.** Network structure scheme built by WEKA to estimate $ET_0$.

However, it is important to note that using multiple hidden layers allows the model to learn more complex representations of the data. While it is possible for a single hidden layer to approximate any function, it may be computationally inefficient or require a large number of neurons in the hidden layer to learn an adequate representation of the data. Additionally, neural networks with a single hidden layer often struggle with generalization, which can lead to inaccurate prediction for unseen data. However, in some situations, it was observed that a single layer performed better than two.

Regarding the initial random assignment of weights, also known as seed, WEKA allows users to change the values as necessary or randomly generate them without the need for an initial introduction. However, if no weight is assigned, the default setting will assign a value of 1 to the seed. Therefore, in this study, the random seed for initialization was set to 1. This allows for consistent comparison of results across different runs of the algorithm and facilitates assessment of the reliability of the results. The other WEKA configuration parameters were kept as standard.

### *2.5. Support Vector Machine (SVM)*

In this study, SVM equations were applied based on Vapnik's theory [22]. SVMs are separated into two main categories: (i) the classifier model and (ii) the regression model (SVR). SVR is used to take a hyperplane suitable for the data used. The distance to any point in this hyperplane shows the error of that point [14]. SVR can be translated into the following equation:

$$y = f(x) = \omega\ \varphi(x_i) + b \quad (5)$$

where x is the input data; $\varphi(x)$ represents a function that can transfer x into high-dimensional feature spaces; and $\omega$ (weight vector) and b are coefficients which are estimated by minimizing the regularized risk function. The error function in the SVM model is minimized based on the mentioned constraints in the equation below. Further details on the application of SVM can be found in [40].

$$\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_1 + \xi_1^*) \quad (6)$$

$$\text{Subject to} \begin{cases} y_i - b[\omega\ \varphi(x_i)] \leq \epsilon + \xi_1 \\ [\omega\ \varphi(X_i)] + b - y_i \leq \epsilon + \xi_1^* \\ \xi_1 \xi_1^*\ \geq 0 \end{cases} \quad (7)$$

where C is the capacity or penalty parameter, $y_i$ is the estimated output by SVM, and $\xi_i$ and $\xi^*_i$ are slack variables which must satisfy the function constraints. The SVM model changes the scale of the problem by using kernel functions to solve non-linear problems. SVM provides four different kernel functions: sigmoid, linear, polynomial, and radial basis functions. In this study, during SVM modelling, all kernel functions were tested. The linear kernel function proved to be more efficient for estimating $ET_0$. The linear kernel function is as follows:

$$K(x_i, y_i) = x_i y_i \quad (8)$$

where $x_i$ and $x_j$ are vectors in the input space. The SVM was implemented by WEKA. The input data consisted of different combinations (I8, I6, I4 and I2) and were evaluated in the three different scenarios. The WEKA configuration parameters in the SVM implementation were: SVM Type, $\epsilon$-SVR; cost parameter C, 0.01. The random seed for initialization was set to 1, and gamma was not utilized as the kernel function was linear. The other WEKA configuration parameters were kept as standard for the libsvm library. The libsvm library simplifies the use of SVM in various studies and has enabled its application in pattern recognition and other machine-learning fields [39]. Eighteen folds of the sample set were used in cross-assessment. The same WEKA configuration parameters were used for all input data combinations and in all scenarios.

### *2.6. Random Forest (RF)*

RF is an ensemble learning technique based on a collection of tree predictors [41]. It is a combination of many predictor trees (forest), in which each tree is generated from a random vector and sampled independently, with the same distribution for all trees in the forest. According to [20], there are three simple steps to building an RF model: (i) build n bootstrap samples from the original data; (ii) build an unpruned regression tree; and (iii) predict new data by aggregating the predictions of the n. More details can be found in [20,42] regarding the representation of the steps used in the RF model following the resampling strategy.

RF was implemented by WEKA. The WEKA configuration that resulted in the greatest predictive capacity was a bag size of 100 (the size of each bag, as a percentage of the training set size); 500 iterations (the number of trees in the random forest); unlimited depth for individual trees (as standard); and a random seed for initialization of 1. The other WEKA configuration parameters were kept as standard. All adjustments were performed with cross-validation, and twenty folds of the sample set were used; therefore, the number of splits is not specified. The same WEKA configuration parameters were used in all input data combinations and in all scenarios as they yielded the best results. Adjustments to the hyperparameters aimed at minimizing the root-mean-square error (RMSE) of the validation set. The methodology applied was similar to that employed by [43]; however, it did not involve a separate test set due to the cross-validation approach.

### *2.7. Multiple Linear Regression (MLR)*

MLR was developed to estimate $ET_0$ based on different combinations of the independent variables. The base regression equation can be expressed as:

$$Y_i = \beta_0 + \beta_1 lat + \beta_2 lon + \beta_3 alt + \beta_4 month + \beta_5 Tmax + \beta_6 Tmean + \beta_7 Tmin + \beta_8 RH \quad (9)$$

where $Y_i$ is the dependent variable ($ET_0$); lat, lon, alt, month, Tmean, Tmax, Tmin, and RH are independent variables; and $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$, $\beta_5$, $\beta_6$, $\beta_7$, and $\beta_8$ are the regression coefficients.

MLR was implemented using WEKA. The attribute selection method used in the WEKA configuration was the M5 method. This method initially builds the MLR model with all independent variables. Then, the independent variables with the smallest standardized coefficients are removed stepwise until no improvement is observed in the estimate of the error given by the Akaike information criterion (AIC). The AIC seeks the best model in terms of complexity and performance. This technique evaluates different models relative to each other; therefore, when adding more parameters, the AIC of the model may show inadequate performance [44]. The other WEKA configuration parameters were kept as standard. Eighteen folds of the sample set were used in cross-assessment. The same WEKA configuration parameters were used for all input data combinations and in all scenarios.

## 3. Results and Discussion

The results presented in this study are essential for more adequate water management, since accurate estimation of $ET_0$ is fundamental for water demand quantification. Moreover, the use of different estimation techniques and combinations of input data in the models allowed us to obtain important results at different spatial scales. It must be noted that while daily $ET_0$ values are useful for conducting irrigation, monthly $ET_0$ provides an overview of how much water is required to maintain plant health over a longer period, such as a month or growth cycle. Monthly $ET_0$ is particularly valuable in irrigation planning, as it helps water managers, designers, development planners, and farmers estimate the total water requirements for a successful harvest and make accurate decisions.

According to the results, it was possible to observe linear correlations between the input data and $ET_0$, with the variables Tmean, Tmax, and Tmin showing the best correlation (Figure 3). The other variables have a low (lat, alt and RH) or no (lon and month) correlation with $ET_0$. Behavior inversely proportional to $ET_0$ was observed for the lat, alt, and RH variables. Higher latitudes tend to be cooler regions, with less energy available for the $ET_0$ process. An increase in altitude also results in a decrease in temperature according to the vertical thermal gradient in the troposphere. An increase in RH increases the potential gradient, increasing the water transfer rate from the soil–plant system to the atmosphere. However, proportional behavior was observed between the Tmean, Tmax, and Tmin variables and $ET_0$. An increase in Tmean, Tmax, or Tmin results in more energy being available for $ET_0$. The authors of [14] observed the same behavior in the variables Tmean, Tmax, Tmin, and RH when estimating $ET_0$. The variables Tmean, Tmax, and Tmin were all highly correlated with $ET_0$, and the RH mean was the least correlated variable.

In this way, the capability of machine-learning approaches using the variables mentioned above was investigated in different conditions and scenarios. The ANN, RF, SVM, and MLR statistical performance indicators for estimating $ET_0$ in any location within the Minas Gerais state (SI: data from the 56 climatological stations—100% of the input data available) are presented in Table 3.

All the models developed with the I8 and I6 input combinations exhibited better performance than versions developed with I3 and I2. The lowest predictive capacity was observed when the RF model was used with the I8 input combination. The greatest predictive capacity, in SI, was observed when the RF and ANN models were used with the I6 and I8 input combinations, respectively. The SVM and MLR models exhibited better performance than ANN and RF when only Tmean and RHmean (I2) were used as input data.

**Table 3.** Statistical performance indicators of the ANN, RF, SVM, and MLR models in SI.

| | SI | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $I_8$ | | | $I_6$ | | | $I_3$ | | | $I_2$ | | |
| | **r** | **MAE** | **RMSE** | **r** | **MAE** | **RMSE** | **r** | **MAE** | **RMSE** | **r** | **MAE** | **RMSE** |
| ANN | ***0.966*** | ***0.167*** | ***0.215*** | 0.963 | 0.178 | 0.224 | *0.943* | *0.210* | *0.278* | 0.860 | 0.332 | 0.429 |
| RF | 0.955 | 0.191 | 0.250 | ***0.966*** | ***0.166*** | ***0.220*** | 0.934 | 0.220 | 0.296 | 0.859 | 0.335 | 0.426 |
| SVM | **0.933** | **0.23** | **0.290** | 0.927 | 0.242 | 0.310 | 0.878 | 0.311 | 0.399 | *0.877* | *0.312* | *0.399* |
| MLR | **0.933** | **0.231** | **0.298** | 0.928 | 0.241 | 0.308 | 0.877 | 0.313 | 0.399 | *0.877* | *0.312* | *0.398* |

Values in bold indicate the best results within each model; values in italics indicate the best results within each input data combination. Data combinations: (I8) latitude, longitude, altitude, month, Tmean, Tmax, Tmin, and UR; (I6) latitude, longitude, altitude, month, Tmean, and RH; (I3) month, Tmean, and UR; and (I2) Tmean and RH. RMSE and MAE are in mm day$^{-1}$.

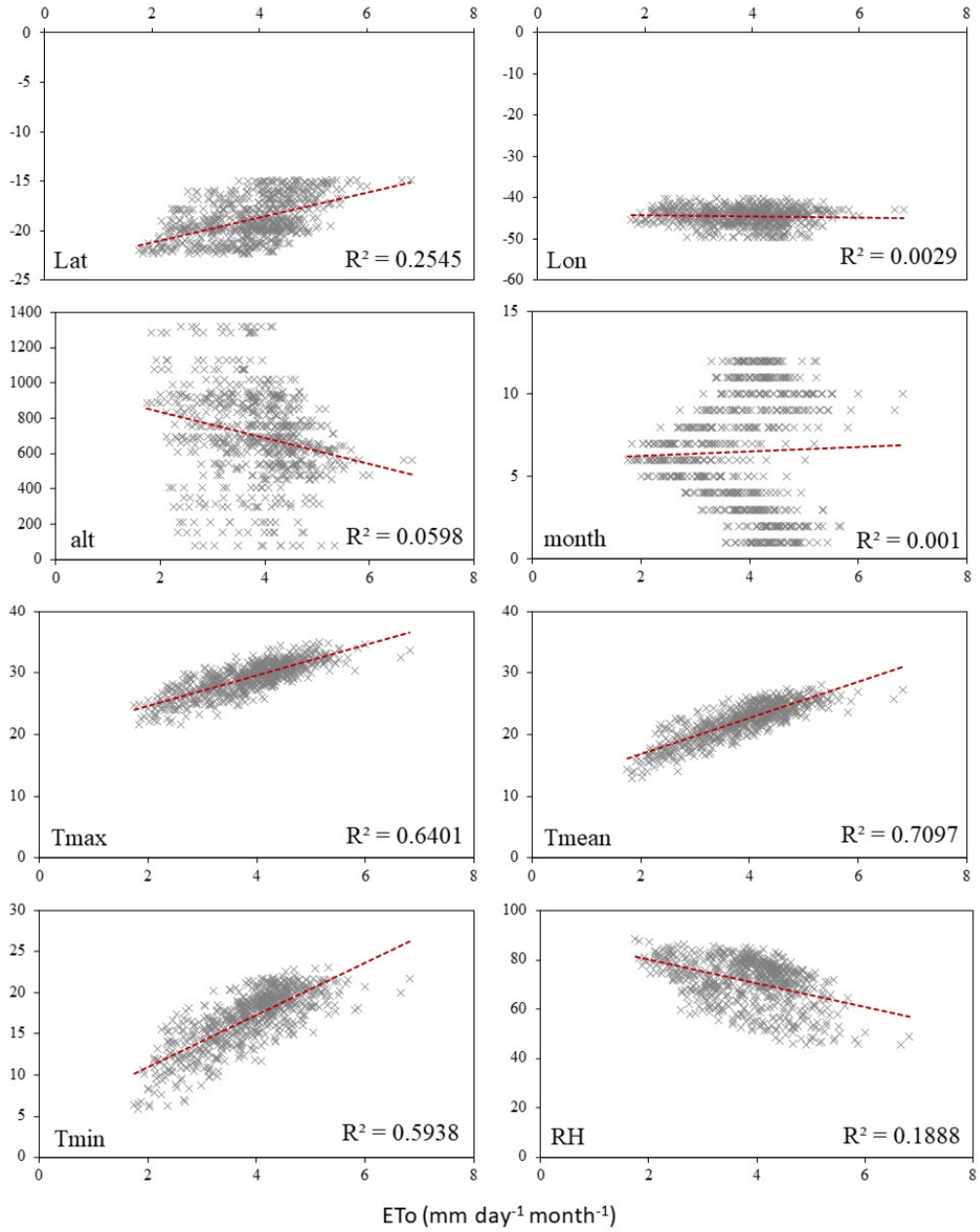


**Figure 3.** Correlation analysis between $ET_0$ and each input variable.

When comparing combination I8 with I6, the average r, MAE, and RMSE values for all models do not show high variation. Removal of the geographic coordinates (I6 to I3) resulted in greater performance reduction for the SVM and MLR models. The greatest impact on performance was observed for ANN and RF when the month variable was removed (I3 to I2). Average r decreased by 8%; MAE and RMSE increased by 52.2% and 43.9%, respectively. The removal of month did not impact the SVM and MLR models' performance.

In the case of SII, a scenario in which the state of Minas Gerais was divided into two areas (Tho1 and Tho2), the statistical performance indicators of the models used in $ET_0$ estimation are shown in Table 4.

**Table 4.** Statistical performance indicators of the ANN, RF, SVM, and MLR models in SII.

| **SII** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tho.1 (A, B4, B3, B2, and B1) | | | | | | | | | | | | |
| | $I_8$ | | | $I_6$ | | | $I_3$ | | | $I_2$ | | |
| | r | MAE | RMSE | r | MAE | RMSE | r | MAE | RMSE | r | MAE | RMSE |
| ANN | ***0.976*** | ***0.135*** | ***0.168*** | 0.965 | 0.156 | 0.196 | 0.959 | 0.169 | 0.212 | 0.904 | 0.266 | 0.322 |
| RF | 0.964 | 0.164 | 0.198 | **0.973** | **0.143** | **0.174** | *0.963* | *0.16* | *0.198* | 0.920 | 0.234 | 0.291 |
| SVM | **0.955** | **0.181** | **0.219** | 0.948 | 0.191 | 0.235 | 0.925 | 0.232 | 0.281 | 0.924 | 0.235 | 0.284 |
| MLR | **0.957** | **0.178** | **0.216** | 0.949 | 0.190 | 0.233 | 0.927 | 0.23 | 0.278 | *0.925* | *0.233* | *0.282* |
| Tho.2 (C2, C1, and D) | | | | | | | | | | | | |
| | $I_8$ | | | $I_6$ | | | $I_3$ | | | $I_2$ | | |
| | r | MAE | RMSE | r | MAE | RMSE | r | MAE | RMSE | r | MAE | RMSE |
| ANN | 0.940 | 0.211 | 0.269 | ***0.944*** | ***0.210*** | ***0.260*** | *0.895* | *0.269* | *0.353* | 0.818 | 0.377 | 0.462 |
| RF | 0.925 | 0.227 | 0.302 | **0.943** | **0.200** | **0.267** | 0.879 | 0.29 | 0.374 | 0.817 | 0.350 | 0.453 |
| SVM | 0.893 | 0.276 | 0.353 | **0.898** | **0.271** | **0.346** | 0.840 | 0.342 | 0.427 | 0.840 | 0.340 | 0.427 |
| MLR | 0.898 | 0.269 | 0.345 | **0.899** | **0.267** | **0.342** | 0.839 | 0.339 | 0.427 | 0.839 | 0.339 | 0.427 |

Values in bold indicate the best results within each model; values in italics indicate the best results within each input data combination. Data combinations: (I8) latitude, longitude, altitude, month, Tmean, Tmax, Tmin, and UR; (I6) latitude, longitude, altitude, month, Tmean, and RH; (I3) month, Tmean, and UR; and (I2) Tmean and RH. RMSE and MAE are in mm $day^{-1}$.

Tho1 and Tho2 had 48.2% and 51.8%, respectively, of the data available as input data. The highest predictive capacities in the Tho1 and Tho2 areas were observed when the ANN model was used with the I8 input combination and RF model was used with the I6 input combination, respectively. The removal of Tmax and Tmin input data (I6) did not increase the models' predictive capacities in the Tho1 area, except for the RF model. This behavior is similar to that observed in SI. However, all models performed better in the Tho2 area when the I6 input combination was used (better results).

Removal of the month variable (I3 to I2) resulted in the greatest impact on the ANN and RF models' quality. When comparing combination I8 with I3, the average r values of the ANN and RF models decreased by 7.2% and 5.7%, respectively. The MAE values of the ANN and RF models increased by 36.4% and 31.6%, respectively. However, no expressive variation was observed in the performance of the SVM and MLR models.

In the case of SIII, the statistical performance indicators of the models for this scenario are presented in Table 5, where the Minas Gerais state was divided in areas K1 and K2, which were characterized by 62.5% and 37.5% of the climatological stations, respectively.

**Table 5.** Statistical performance indicators of the ANN, RF, SVM, and MLR models in SIII.

| **SIII** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **K1 (Cwa, Cwb, and Cfb)** | | | | | | | | | | | | |
| | $I_8$ | | | $I_6$ | | | $I_3$ | | | $I_2$ | | |
| | r | MAE | RMSE | r | MAE | RMSE | r | MAE | RMSE | r | MAE | RMSE |
| ANN | 0.966 | 0.170 | 0.209 | **0.968** | **0.163** | **0.204** | 0.961 | 0.180 | 0.225 | 0.912 | 0.270 | 0.334 |
| RF | 0.963 | 0.175 | 0.221 | ***0.973*** | ***0.15*** | ***0.191*** | 0.962 | 0.174 | 0.222 | 0.920 | 0.261 | 0.318 |
| SVM | **0.949** | **0.199** | **0.256** | 0.944 | 0.209 | 0.267 | 0.927 | 0.247 | 0.305 | 0.926 | 0.248 | 0.306 |
| MLR | **0.950** | **0.204** | **0.253** | 0.945 | 0.212 | 0.266 | 0.928 | 0.245 | 0.303 | 0.917 | 0.247 | 0.305 |
| K2 (Am and Aw) | | | | | | | | | | | | |
| | $I_8$ | | | $I_6$ | | | $I_3$ | | | $I_2$ | | |
| | r | MAE | RMSE | r | MAE | RMSE | r | MAE | RMSE | r | MAE | RMSE |
| ANN | ***0.964*** | ***0.16*** | ***0.201*** | 0.889 | 0.269 | 0.350 | *0.895* | *0.263* | *0.340* | 0.817 | 0.360 | 0.447 |
| RF | 0.924 | 0.23 | 0.294 | **0.943** | **0.203** | **0.258** | 0.885 | 0.285 | 0.352 | 0.826 | 0.347 | 0.429 |
| SVM | 0.889 | 0.270 | 0.347 | **0.890** | **0.274** | **0.347** | 0.846 | 0.329 | 0.405 | 0.847 | 0.326 | 0.403 |
| MLR | 0.892 | 0.269 | 0.343 | **0.894** | **0.269** | **0.340** | 0.846 | 0.325 | 0.404 | *0.848* | *0.323* | *0.403* |

Values in bold indicate the best results within each model; values in italics indicate the best results within each input data combination. Data combinations: (I8) latitude, longitude, altitude, month, Tmean, Tmax, Tmin, and UR; (I6) latitude, longitude, altitude, month, Tmean, and RH; (I3) month, Tmean, and UR; and (I2) Tmean and RH. RMSE and MAE are in mm $day^{-1}$.

In general, the ANN and RF models were better than the SVM and RLM models with the input combinations I8, I6, and I3. When the I2 combination was used, the SVM and RLM models were superior. The model with highest predictive capacity in the K1 area was ANN with the I8 input combination. The RF model with the I6 input combination showed the highest predictive capacity in the K2 area.

In the K1 area, removal of the month variable resulted in the greatest impact on the ANN and RF models' performance. Removal of the alt, lat, and lon variables resulted in the highest impact on the SVM and MLR models' performance. In the K2 area, the behavior of RF, SVM, and MLR was similar to that observed in the K1 area. However, withdrawal of the alt, lat, and lon variables resulted in the highest impact on ANN in the K2 area.

The ANN and RF models showed greater predictive capacity in all scenarios when compared with the SVM and MLR models. This high capacity is achieved with the data input combinations I8 and I6. Both models had similar performance, but, on average, RF showed slight superiority. In [12,15], the authors evaluated the performance of different machine-learning models in $ET_0$ estimation in Brazil. In these studies, it was observed that, in general, ANN performed slightly better than the other traditional machine-learning models (i.e., RF and extreme gradient boosting—XGBoost). However, in some studies, the RF model performed slightly better than other models (i.e., generalized regression neural networks—GRNN) in estimating $ET_0$ [20,45]. There are papers suggesting better performance than other machine-learning models in different situations and regions [24,46]. Therefore, there is a need for studies that address more than one model.

The SVM and MLR models showed similar statistical indices and responses in all scenarios. These results can be explained by the use of the linear kernel function in SVM, which probably presented behavior similar to an MLR. Tests with the nonlinear kernel function did not result in improvements in prediction. Possibly, the data used does not present complexity that justifies the use of SVM.

The SVM and MLR models showed greater predictive capacity in all scenarios when the input data were limited to only Tmean and RH (I2). This result may indicate a low predictive capacity of the ANN and RF models in situations of low variability in the input data. This low variability may hinder the search for patterns that justify variations in $ET_0$.

In some scenarios, the removal of Tmax and Tmin improved the $ET_0$ estimation results. According to [14], the authors observed an increase in the accuracy of the support vector

regression (SVR) and Gaussian process regression (GPR) models with the removal of some input data, including Tmax and Tmin.

Although Tmax and Tmin showed a good correlation with $ET_0$ (Figure 3), the weight of Tmax and Tmin is diluted in the calculation of Tmean used in the calculation of $ET_0$. Thus, adding Tmax and Tmin can make $ET_0$ estimation more complex or confusing. This fact can decrease the accuracy of the models, and the removal of this input data can improve the prediction. Determining the input data is critical to the success of the models. This selection can facilitate the training and testing process, improving the understanding of the system [47,48]. However, this result shows that linear regression alone is not sufficient to decide which input data should be removed in order to increase predictive performance.

When the independent variables lat, lon, and alt were removed (I3), a reduction in the statistical indexes of all models was observed. These variables are related to the spatial location of the observed data. Although the correlation observed between these variables and $ET_0$ is low (Figure 4), the joint removal of these data negatively impacted the model's performance. The air temperature and solar radiation variables are among the main data impacting $ET_0$ [1,46]. Several studies have indicated the influence of lat, lon, and alt variables on air temperature and solar radiation [49,50]. Therefore, variations in lat, lon, and alt may indirectly impact $ET_0$. This can explain these observed results.

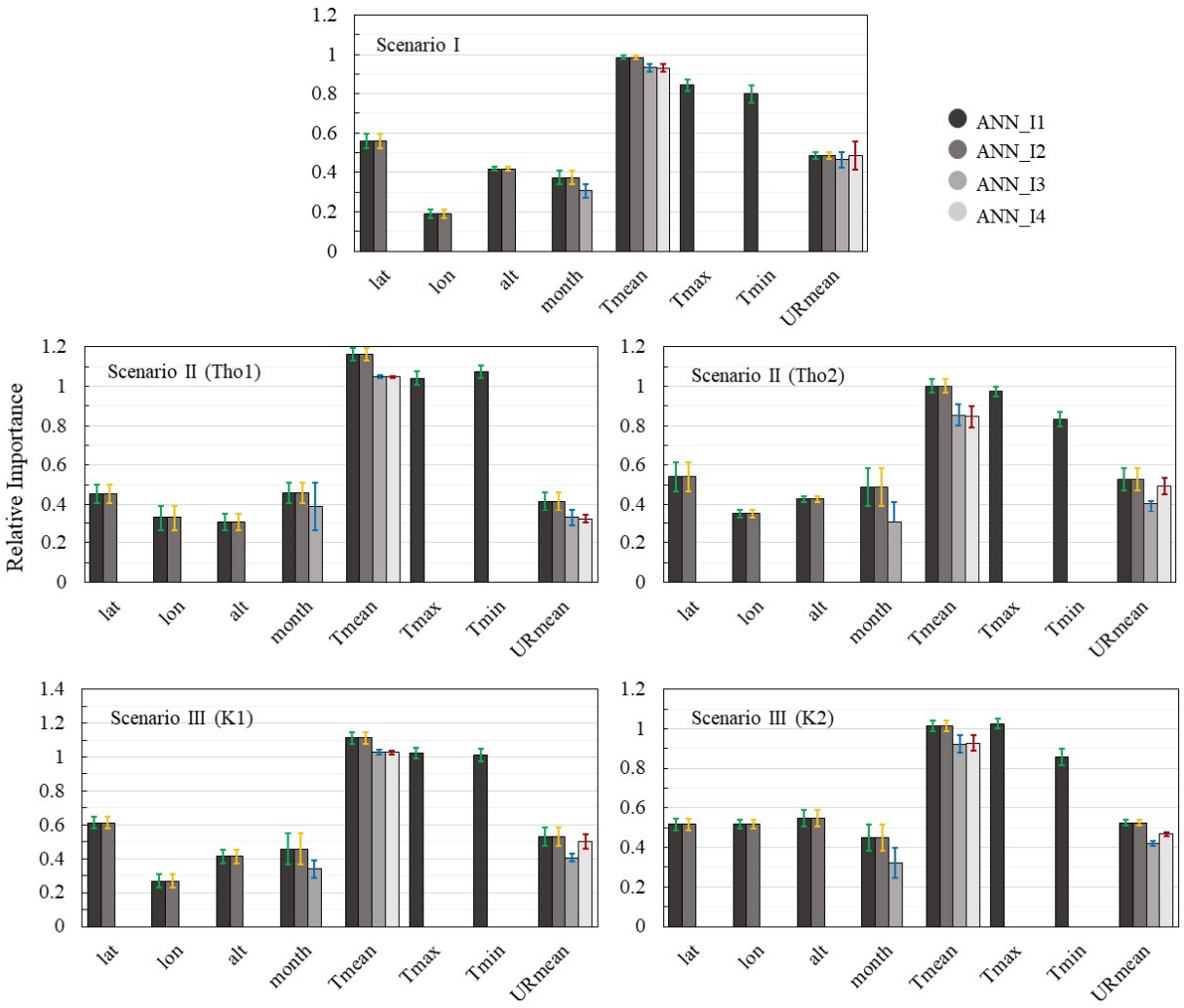


**Figure 4.** Importance of the input variables in ANN models. The bars represent standard deviation.

The division of the input data into two areas with climatic similarity aimed to increase the performance of the models. The division presented in SII and SIII managed to slightly increase the capacities of the models in relation to SI. However, this increase was only

observed in the Tho1 and K1 areas. Thus, we can infer that, although the division into areas with climatic similarity can reduce the amount of input data for training, in some situations this division is valid, and the models can respond more accurately. Machine-learning models developed for broader scenarios (e.g., SI) typically have reduced predictive capacity due to the high nonlinearity and low similarity of their input data; however, these models have greater ability to generalize [24]. According to [12], although the models developed locally perform better, these models may have low predictive capacity when used in other regions, since they may be highly specific to the location.

Regarding the importance of each input variable to the response variable of the evaluated algorithms, WEKA was used to select the attributes (Figures 4–6). Attributes were selected using the "ClassifierAttributeEval" tool associated with the "Ranker" method. These tools rank attributes by their individual evaluations. Correlation coefficient was the measure used to evaluate the performance of attribute combinations in the Ranker configuration. The same ranking method in WEKA was used by [51] in order to verify the importance of each input variable in solar radiation prediction.

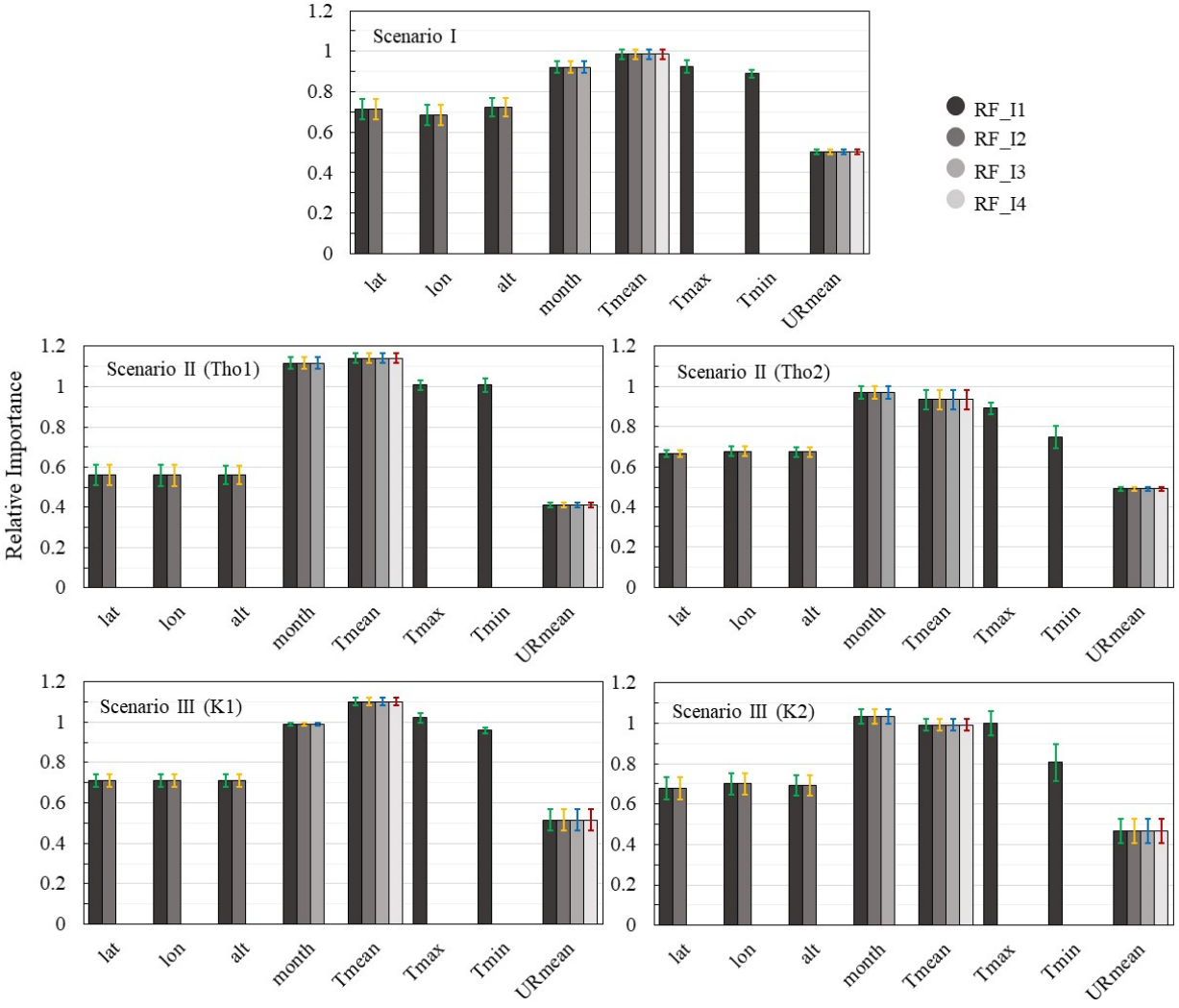


**Figure 5.** Importance of the input variables in RF models. The bars represent standard deviation.

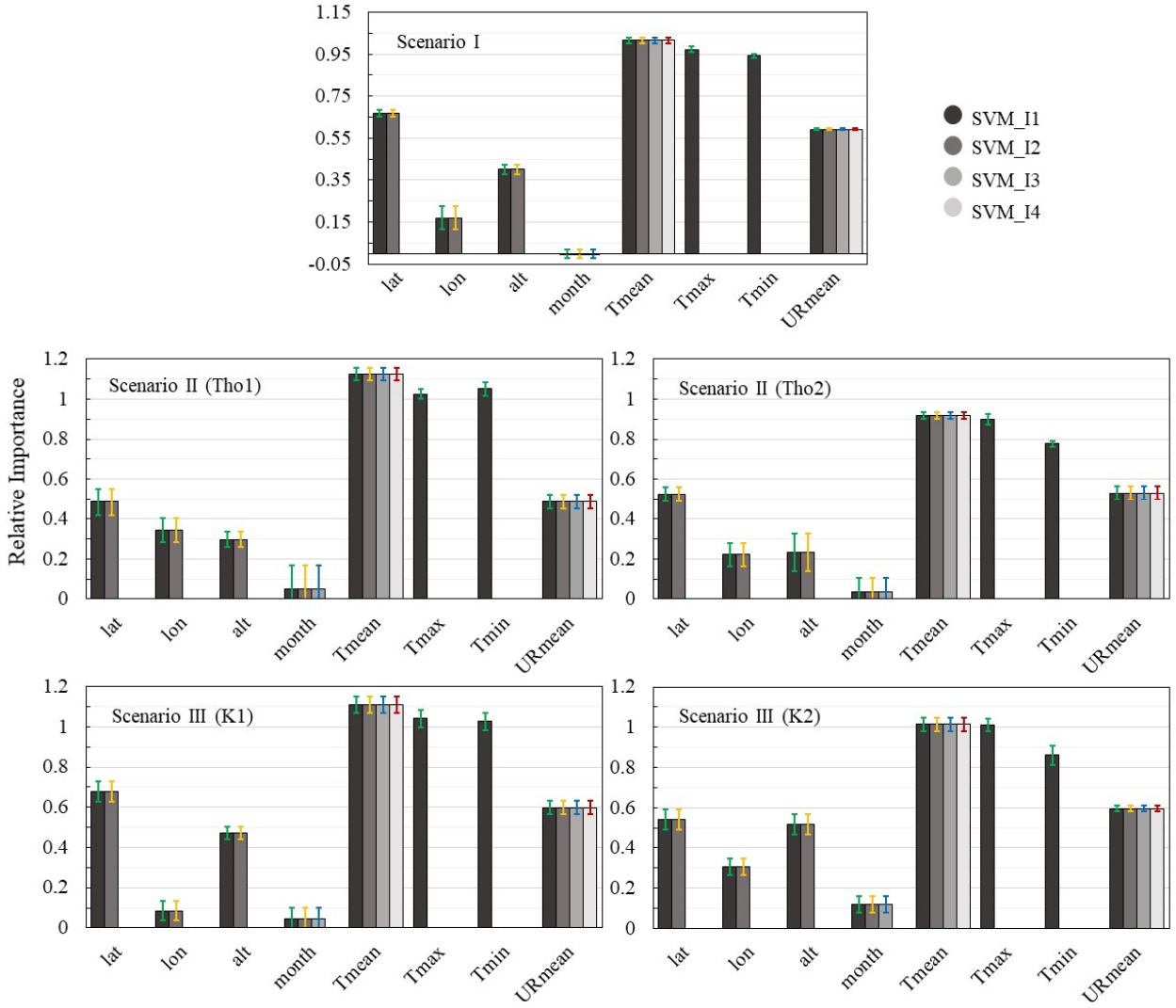


**Figure 6.** Importance of the input variables in SVM models. The bars represent standard deviation.

Different ANN settings were used for different input data (Table 2). These ANN settings resulted in different weights for each input attribute (Figure 4). However, similar behavior was observed in the different configurations. In all scenarios, Tmean, Tmax, and Tmin had greater weight in producing the estimate. In SIII K2, the relative importance of Tmax surpassed Tmed (Figure 4). This result may explain the decrease in ANN's performance in this scenario when Tmax and Tmin were removed (Table 4). The variables lat and month had a similar weight in all scenarios. Although similar, the removal of the month variable resulted in a greater reduction in ANN's performance when compared with the removal of the variables lat, lon, and alt.

The ranked values of each input variable in RF are shown in Figure 5. The Tmean and month variables had a higher weight in the $ET_0$ estimate. In SII Tho1 and SIII K2, the month variable was more important than the Tmean variable. This result may explain the drop in the RF model's performance when it removed the month variable ($I_3$ to $I_2$). The Tmax and Tmin variables also had a high weight in the $ET_0$ estimate. However, the removal of these variables increased the capacity of the RF model as observed (Tables 3–5) and discussed previously.

The relative importance of each input variable in SVM is shown in Figure 6. It was possible to observe that the Tmean, Tmax, and Tmin variables had a higher weight in the $ET_0$ estimate, followed by HR and lat. The month variable was of low importance in the $ET_0$ estimate. In SI, the month showed a negative weight. Therefore, this input data

can negatively impact the $ET_0$ estimate. In the performance results for the SVM model (Tables 3–5), there was no significant variation in performance when the month variable was removed. Both results make it possible to highlight that, for this region, the month variable does not contribute to the performance of the SVM model.

Although each model has a different pattern in the ranking of the input variables (Figures 4–6), air temperature was the most important attribute. The observed correlation between air temperature and $ET_0$ (Figure 4) may explain the importance of air temperature in this estimate. This behavior was not observed in SIII K2 or SII Tho2. However, in these scenarios, no significant difference was observed between the month and Tmean variables. Studying the ranking of the importance of meteorological variables based on the RF method, the three most important variables were insolation (n), Tmax, and RH [20]. The high relative importance observed corroborates the results of the present study.

The other variables presented different weights according to each model applied. These results indicate a peculiarity of the models. Hence, new research and applications can be based on these results, choosing the best method to suit the conditions of the input data. However, it is recommended that the models be previously experimented with using different input data; as noted, some variables may have a relatively high weight in the $ET_0$ estimate, but their use can decrease the predictive performance of the model. This behavior was observed when using the RF model. In this model, removal of the variables Tmax and Tmin increased predictive capacity, although these variables have shown high relative importance.

It is important to note that the month variable was highly important in estimation with RF. However, low importance was observed when the SVM model was used, since this variable was not correlated with $ET_0$ (Figure 3). These results highlight the need for more techniques to select the meteorological variables used in modeling. Linear regression alone is not sufficient to identify the relevance of the input data. Furthermore, different models may present different behaviors regarding classification of the importance of the input variable and still present satisfactory results.

Differently from the evaluation of the importance of the ANN, RF, and SVM attributes, for the MLR method, the attribute selection method was applied (the M5 method), which indicates the importance of each input attribute in the generated model. The adjusted coefficients are shown in Table 6. It was observed that, in some models, the method used (the M5 method) excluded the month variable. This behavior indicates a low importance of this variable in the MLR estimate. This result was similar to that observed in the analysis of the importance of the input variables in SVM. The exclusion of lat and Tmax was also observed in some cases.

The results presented in this study reveal that, for locations in Minas Gerais state, these models can be used safely. The ANN and RF models are recommended to estimate $ET_0$ when considering a wider range of input data, as they have better predictive capacity in this situation. The SVM and MLR models are recommended in situations where only temperature and relative humidity data are available. However, between these two models, MLR is recommended because it requires less computational effort. These models, although they have a high predictive capacity, cannot be perfect. Other meteorological variables not considered as input data (e.g., solar radiation, wind speed, and vapour-pressure deficit) and other factors (e.g., data recorded in error) contributed to a decrease in the predictive capacity of these models.

No statistical method or machine-learning method can produce results that are the same as the observed and/or recorded data. There will always be some error, no matter how small. Therefore, it is important that the meteorological stations function continuously. As in all studies, some limitations were noted in this study. One of the main limitations is related to difficulties in the availability of quality meteorological data. The malfunction and limited collection of meteorological data has been a limitation in several countries. Another limitation that can be observed is the difficulty and complexity of using some models. In

this context, it is recommended to evaluate and use models with good results and that present greater simplicity in their use.

**Table 6.** Coefficients of the multiple linear regression models in SI, SII, and SIII.

| | | **MLR Method Coefficients** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | lat | lon | alt | month | Tmax | Tmean | Tmin | RH | |
| | | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ | $\beta_0$ |
| SI | $I_8$ | −0.0208 | 0.0579 | 0.0016 | 0.0091 | 0.0758 | 0.2966 | −0.0396 | −0.02 | −1.8209 |
| | $I_6$ | −0.0222 | 0.0402 | 0.0013 | 0.0065 | - | 0.2972 | - | −0.0264 | −0.4453 |
| | $I_3$ | - | - | - | ø | - | 0.2262 | - | −0.0234 | 0.4921 |
| | $I_2$ | - | - | - | - | - | 0.2262 | - | −0.0234 | 0.4921 |
| SII | | Tho.1 (A, B4, B3, B2, and B1) | | | | | | | | |
| | $I_8$ | −0.0343 | 0.06 | 0.0012 | 0.0172 | 0.0633 | 0.3096 | −0.0532 | −0.0229 | −1.2174 |
| | $I_6$ | −0.0523 | 0.0294 | 0.0008 | 0.0159 | - | 0.2807 | - | −0.0278 | −0.7404 |
| | $I_3$ | - | - | - | 0.0159 | - | 0.2521 | - | −0.0234 | −0.0037 |
| | $I_2$ | - | - | - | - | - | 0.2498 | - | −0.0251 | 0.2709 |
| | | Tho.2 (C2, C1, and D) | | | | | | | | |
| | $I_8$ | ø | 0.0517 | 0.0017 | ø | ø | 0.3857 | −0.0447 | −0.0215 | −1.344 |
| | $I_6$ | ø | 0.0511 | 0.0016 | ø | - | 0.331 | - | −0.0247 | −0.6378 |
| | $I_3$ | - | - | - | ø | - | 0.2858 | - | −0.0297 | −0.6149 |
| | $I_2$ | - | - | - | - | - | 0.2858 | - | −0.0297 | −0.6149 |
| SIII | | K1 (Cwa and Cwb) | | | | | | | | |
| | $I_8$ | ø | 0.0713 | 0.0013 | 0.0149 | 0.091 | 0.2515 | −0.0203 | −0.0231 | −0.1477 |
| | $I_6$ | −0.0166 | 0.0461 | 0.0009 | 0.0122 | - | 0.2861 | - | −0.029 | 0.6759 |
| | $I_3$ | - | - | - | 0.0128 | - | 0.256 | - | −0.0243 | −0.0213 |
| | $I_2$ | - | - | - | - | - | 0.254 | - | −0.0257 | 0.2013 |
| | | K2 (Am and Aw) | | | | | | | | |
| | $I_8$ | −0.0428 | 0.0595 | 0.002 | ø | 0.066 | 0.3543 | −0.0588 | −0.0139 | −3.4505 |
| | $I_6$ | ø | 0.0316 | 0.0014 | ø | - | 0.3329 | - | −0.0208 | −1.7699 |
| | $I_3$ | - | - | - | ø | - | 0.3172 | - | −0.029 | −1.5306 |
| | $I_2$ | - | - | - | - | - | 0.3172 | - | −0.029 | −1.5306 |

ø: input data excluded by the M5 method.

The models developed in this study are expected to help decision-making by different professionals, mainly farmers. Agricultural companies are responsible for a considerable part of the Brazilian gross domestic product [52], and the Minas Gerais state had the third-largest gross domestic product in Brazil in 2018 [33]. The results of these models can assist in irrigation management, climatic zoning, and the construction of productivity models, among other applications. In addition, the approaches used in the present study have the potential to benefit the development of other types of models and studies from other regions.

## 4. Conclusions

The results and information presented in this study are important for planning and determining the use of the best model to estimate $ET_0$ for a region with limited climate data. The use of input data combination I6 (alt, lat, lon, month, Tmean, and RH) in the scenarios SI, SII, and SIII provided in general the best results in $ET_0$ estimation between the evaluated models, so this data combination is recommended to be used. The RF and ANN models presented the highest predictive ability for the $ET_0$ estimate. Both models, in best-case scenarios, with the input data combination I6 or I8, explain more than 96% of the variability of the variables estimated using the independent dataset. If experimental data are available, machine-learning algorithms represent a powerful tool able to provide accurate predictions. The SVM and MLR models are recommended for all scenarios in

situations with limited climatic data where only air temperature and relative humidity data are available. Although dividing into scenarios results in less input data for model training, the SII and SIII scenarios showed slightly better results in the southern areas of the Minas Gerais state.

Air temperature was the meteorological input variable that presented the greatest relative importance, while the month variable presented the greatest variation in importance in relation to the model used. Therefore, it is concluded that although temperature is fundamental for the estimation of $ET_0$, other variables can present different levels of importance in the prediction of $ET_0$.

The results presented in this study contribute relevant information, and, together with other studies, can serve as a basis for the estimation of reference evapotranspiration. However, new studies are necessary in order to evaluate new models and their performance with limited climate data, based on other machine-learning algorithms that contemplate different climatic conditions and that subsequently take into account the effects of climate change.

**Author Contributions:** Conceptualization, P.A.B.d.S. and F.S.; methodology and formal analysis, L.G.d.C., V.B.d.S.B., D.B.M., and G.A.e.S.F.; data curation, P.A.B.d.S., D.B.M., and F.S.; writing—original draft preparation, P.A.B.d.S.; writing—review and editing, L.C., G.B., G.R., and F.S.; visualization, D.B.M. and V.B.d.S.B.; supervision, G.A.e.S.F., L.C., G.B., G.R., and F.S.; project administration, P.A.B.d.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES—Finance Code 001) for the research and study grants.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Souza, L.S.B.; Silva, M.T.L.; Alba, E.; Moura, M.S.B.; Cruz Neto, J.F.; Souza, C.A.A.; Silva, T.G.F. New method for estimating reference evapotranspiration and comparison with alternative methods in a fruit-producing hub in the semi-arid region of Brazil. *Theor. Appl. Climatol.* **2022**, *149*, 593–602. [CrossRef]
2. Wen, X.; Si, J.; He, Z.; Wu, J.; Shao, H.; Yu, H. Support-vector-machine-based models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions. *Water Resour. Manag.* **2015**, *29*, 3195–3209. [CrossRef]
3. Yassin, M.A.; Alazba, A.A.; Mattar, M.A. Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. *Agric. Water Manag.* **2016**, *163*, 110–124. [CrossRef]
4. Kumar, M.; Raghuwanshi, N.S.; Singh, R.; Wallender, W.W.; Pruitt, W.O. Estimating evapotranspiration using artificial neural network. *J. Irrig. Drain. Eng.* **2002**, *128*, 224–233. [CrossRef]
5. Ning, M.; Jozsef, S.; Zhang, Y. Calibration-free complementary relationship estimates terrestrial evapotranspiration globally. *Water Resour. Res.* **2021**, *57*, e2021WR029691.
6. Yu, L.; Qiu, G.Y.; Yan, C.; Zhao, W.; Zou, Z.; Ding, J.; Xiong, Y. A global terrestrial evapotranspiration product based on the three-temperature model with fewer input parameters and no calibration requirement. *Earth Syst. Sci. Data* **2022**, *14*, 3673–3693. [CrossRef]
7. Almorox, J.; Quej, V.H.; Martí, P. Global performance ranking of temperature-based approaches for evapotranspiration estimation considering Köppen climate classes. *J. Hydrol.* **2015**, *528*, 514–522. [CrossRef]
8. Yang, Q.; Ma, Z.; Zheng, Z.; Duan, Y. Sensitivity of potential evapotranspiration estimation to the Thornthwaite and Penman–Monteith methods in the study of global drylands. *Adv. Atmos. Sci.* **2017**, *34*, 1381–1394. [CrossRef]
9. Ewaid, S.H.; Abed, S.A.; Al-Ansari, N. Crop water requirements and irrigation schedules for some major crops in Southern Iraq. *Water* **2019**, *11*, 756. [CrossRef]
10. Xiang, K.; Li, Y.; Horton, R.; Feng, H. Similarity and difference of potential evapotranspiration and reference crop evapotranspiration—A review. *Agric. Water Manag.* **2020**, *232*, 106043. [CrossRef]
11. Benali, L.; Notton, G.; Fouilloy, A.; Voyant, C.; Dizene, R. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renew. Energy* **2019**, *132*, 871–884. [CrossRef]
12. Ferreira, L.B.; Cunha, F.F.; Oliveira, R.A.; Fernandes Filho, E.I. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM–A new approach. *J. Hydrol.* **2019**, *572*, 556–570. [CrossRef]
13. Malik, A.; Kumar, A.; Ghorbani, M.A.; Kashani, M.H.; Kisi, O.; Kim, S. The viability of co-active fuzzy inference system model for monthly reference evapotranspiration estimation: Case study of Uttarakhand State. *Hydrol. Res.* **2019**, *50*, 1623–1644. [CrossRef]

14. Sattari, M.T.; Apaydin, H.; Band, S.S.; Mosavi, A.; Prasad, R. Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 603–618. [CrossRef]
15. Ferreira, L.B.; Da Cunha, F.F. New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. *Agric. Water Manag.* **2020**, *234*, 106–113. [CrossRef]
16. Yin, Z.; Wen, X.; Feng, Q.; He, Z.; Zou, S.; Yang, L. Integrating genetic algorithm and support vector machine for modeling daily reference evapotranspiration in a semi-arid mountain area. *Hydrol. Res.* **2017**, *48*, 1177–1191. [CrossRef]
17. Martí, P.; González-Altozano, P.; Gasque, M. Reference evapotranspiration estimation without local climatic data. *Irrig. Sci.* **2011**, *29*, 479–495. [CrossRef]
18. Nourani, V.; Elkiran, G.; Abdullahi, J. Multi-station artificial intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements. *J. Hydrol.* **2019**, *577*, 123958. [CrossRef]
19. Feng, Y.; Cui, N.; Gong, D.; Zhang, Q.; Zhao, L. Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. *Agric. Water Manag.* **2017**, *193*, 163–173. [CrossRef]
20. Wang, S.; Lian, J.; Peng, Y.; Hu, B.; Chen, H. Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China. *Agric. Water Manag.* **2019**, *221*, 220–230. [CrossRef]
21. Zhou, X.; Zhu, X.; Dong, Z.; Guo, W. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop. J.* **2016**, *4*, 212–219.
22. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: Berlin/Heidelberg, Germany, 2013; p. 188.
23. Mohammadrezapour, O.; Piri, J.; Kisi, O. Comparison of SVM, ANFIS and GEP in modeling monthly potential evapotranspiration in an arid region (Case study: Sistan and Baluchestan Province, Iran). *Water Supply* **2019**, *19*, 392–403. [CrossRef]
24. Shiri, J.; Nazemi, A.H.; Sadraddini, A.A.; Landeras, G.; Kisi, O.; Fard, A.F.; Marti, P. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. *Comput. Electron. Agric.* **2014**, *108*, 230–241. [CrossRef]
25. Valipour, M.; Gholami Sefidkouhi, M.A.; Raeini−Sarjaz, M. Selecting the best model to estimate potential evapotranspiration with respect to climate change and magnitudes of extreme events. *Agric. Water Manag.* **2017**, *180*, 50–60. [CrossRef]
26. Wang, Z.; Xie, P.; Lai, C.; Chen, X.; Wu, X.; Zeng, Z.; Li, J. Spatiotemporal variability of reference evapotranspiration and contributing climatic factors in China during 1961–2013. *J. Hydrol.* **2017**, *544*, 97–108. [CrossRef]
27. Dou, X.; Yang, Y. Evapotranspiration estimation using four different ma-chine learning approaches in different terrestrial ecosystems. *Comput. Electron. Agric.* **2018**, *148*, 95–106. [CrossRef]
28. Pozníková, G.; Fischer, M.; van Kesteren, B.; Orság, M.; Hlavinka, P.; Žalud, Z.; Trnka, M. Quantifying turbulent energy fluxes and evapotranspiration in agricultural field conditions: A comparison of micrometeorological methods. *Agric. Water Manag.* **2018**, *209*, 249–263. [CrossRef]
29. Tang, D.; Feng, Y.; Gong, D.; Hao, W.; Cui, N. Evaluation of artificial intelligence models for actual crop evapotranspiration modeling in mulched and non-mulched maize croplands. *Comput. Electron. Agric.* **2018**, *152*, 375–384. [CrossRef]
30. Zhang, Z.; Gong, Y.; Wang, Z. Accessible remote sensing data based reference evapotranspiration estimation modelling. *Agric. Water Manag.* **2018**, *210*, 59–69. [CrossRef]
31. Granata, F. Evapotranspiration evaluation models based on machine learning algorithms—A comparative study. *Agric. Water Manag.* **2019**, *217*, 303–315. [CrossRef]
32. Chen, Z.; Zhu, Z.; Jiang, H.; Sun, S. Estimating daily reference evapotran-spiration based on limited meteorological data using deep learning and classical machine learning methods. *J. Hydrol.* **2020**, *591*, 125286. [CrossRef]
33. IBGE. Instituto Brasileiro de Geografia e Estatística. 2022. Available online: https://cidades.ibge.gov.br/brasil/mg (accessed on 28 April 2022).
34. Allen, R.G.; Pereira, L.S.; Raes, D.; Smith, M. *Crop Evapotranspiration—Guidelines for Computing Crop Water Requirements*; FAO Irrigation and Drainage Paper 56; FAO: Rome, Italy, 1998; 297p.
35. Thornthwaite, C.W. An approach toward a rational classification of climate. *Geogr. Rev.* **1948**, *38*, 55–94. [CrossRef]
36. Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; Gonçalves, J.D.M.; Sparovek, G. Köppen's climate classification map for Brazil. *Meteorol. Z.* **2013**, *22*, 711–728. [CrossRef] [PubMed]
37. Martí, P.; González-Altozano, P.; López-Urrea, R.; Mancha, L.A.; Shiri, J. Modeling reference evapotranspiration with calculated targets. Assessment and implications. *Agric. Water Manag.* **2015**, *149*, 81–90. [CrossRef]
38. Fausett, L. *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*; Pearson Education India Editora: Chennai, India, 1994; 461p.
39. Berar, D. Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 542–545.
40. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
41. Xu, Y.; Knudby, A.; Ho, H.C. Estimating daily maximum air temperature from MODIS in British Columbia, Canada. *Int. J. Remote Sens.* **2014**, *35*, 8108–8121. [CrossRef]
42. Wang, H.; Lei, M.; Chen, Y.; Li, M.; Zou, L. Intelligent identification of maceral components of coal based on image segmentation and classification. *Appl. Sci.* **2019**, *9*, 3245. [CrossRef]
43. Schumacher, B.L.; Burchfield, E.K.; Bean, B.; Yost, M.A. Leveraging Important Covariate Groups for Corn Yield Prediction. *Agriculture* **2023**, *13*, 61. [CrossRef]

44. Samadianfard, S.; Asadi, E.; Jarhan, S.; Kazemi, H.; Kheshtgar, S.; Kisi, O.; Manaf, A.A. Wavelet neural networks and gene expression programming models to predict short-term soil temperature at different depths. *Soil Tillage Res.* **2018**, *175*, 37–50. [CrossRef]
45. Feng, Q.; Wen, X.; Li, J. Wavelet analysis-support vector machine coupled models for monthly rainfall forecasting in arid regions. *Sustain. Water Resour. Manag.* **2015**, *29*, 1049–1065. [CrossRef]
46. Mehdizadeh, S.; Behmanesh, J.; Khalili, K. Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. *Comput. Electron. Agric.* **2017**, *139*, 103–114. [CrossRef]
47. Bowden, G.J.; Dandy, G.C.; Maier, H.R. Input determination for neural network models in water resources applications. Part 1—Background and methodology. *J. Hydrol.* **2005**, *301*, 75–92.
48. Maier, H.R.; Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environ. Model Softw.* **2000**, *15*, 101–124. [CrossRef]
49. Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; Moraes Gonçalves, J.L. Modeling monthly mean air temperature for Brazil. *Theor. Appl. Climatol.* **2013**, *113*, 407–427. [CrossRef]
50. Ozgoren, M.; Bilgili, M.; Sahin, B. Estimation of global solar radiation using ANN over Turkey. *Expert. Syst. Appl.* **2012**, *39*, 5043–5051. [CrossRef]
51. Yadav, A.K.; Malik, H.; Chandel, S.S. Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models. *Renew. Sust. Energ. Rev.* **2014**, *31*, 509–519. [CrossRef]
52. Brugnaro, R.; Bacha, C.J.C. Analysis of increased participation of agriculture in the Brazilian GDP from 1994 a 2004. In Proceedings of the Congress of the European Regional Science Association, Volos, Greece, 30 August–3 September 2006; Volume 46, p. 19.