PIETROS ANDRE BALBINO DOS SANTOS

ESTIMATION OF AIR TEMPERATURE AND REFERENCE EVAPOTRANSPIRATION IN THE MINAS GERAIS STATE BY DIFFERENT METHODS

LAVRAS – MG

2021

PIETROS ANDRE BALBINO DOS SANTOS

ESTIMATION OF AIR TEMPERATURE AND REFERENCE EVAPOTRANSPIRATION IN THE MINAS GERAIS STATE BY DIFFERENT METHODS.

Thesis submitted for the degree of Doctor of Science in Irrigation and Drainage Engineering, Water Resources Graduate Program, Federal University of Lavras, Brazil.

Prof. Dr. Luiz Gonsaga de Carvalho (UFLA)
Supervisor

LAVRAS – MG

2021

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Santos, Pietros Andre Balbino dos.
Estimation of air temperature and reference evapotranspiration in the Minas Gerais state by different methods / Pietros Andre Balbino dos Santos. - 2021.
88 p. : il.

Orientador(a): Luiz Gonsaga de Carvalho.

Tese (doutorado) - Universidade Federal de Lavras, 2021.
Bibliografia.

1. Air temperature. 2. Evapotranspiration. 3. Machine learning models. I. Carvalho, Luiz Gonsaga de. II. Título.

PIETROS ANDRÉ BALBINO DOS SANTOS

ESTIMATION OF AIR TEMPERATURE AND REFERENCE EVAPOTRANSPIRATION IN THE MINAS GERAIS STATE BY DIFFERENT METHODS.

ESTIMATIVA DA TEMPERATURA DO AR E DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA NO ESTADO DE MINAS GERAIS POR DIFERENTES MÉTODOS

Thesis submitted for the degree of Doctor of Science in Irrigation and Drainage Engineering, Water Resources Graduate Program, Federal University of Lavras, Brazil.

APROVED in May 18, 2021.

Dr. Adriano Valentim Diotto	Federal University of Lavras (UFLA)
Dr. Felipe Schwerz	Federal University of Lavras (UFLA)
Dr. Luiz Gonsaga de Carvalho	Federal University of Lavras (UFLA)
Dr. Wezer Lismar Miranda	Federal Institute of Baiano (IFBaiano)
Dr. Wilian Soares Lacerda	Federal University of Lavras (UFLA)

Prof. Dr. Luiz Gonsaga de Carvalho (UFLA)
Supervisor

LAVRAS – MG

2021

This thesis is dedicated to my father, Edilson "*Canhoto"* (*In memoriam*), and to my grandmother, Terezinha (*In memoriam*), with love.

DECLARATION

I hereby declare that this work has been originally produced by myself for this thesis and it has not been submitted for the award of a higher degree to any other institution. Collaborations with other researchers, as well as publications or submissions for publication are properly acknowledged throughout the document.

Pietros André Balbino dos Santos, Lavras, May 2021.

ACKNOWLEDGEMENTS

A special thanks to my parents Francisca and Edilson (In memoriam), who supported me in every step. My love Jéssica, my brother Pacelly and his family (Killiian and Clarice), who always believed in me, and gave me bravery to finish this task. And everyone in the Santos and Balbino family.

I would like to thank Federal University of Lavras, in particular the Water Resources Graduate Program, the Engineering Department, and to its faculty and administrative technicians for the knowledge transmitted, and for the physical space and equipment made available for studies.

I also recognize my gratitude to CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil) for their financial support and scholarships. This study was financed in part by the CAPES – Finance Code 001.

I would like to express my eternal gratitude to my supervisor Luiz Gonsaga, for friendship, teaching, and support, especially in times of difficulty.

A special thanks for Cássio Ussi Monti, Felipe Schwerz, Wilian Soares Lacerda, Thiago Henrique, Gustavo Jardim, Matheus, Victor Buono, Inara, Stefany, Jéssica Peres and Kaka for supporting in the experiments and writing the papers.

I would like to thank my judge committee: Adriano Valentim Diotto, Felipe Schwerz, Luiz Gonsaga de Carvalho, Wezer Lismar Miranda, and Wilian Soares Lacerda.

Special thanks to my childhood friends, from my city, my friends I conquered at UFLA and everyone who contributed to the realization of this work.

ABSTRACT

Consistent weather data is obtained by weather stations. These data are important for different fields of science such as climatology, irrigation, and hydrology. The meteorological element and the agrometeorological parameter air temperature and evapotranspiration (ET), respectively, are fundamental for studies in these fields. The temperature indicates the amount of energy available in the water-soil-atmosphere system. This energy can influence various processes on the Earth's surface, among them the growth and development of plants. ET is the process of water transportation from a vegetated surface to the atmosphere including the evaporation and transpiration process. These meteorological variables and agrometeorological parameter can be monitored daily in weather stations, however, in the Minas Gerais State, the coverage of the weather stations network is limited. Besides, interruptions and errors in the database are quite common. In this sense, this research aimed to develop models that can reliably estimate air temperature and evapotranspiration through easily obtained input data such as geographic coordinates. As described in paper 1, the aim was to develop models of multiple linear regression (MLR), artificial neural network (ANN), and random forest (RF) to estimate the mean (Tmean), maximum (Tmax), and minimum (Tmin) monthly air temperatures as a function of geographic coordinates, altitude, and month for different localities in the Minas Gerais State, Brazil, with Köppen's climatic classification Cwa or Cwb. The Tmax, Tmean and Tmin data were extracted from national network of climatological stations (INMET). MLR was implemented using the data analysis tool in Microsoft Excel®. ANN and RF were implemented using the WEKA. The results showed that the algorithms RF and ANN were used to estimate Tmean, Tmax, and Tmin with high accuracy. The best results were obtained using the RF model. The MLR did not present a good accuracy. In paper 2, the aim was to evaluates the performance of ANN, RF, Support Vector Machine (SVM) and MLR to estimate the monthly mean reference evapotranspiration (ET_0) with four different input data combinations (I_8, I_6, I_3 and I_2) and in three scenarios: (SI) at the state level, where all climatological stations were used; and at regional level (SII and SIII), where the Minas Gerais state was divided into two areas according to the climatic classification of each climatological stations. The climatic classifications proposed by Thornthwaite (SII) and by Köppen (SIII) were used. All models were implemented by WEKA. The results showed that ANN and RF performed better in SI, II, III with the I_8 (latitude, longitude, altitude, month, Tmean, Tmax, Tmin, and relative humidity) or I_6 (latitude, longitude, altitude, month, Tmean, and relative humidity) input data. The SVM and MLR performed better in all scenarios when only two input variables were used (I_2 - mean temperature and relative humidity). Although dividing into scenarios results in less input data for models training, SII and SIII showed a slightly better result in the southern areas of the Minas Gerais state.

Keywords: Artificial Neural Network. Random Forest. Support Vector Machine**.** Multiple Linear Regression.

RESUMO

Dados meteorológicos consistentes são obtidos por estações meteorológicas. Esses dados são importantes para diferentes campos da ciência, como climatologia, irrigação e hidrologia. O elemento meteorológico e o parâmetro agrometeorológico: temperatura do ar e a evapotranspiração (ET), respectivamente, são fundamentais para estudos nesses campos. A temperatura indica a quantidade de energia disponível no sistema água-solo-atmosfera. Essa energia pode influenciar vários processos na superfície da Terra, entre eles o crescimento e o desenvolvimento das plantas. ET é o processo de transporte de água de uma superfície vegetada para a atmosfera, que inclui o processo de evaporação e de transpiração. Essas variáveis meteorológicas podem ser monitoradas diariamente em estações meteorológicas, porém, no Estado de Minas Gerais, a cobertura da rede de estações meteorológicas é limitada. Além disso, interrupções e erros no banco de dados são bastante comuns. Nesse sentido, com esta pesquisa objetivou-se desenvolver modelos que possam estimar com segurança a temperatura do ar e a evapotranspiração por meio de dados de entrada de fácil obtenção, como coordenadas geográficas. Conforme descrito no artigo 1, o objetivo foi desenvolver modelos de regressão linear múltipla (RLM), rede neural artificial (RNA) e floresta aleatória (FA) para estimar as temperaturas média (Tmean), máxima (Tmax) e mínima (Tmin) mensais do ar em função de coordenadas geográficas, altitude e mês para diferentes localidades do Estado de Minas Gerais, Brasil, com classificação climática, segundo Köppen, Cwa ou Cwb. Os dados de Tmax, Tmean e Tmin foram extraídos da rede nacional de estações climatológicas (INMET). A RLM foi implementada por meio da ferramenta de análise de dados do Microsoft Excel®. RNA e FA foram implementadas usando o WEKA. Os resultados mostraram que os algoritmos FA e RNA foram usados para estimar Tmean, Tmax e Tmin com alta precisão. Os melhores resultados foram obtidos com o modelo FA. A RLM não apresentou uma boa acurácia. No artigo 2, o objetivo foi avaliar o desempenho da RNA, FA, Máquina de vetor de suporte (MVS) e RLM para estimar a evapotranspiração de referência média mensal (ET_0) com quatro combinações diferentes de dados de entrada (I_8, I_6, I_3 e I_2) e em três cenários: (SI) a nível estadual, onde todas as estações climatológicas foram utilizadas; a nível regional (SII e SIII), onde o estado de Minas Gerais foi dividido em duas áreas de acordo com a classificação climática de cada estação climatológica. Foram utilizadas as classificações climáticas propostas por Thornthwaite (S II) e por Köppen (SIII). Todos os modelos foram implementados no software WEKA. Os resultados mostraram que RNA e a FA tiveram melhor desempenho nos SI, SII, SIII com I_8 (latitude, longitude, altitude, mês, Tmédia, Tmax, Tmin, e umidade relativa do ar) ou I_6 (latitude, longitude, altitude, mês, Tmédia e umidade relativa do ar). A MVS e a RLM tiveram melhor desempenho em todos os cenários quando apenas duas variáveis de entrada foram usadas (I_2 - Tmédia e umidade relativa). Embora a divisão em cenários resulte em menos dados de entrada para o treinamento de modelos, os SII e SIII mostraram um resultado ligeiramente melhor nas áreas mais ao Sul do estado de Minas Gerais.

Palavras-Chave: Rede Neural Artificial, Floresta Aleatória, Máquina de Vetor de Suporte, Regressão Linear Múltipla.

SUMARY

1ST CHAPTER

1 GENERAL INTRODUCTION

Minas Gerais state is the fourth largest in territorial extension in Brazil. Minas Gerais has a territory of 586,513.993 km² (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATISTICA - IBGE, 2020a). It is in the southeastern region of Brazil, between the parallels of 14° 13' 58" and 22° 54' 00", of south latitude, and the meridians of 39° 51' 32" and 51° 02' 35" a west of Greenwich. The Minas Gerais state is the second most populous in the country and has the third largest Gross Domestic Product in Brazil of 2018 (IBGE, 2020b). The state has agriculture as its primary sector. In 2018, the economic result of Minas Gerais was strongly related to the performance of the agricultural sector within the scope of agricultural activity. Among the agricultural activities, the coffee crop stands out, which represents more than a third of the gross value of state agricultural production (FUNDAÇÃO JOÃO PINHEIRO, 2020). Thus, several researches are being conducted with the aim of further expanding the productive capacity of the state.

Climatic studies are important allies in the state's agricultural growth. These studies are significant for producers to make decisions regarding activities and property protection (SOARES et al., 2018). In general, the meteorological data used in surveys are obtained from weather stations networks from agencies such as National Institute of Meteorology (INMET) and National Water Agency (ANA). Although the climatological data series from these weather stations networks are extensive and reliable, the coverage of the network is limited. In addition, failures, interruptions, and errors are quite common. The errors can be attributed to reading errors, damaged devices, and other unintended observational problems (DUMEDAH; COULIBALY, 2011; MWALE; ADELOYE; RUSTUM, 2012). According to Thom (1966), the interruptions that occur in the climatological data series do not make them unfeasible. These gaps are not filled since it is not possible to estimate the lost data without changing the frequency distribution dispersion scale.

The weather stations provide different quantitative data that indicate the current weather state in each location (KUSRIYANTO; PUTRA, 2018). Among the data obtained are air temperature, relative humidity, precipitation, and atmospheric pressure. The air temperature, typically measured at the height of about 2 m above ground, is one of the most important meteorological elements (JANATIAN et al., 2017). The air temperature exhibits spatiotemporal patterns that can be highly variable and complex (BENALI et al., 2012). This

variable indicates the amount of energy available to the system. Changes in air temperature can alter all processes on the Earth's surface. These changes are visible in thermal comfort of people in urban spaces (TALEGHANI, 2018), and physiological processes occurring in a plant, such as the speed of chemical reactions (BENAVIDES et al., 2007).

All plant species have an ideal air temperature range for development. This temperature range represented by a minimum, maximum, and optimum (HATFIELD; PRUEGER, 2015). When the air temperature exceeds the ideal range for each species, morphological, physiological, and biochemical changes may be induced, leading to adverse effects on plant growth (WAHID et al., 2007). Schlenker and Roberts (2009) indicated that in the temperature range up to 29 °C to 32 °C, the yield for corn, soybean, and cotton would gradually increase and then yield would decline sharply with temperature increases beyond this range. In coffee crop, the optimum mean annual temperature falls in the range of 18 to 23 °C for the proper growth of *C. Arabica* specie. The optimum temperature falls in the range of 22 to 26 °C for the proper growth of *C. Canephora* (DAMATTA et al., 2018). Temperature extremely low or very high may cause a reduction in activity of the coffee crop and decrease in the net photosynthetic rates of the leaves (BATISTA-SANTOS et al., 2011; CANNELL, 1985; PARTELLI et al., 2009). These studies show the importance of air temperature for the growth of agriculture. However, not only meteorological variables are important, but agrometeorological parameter such as evapotranspiration (ET) are of great importance in crop development and hydrological studies.

ET is an important component of the water balance (KUMAR; RAGHUWANSHI; SINGH, 2011; WILCOX et al., 2003; XIANG et al., 2020). ET divides into two individual components: water evaporation from the soil and from water intercepted by the plant canopy and transpiration through the stomata of plants. Although both components are important, only transpiration associated with plant yield. Transpiration is commonly considered the more desirable component aiming to increase the water use efficiency (AGAM et al., 2012). Nevertheless, evaporation is a micro-climate moderator, and some studies claim that this ability may indirectly benefit growth and yield of the plants (BURT et al., 2005).

ET can be measured using the lysimeter (evaporimeter) or water balance approach. These methods are ways to get ET directly and are typically used in the development and validation of other methods (ALLEN et al., 2011). However, not always possible to use. The lysimeter and water balance are a time-consuming method and needs precisely and carefully

planned experiments (KUMAR et al., 2002). Furthermore, these approaches may require skilled labor and high investments. Therefore, indirect estimation methods from weather data are used. These methods vary from an empirical relationship to combination methods based on physical processes. These indirect estimation methods have been the main way to obtain ET.

ET can be found in the literature as crop evapotranspiration (ET_c) or reference crop evapotranspiration (ET_0). Both concepts measure the transfer rate of water from the soil plant system to the atmosphere. However, ET_c measures ET for any crop, while ET_0 is the ET rate from a reference crop surface. One of the ways to obtain the ET_c is through the ET_0 and the crop coefficient (Kc) (CARVALHO et al., 2011; SALAM et al., 2020).

ET_0 has also been applied in climatology (ALMOROX; QUEJ; MARTÍ, 2015; YANG et al., 2017), agronomy field (EWAID; ABED; AL-ANSARI, 2019; ISMAIL; EL-NAKHLAWY, 2018), hydrology (MOJID; RANNU; KARIM, 2015; PHAM et al., 2018) among other science fields. ET_0 originated from the of Penman (1948) equation. This equation is based on physical processes and compute the evaporation from an open water surface from climatological data. Later, Monteith (1965) introduced a surface conductance term that accounted for the response of leaf stomata to its hydrologic environment to the Penman equation. This combination method extended the application of the equation to cropped surfaces and gave rise to Penman-Monteith evapotranspiration equation.

In the 1990s, the Food and Agriculture Organization (FAO) held a meeting with experts to analyze the concepts and procedures in ET calculate. From then on, the Penman-Monteith equation was established as the recommended method for ET_0 estimate. However, the Penman-Monteith equation was adapted. Specific and invariable parameters for crop were established according to the suggestions proposed by Allen et al. (1998), creating a reference crop. The FAO adopted this ET_0 concept formally in FAO Irrigation and Drainage Paper N 56 (ALLEN et al., 1998). Thus, with the crop parameterization, ET_0 is obtained only through meteorological data.

The FAO Penman-Monteith equation is a nonlinear and complex method. This method is considered more realistic physically and is well accepted in practical applications and academic research. However, it requires some additional meteorological variables when compared to other methods (YANG et al., 2017). This dependence on several meteorological data (temperature, relative humidity, solar radiation, soil heat flow, atmospheric pressure, and

wind speed) associated with the limited of surface weather stations network makes it difficult to measure ET_0.

Given the importance of T and of ET_0, besides the difficulties already mentioned in obtaining this data for any location and time, several studies have investigated alternative equations and methods to estimate T and ET_0 with reduced data requirement. Classical regression analysis (e.g., Linear regression and Multiple linear regression -MLR) has been used, however it has some limitations (ALVARES et al., 2013; MALIK et al., 2019). Classical regressions analysis is recommended as a reference method. The MLR was used as a reference classical statistical method in this study. In recent years, machine learning models have been high capacity in estimating and forecasting meteorological data. These models can capture complex relationships between input and output data. In this study, we selected this machine learning models: Artificial Neural Network (ANN), Random Forest (RF) and Support Vector Machine (SVM) because these models showed high predictive capacity in the T and ET0 estimate in some articles (FERREIRA; DA CUNHA, 2020; HUANG et al., 2019; MOREIRA; CECÍLIO, 2016; NOI; DEGENER; KAPPAS, 2017).

MLR is developed to formulate the complex input–output data relationship (WANG; HUANG; HE, 2012). MLR aim at explaining the collinearity between a dependent variable and independent by means of a linear combination of predictors independent variables (more than one). This model requires data from the past projects in order to evaluate the current projects (LEUNG; FAN, 2002). To assess the adequacy of the model, the statistical indicators as coefficient of determination R^2 they are used. According to (MARTÍ; GONZÁLEZ-ALTOZANO; GASQUE, 2011), using the same input data, the studied ANNs present very similar accuracy indicators and performance trends as the MLR models in the ET_0 estimate. This model also had good accuracy in estimating monthly air temperature (ALVARES et al., 2013). However, the complexity of the input data can change this performance.

The ANN is a promising and effective tool for non-linear modeling and complex time-series. The ANN has performance characteristics resembling biology of the human brain that learn from trial and error. ANNs, in general, have an architecture with connections between neurons (neural networks) and methods to determine the connections weight (PATIL; DEKA, 2016). Its customary architecture is composed of three layers: input, hidden, and output layers and each layer include an array of processing elements (FERREIRA et al., 2019; KUMAR et al., 2002; YIN et al., 2017). However, different architectures must be tested to achieve

maximum predictive performance. In this study, the ANN of the feed-forward multilayer perceptron (MLP) type was used (FAUSETT, 1994). In the training process of this ANN the input sign spreads layer-by-layer forward (forward pass) and, posteriorly, the sign is backpropagated for the correction of the error (reverse pass). ANN showed performed better than the others traditional machine learning models in certain scenarios and regions (SATTARI et al., 2021). However, in some situations the model does not perform well (BENALI et al., 2019).

The RF method belongs to the regression tree (RT) family. RF is an ensemble learning technique based on a collection of tree predictors (XU; KNUDBY; HO, 2014). RF uses the Breiman's "bagging" idea to ensemble many decision trees into a single but strong model (BREIMAN, 2001). It is a combination of many predictor trees (forest), in which each tree is generated from a random vector, sampled independently and with the same distribution for all trees in the forest. According to Wang et al. (2019), 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; (iii) and predict new data by aggregating the predictions of the n. As reported by (BENALI et al., 2019), this technique is recognized as one of the most effective machine learning models for forecasting. But, like ANN, in some scenarios the RF did not perform better than other models (FERREIRA; DA CUNHA, 2020).

SVM is a supervised machine learning algorithm developed by Vapnik (2013). This algorithm is very powerful at recognizing subtle patterns in complex datasets with balanced accuracy and reproducibility (PISNER; SCHNYER, 2020). SVM is an optimal "hyperplane" that it aims to separate (i.e., "classify") observed data according to its class based on patterns of information about those observations called features. An SVM can be linear or nonlinear, however the linear is the most used (PISNER; SCHNYER, 2020). SVM based on a statistical learning theory and concept of the structural risk minimization principle, which reduces the upper bound generalization error rather than the local training error (FENG; WEN; LI, 2015; SHIRI et al., 2014). Mehdizadeh, Behmanesh and Khalili (2017) observed a correlation coefficient above 0.97 (testing stage) in the ET_0 estimate through the SVM. However, another algorithm used in the research showed better results. This pattern was observed by other authors (KUMAR et al., 2016). Thus, evaluating other models is necessary.

The WEKA (Waikato Environment for Knowledge Analysis), developed by the University of Waikato, Hamilton, New Zealand (WITTEN; FRANK, 2002), was data mining

tool used for the study to generate the models. WEKA is an open-source tool written in Java that is widely used by the data miners (KABAKCHIEVA, 2013). Weka has several classification and regression algorithms such as J48, Bayesian Network, Random Tree and Simple Linear Regression (in addition to the algorithms used in the present study). This tool facilitates the use of the algorithms, making it possible to configure the model's boundary conditions. Besides, it allows other analyzes such as the rank of the input attributes and the comparison of different algorithms. WEKA has been used frequently in studies related to agrometeorology, climatology, hydrology, and irrigation (ERECHTCHOUKOVA; KHAITER, 2017; PATEL et al., 2014; SATTARI et al., 2021). Furthermore, these papers presented a positive evaluation of the tool.

Two articles were developed to compose the present thesis that has as general objective the study the predictive capacity of classical statistical method (MLR) and machine learning models (ANN, RF and SVM) to estimate the mean (Tmean), maximum (Tmax), and minimum (Tmin) air temperatures and the mean reference evapotranspiration (ET_0), in the Minas Gerais State, with different input meteorological data combinations.

The first manuscript, entitled "Air temperature estimation techniques in the Minas Gerais state, Brazil, Cwa and Cwb climate regions according to the Koppen- Geiger climate classification system" was accepted for publication by journal *Ciência e Agrotecnologia*. This article aimed to determine a model that is efficient in estimating the mean, maximum, and minimum monthly air temperatures in any location in the Minas Gerais state with climatic classification Cwa or Cwb (KÖPPEN; GEIGER, 1928). The main result of this study was to analyze the ability of classical statistical methods (MLR) and machine learning methods (ANN and RF) when input data is restricted to geographic coordinates, altitude, and month. In addition, to assessing the importance of each input data in the estimate.

The second manuscript, entitled "Evaluation of monthly mean reference evapotranspiration estimation techniques in the Minas Gerais state, Brazil" is being prepared with the aim of submitting it to a high impact scientific journal. This article aimed to develop effective models to estimate the monthly mean ET_0 with different input data combinations and in different scenarios. The different combinations of input data make it possible to analyze the model's performance under limited data conditions. Three different scenarios were analyzed: At state level (scenario II – SI) and at regional level. At the regional level, the Minas Gerais state was divided into two areas with climatic similarity according to the Classification

Systems proposed by Thornthwaite (1948) (scenario II - SII) and Köppen (1936) (scenario III - SIII). The scenarios aimed to group data from similar regions and thus increase the capacity of the models.

2 GENERAL CONSIDERATIONS

This thesis focused on improving techniques that provide conditions of access to estimated data with high efficiency for any location in Minas Gerais state. Thus, it is intended to offer subsidies for different areas of knowledge such as irrigation, hydrology, and climatology.

Studies of estimation of meteorological data are not able to remedy the deficit of climatic data. The filling of these deficits by estimated data would cause a change in the frequency distribution dispersion scale (THOM, 1966). Therefore, the estimation models aim to find patterns of behavior, learn from these patterns, and expand their application capacity to different locations.

The development of models requires a solid and reliable database, however, according to Thom (1966), the interruptions that occur in the series of climatological data do not make them unfeasible. These models, most of the time, need knowledge of the technology used or of palpable forms of application such as a web site or application so that its applicability is available to those who need it (farmers, technicians, researchers, government officials, etc.). However, studies like the ones presented are fundamental to understand the behavior and the relationships between the input data and estimated data. In addition, the results obtained may guide similar future studies in other regions. We emphasize that special attention should be paid to regions where data are not available and/or regions with low development and high potential

The meteorological variables and agrometeorological parameter analyzed are fundamental for several areas, as already mentioned. Although the thesis focuses on these variables and parameter, the knowledge produced can be expanded to other meteorological variables such as relative humidity and solar radiation.

Future scenarios indicate a reduction in rainfall and an increase in temperature and, consequently, an increase in evapotranspiration. Updates to these scenarios are essential for good management of natural resources and urban planning. Thus, ways of estimating climate data with excellence can be a way to improve and expand these forecasts, since these data are the basis for the forecasts.

New technologies such as machine learning methods are being highly efficient in estimating and classifying climate data. There is still a lot to develop, however tools such as

WEKA help in the implementation of these models in an easier and safer way. However, knowledge of the techniques is indispensable when using WEKA. Using these models and tools incorrectly can lead to inconsistent results and conclusions.

REFERENCES

AGAM, N. et al. Evaporative loss from irrigated interrows in a highly advective semi-arid agricultural area. **Advances in Water Resources**, v. 50, p. 20–30, 2012.

ALLEN, R. G. et al. **Crop evapotranspiration** - guidelines for computing crop water requirements**,** Rome: FAO, 1998. 297p. (FAO Irrigation and drainage paper 56).

ALLEN, R. G. et al. Evapotranspiration information reporting: I. Factors governing measurement accuracy. **Agricultural Water Management**, v. 98, n. 6, p. 899–920, 2011.

ALMOROX, J.; QUEJ, V. H.; MARTÍ, P. Global performance ranking of temperature-based approaches for evapotranspiration estimation considering Köppen climate classes. **Journal of Hydrology**, v. 528, p. 514–522, 2015.

ALVARES, C. A. et al. Modeling monthly mean air temperature for Brazil. **Theoretical and Applied Climatology**, v. 113, n. 3–4, p. 407–427, 2013.

BATISTA-SANTOS, P. et al. The impact of cold on photosynthesis in genotypes of Coffea spp.-Photosystem sensitivity, photoprotective mechanisms and gene expression. **Journal of Plant Physiology**, v. 168, n. 8, p. 792–806, 2011.

BENALI, A. et al. Estimating air surface temperature in Portugal using MODIS LST data. **Remote Sensing of Environment**, v. 124, p. 108–121, 2012.

BENALI, L. et al. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. **Renewable Energy**, v. 132, p. 871–884, 2019.

BENAVIDES, R. et al. Geostatistical modelling of air temperature in a mountainous region of Northern Spain. **Agricultural and Forest Meteorology**, v. 146, n. 3–4, p. 173–188, 2007.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

BURT, C. M. et al. Evaporation research: Review and interpretation. **Journal of irrigation and drainage engineering**, v. 131, n. 1, p. 37–58, 2005.

CANNELL, M. G. R. Physiology of the coffee crop. In: **Coffee**. [s.l.] Springer, 1985. p. 108–134.

CARVALHO, L. G. de et al. Evapotranspiração de referência: uma abordagem atual de diferentes métodos de estimativa. **Pesquisa Agropecuária Tropical**, v. 41, n. 3, p. 456–465, 2011.

DAMATTA, F. M. et al. Physiological and Agronomic Performance of the Coffee Crop in the Context of Climate Change and Global Warming: A Review. **Journal of Agricultural and Food Chemistry**, v. 66, n. 21, p. 5264–5274, 2018.

DUMEDAH, G.; COULIBALY, P. Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. **Journal of Hydrology**, v. 400, n. 1–2, p. 95–102, 2011.

ERECHTCHOUKOVA, M. G.; KHAITER, P. A. The effect of data granularity on prediction of extreme hydrological events in highly urbanized watersheds: A supervised classification approach. **Environmental modelling & software**, v. 96, p. 232–238, 2017.

EWAID, S. H.; ABED, S. A.; AL-ANSARI, N. Crop water requirements and irrigation schedules for some major crops in Southern Iraq. **Water**, v. 11, n. 4, p. 756, 2019.

FAUSETT, L. **Fundamentals of neural networks: architectures, algorithms, and applications**. [s.l.] Prentice-Hall, Inc., 1994.

FENG, Q.; WEN, X.; LI, J. Wavelet analysis-support vector machine coupled models for monthly rainfall forecasting in arid regions. **Water resources management**, v. 29, n. 4, p. 1049–1065, 2015.

FERREIRA, L. B. et al. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM–A new approach. **Journal of hydrology**, v. 572, p. 556–570, 2019.

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. **Agricultural Water Management**, v. 234, p. 106113, 2020.

FUNDAÇÃO JOÃO PINHEIRO. Contas regionais de Minas Gerais: ano de referência 2018. 2020. Available in: < http://novosite.fjp.mg.gov.br/wp-content/uploads/2020/10/11.01_Serie-Estatistica-Informacoes-V.-35-FINAL-110120.pdf> Access in: April, 27, 2021

HATFIELD, J. L.; PRUEGER, J. H. Temperature extremes: Effect on plant growth and development. **Weather and climate extremes**, v. 10, p. 4–10, 2015.

HUANG, G. et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. **Journal of Hydrology**, v. 574, p. 1029–1041, 2019.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Panorama.** 2020a. Available in:< https://cidades.ibge.gov.br/brasil/mg > Access in: April, 28, 2021.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Produto Interno Bruto dos municípios**. 2018. 2020b Available in:< https://www.ibge.gov.br/explica/pib.php > Access in: April, 28, 202.

ISMAIL, S. M.; EL-NAKHLAWY, F. S. Measuring crop water requirement and crop

coefficient for blue panic crop under arid conditions using draining lysimeters. **Irrigation and Drainage**, v. 67, n. 3, p. 454–460, 2018.

JANATIAN, N. et al. A statistical framework for estimating air temperature using MODIS land surface temperature data. **International Journal of Climatology**, v. 37, n. 3, p. 1181–1194, 2017.

KABAKCHIEVA, D. Predicting student performance by using data mining methods for classification. **Cybernetics and information technologies**, v. 13, n. 1, p. 61–72, 2013.

KOPPEN, W. das. Das geographische system der klimat. **Handbuch der klimatologie**, p. 46, 1936.

KÖPPEN, W.; GEIGER, R. Klimate der Erde. Gotha: Verlag Justus Perthes. **Wall-map 150cmx200cm**, 1928.

KUMAR, D. et al. Estimating evapotranspiration using an extreme learning machine model: case study in north Bihar, India. **Journal of Irrigation and Drainage Engineering**, v. 142, n. 9, p. 4016032, 2016.

KUMAR, M. et al. Estimating evapotranspiration using artificial neural network. **Journal of Irrigation and Drainage Engineering**, v. 128, n. 4, p. 224–233, 2002.

KUMAR, M.; RAGHUWANSHI, N. S.; SINGH, R. Artificial neural networks approach in evapotranspiration modeling: A review. **Irrigation Science**, v. 29, n. 1, p. 11–25, 2011.

KUSRIYANTO, M.; PUTRA, A. A. Weather Station Design Using IoT Platform Based On Arduino Mega. In: 2018 International Symposium on Electronics and Smart Devices (ISESD), **Anais**...IEEE, 2018.

LEUNG, H.; FAN, Z. Software cost estimation. In: **Handbook of Software Engineering and Knowledge Engineering: Volume II: Emerging Technologies**. [s.l.] World Scientific, 2002. p. 307–324.

MALIK, A. et al. The viability of co-active fuzzy inference system model for monthly reference evapotranspiration estimation: case study of Uttarakhand State. **Hydrology Research**, v. 50, n. 6, p. 1623–1644, 2019.

MARTÍ, P.; GONZÁLEZ-ALTOZANO, P.; GASQUE, M. Reference evapotranspiration estimation without local climatic data. **Irrigation Science**, v. 29, n. 6, p. 479–495, 2011.

MEHDIZADEH, S.; BEHMANESH, J.; KHALILI, K. Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. **Computers and electronics in agriculture**, v. 139, p. 103–114, 2017.

MOJID, M. A.; RANNU, R. P.; KARIM, N. N. Climate change impacts on reference crop evapotranspiration in North-West hydrological region of Bangladesh. **International Journal of Climatology**, v. 35, n. 13, p. 4041–4046, 2015.

MONTEITH, J. L. Evaporation and environment. In: Symposia of the society for experimental biology, **Anais**...Cambridge University Press (CUP) Cambridge, 1965.

MOREIRA, M. C.; CECÍLIO, R. A. SOFTWARE TO ESTIMATE AIR TEMPERATURE IN THE BRAZILIAN NORTHEASTERN REGION USING ARTIFICIAL NEURAL NETWORKS. **REVISTA ENGENHARIA NA AGRICULTURA-REVENG**, v. 24, n. 2, p. 164–171, 2016.

MWALE, F. D.; ADELOYE, A. J.; RUSTUM, R. Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi - A self organizing map approach. **Physics and Chemistry of the Earth**, v. 50–52, p. 34–43, 2012. Available in: <http://dx.doi.org/10.1016/j.pce.2012.09.006>.

NOI, P. T.; DEGENER, J.; KAPPAS, M. Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. **Remote Sensing**, v. 9, n. 5, 2017.

PARTELLI, F. L. et al. Low temperature impact on photosynthetic parameters of coffee genotypes. **Pesquisa Agropecuária Brasileira**, v. 44, n. 11, p. 1404–1415, 2009.

PATEL, P. K. et al. Flooding: abiotic constraint limiting vegetable productivity. **Advances in Plants and Agriculture Research**, v. 1, n. 3, p. 96–103, 2014.

PATIL, A. P.; DEKA, P. C. An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. **Computers and Electronics in Agriculture**, v. 121, p. 385–392, 2016.

PENMAN, H. L. Natural evaporation from open water, bare soil and grass. **Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences**, v. 193, n. 1032, p. 120–145, 1948.

PHAM, M. T. et al. A coupled stochastic rainfall–evapotranspiration model for hydrological impact analysis. **Hydrology and Earth System Sciences**, v. 22, n. 2, p. 1263–1283, 2018.

PISNER, D. A.; SCHNYER, D. M. Support vector machine. In: **Machine Learning**. [s.l.] Elsevier, 2020. p. 101–121.

SALAM, R. et al. The optimal alternative for quantifying reference evapotranspiration in climatic sub-regions of Bangladesh. **Scientific reports**, v. 10, n. 1, p. 1–21, 2020.

SATTARI, M. T. et al. Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. **Hydrology and Earth System Sciences**, v. 25, n. 2, p. 603–618, 2021.

SCHLENKER, W.; ROBERTS, M. J. Nonlinear temperature effects indicate severe damages to US crop yields under climate change. **Proceedings of the National Academy of sciences**, v. 106, n. 37, p. 15594–15598, 2009.

SHIRI, J. et al. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. **Computers and Electronics in Agriculture**, v. 108, p. 230–241, 2014.

SOARES, E. et al. Ensemble of evolving data clouds and fuzzy models for weather time series prediction. **Applied Soft Computing**, v. 64, p. 445–453, 2018.

TALEGHANI, M. Outdoor thermal comfort by different heat mitigation strategies- A review. **Renewable and Sustainable Energy Reviews**, v. 81, n. March, p. 2011–2018, 2018. Available in: <http://dx.doi.org/10.1016/j.rser.2017.06.010>.

THOM, H. C. S. **Some methods of climatological analysis**. WMO, Techi ed. [s.l: s.n.]. 1966.

THORNTHWAITE, C. W. An approach toward a rational classification of climate. **Geographical review**, v. 38, n. 1, p. 55–94, 1948.

VAPNIK, V. **The nature of statistical learning theory**. [s.l.] Springer science & business media, p. 188, 2013.

WAHID, A. et al. Heat tolerance in plants: An overview. **Environmental and Experimental Botany**, v. 61, n. 3, p. 199–223, 2007.

WANG, S. et al. Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China. **Agricultural Water Management**, v. 221, p. 220–230, 2019.

WANG, S.; HUANG, G. H.; HE, L. Development of a clusterwise-linear-regression-based forecasting system for characterizing DNAPL dissolution behaviors in porous media. **Science of the total environment**, v. 433, p. 141–150, 2012.

WILCOX, B. P. et al. The water balance on rangelands. **Encyclopedia of water science**, v. 791, p. 4, 2003.

WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with Java implementations. **Acm Sigmod Record**, v. 31, n. 1, p. 76–77, 2002.

XIANG, K. et al. Similarity and difference of potential evapotranspiration and reference crop evapotranspiration–a review. **Agricultural Water Management**, v. 232, p. 106043, 2020.

XU, Y.; KNUDBY, A.; HO, H. C. Estimating daily maximum air temperature from MODIS in British Columbia, Canada. **International Journal of Remote Sensing**, v. 35, n. 24, p. 8108–8121, 2014.

YANG, Q. et al. Sensitivity of potential evapotranspiration estimation to the Thornthwaite and Penman–Monteith methods in the study of global drylands. **Advances in Atmospheric Sciences**, v. 34, n. 12, p. 1381–1394, 2017.

YIN, Z. et al. Integrating genetic algorithm and support vector machine for modeling daily reference evapotranspiration in a semi-arid mountain area. **Hydrology Research**, v. 48, n. 5, p. 1177–1191, 2017.

2nd CHAPTER - ARTICLES

ARTICLE 1 - AIR TEMPERATURE ESTIMATION TECHNIQUES IN THE MINAS GERAIS STATE, BRAZIL, Cwa AND Cwb CLIMATE REGIONS ACCORDING TO THE KOPPEN- GEIGER CLIMATE CLASSIFICATION SYSTEM

Article elaborated according to standards of the journal *Ciência e Agrotecnologia,* ISSN: 1981-1829 (Ciênc. agrotec. 45, 2021; https://doi.org/10.1590/1413-7054202145023920)

Air temperature estimation techniques in the Minas Gerais state, Brazil, Cwa and Cwb climate regions according to the Koppen- Geiger climate classification system

Estimativa da temperatura do ar no estado de Minas Gerais, Brasil, em regiões de clima Cwa e Cwb segundo sistema de classificação climática de Koppen-Geiger.

Pietros André Balbino dos Santos[1]

Cassio Augusto Ussi Monti[2]

Luiz Gonsaga de Carvalho[1]

Wilian Soares Lacerda[3]

Felipe Schwerz[1]

[1]Universidade Federal de Lavras/UFLA, Departamento de Engenharia Agrícola, Lavras, MG, Brasil

[2]North Carolina State University, Department of Forestry and Environmental Resources, Jordan Hall, 2800 Faucette Dr 3120, 27607, Raleigh, NC, U.S.A.

[3]Universidade Federal de Lavras/UFLA, Departamento de Automática, Lavras, MG, Brasil

Abstract: Air temperature significantly affects the processes involving agricultural and human activities. The knowledge of the temperature of a given location is essential for agricultural planning. It also helps to make decisions regarding human activities. However, it is not always possible to determine this variable. It is necessary to make a precise estimate, using methods that are capable of detecting the existing variations. The aim of this study was to develop models of multiple linear regression (MLR), artificial neural network (ANN), and random forest (RF) to estimate the mean (Tmean), maximum (Tmax), and minimum (Tmin) monthly air temperatures as a function of geographic coordinates and altitude for different

localities in Minas Gerais state, Brazil, with climatic classification Cwa or Cwb. The average monthly data (Tmean, Tmax, and Tmin), over a period of 30 years, were collected from 20 climatological stations. The MLR was able to estimate the Tmax with accuracy. However, the predictive capacity of estimating Tmean and Tmin was low. The algorithms RF and ANN were used to estimate Tmean, Tmax, and Tmin with high accuracy. The best results were obtained using the RF model.

Keywords: Artificial Neural Network, Random Forest, Multiple Linear Regression, Geographic Coordinates.

Resumo: A temperatura do ar afeta significativamente os processos que envolvem atividades agrícolas e humanas. O conhecimento da temperatura de um determinado local é fundamental para o planejamento agrícola. Também ajuda a tomar decisões sobre as atividades humanas. No entanto, nem sempre é possível determinar essa variável. É necessário fazer uma estimativa precisa, utilizando métodos que sejam capazes de detectar as variações existentes. O objetivo deste estudo foi desenvolver modelos de regressão linear múltipla (RLM), rede neural artificial (RNA) e floresta aleatória (FA) para estimar a temperatura média (Tmean), máximo (Tmax), e mínimo (Tmin) mensal do ar em função de coordenadas geográficas e altitude para diferentes áreas do Estado de Minas Gerais, Brasil, com classificação climática Cwa ou Cwb. Os dados médios mensais (Tmean, Tmax e Tmin), ao longo de um período de 30 anos, foram coletados em 20 estações climatológicas. O RLM foi capaz de estimar o Tmax com precisão. Porém, a capacidade preditiva de estimar Tmean e Tmin foi baixa. Os algoritmos FA e RNA foram usados para estimar Tmean, Tmax e Tmin com alta precisão. Os melhores resultados foram obtidos com o modelo RF.

Palavras-chave: Rede Neural Artificial, Floresta Aleatória, Regressão Linear Múltipla, Coordenadas Geográficas.

INTRODUCTION

It is important to monitor the meteorological elements to achieve proper growth and yield of crops. Efficient monitoring can help in evapotranspiration estimates, irrigation planning, pest and disease risk zoning, animal comfort index mapping, etc. One of the most important meteorological elements is air temperature, which influences plant physiology. Changes in air temperature can lead to change in the growth and development of plants (Benlloch-González et al., 2016; Cardoso et al., 2012; Wahid et al., 2007). The air temperature influences various physiological processes occurring in a plant, such as the speed of chemical reactions (Benavides et al., 2007) that occur in the temperature range of 0 – 40 °C. The extent of influence exerted depends on the plant species. When the air temperature exceeds the ideal range for each species, morphological, physiological, and biochemical changes may be induced, leading to adverse effects on plant growth (Wahid et al., 2007). Studies on the characterization of air temperature, precipitation, and the climatic classification of the regions where agriculture predominates should be conducted to improve crop yields (Cardoso et al., 2015; Costa et al., 2012).

In coffee crop science, one of the main crop types grown in the Minas Gerais State, Brazil (Compahia Brasileira de Abastecimento - CONAB, 2020), the optimum mean annual temperature falls in the range of 18 – 23 °C for the proper growth of *C. Arabica* specie. The optimum temperature falls in the range of 22 – 26 °C for the proper growth of *C. Canephora* (Damatta et al., 2018). Temperatures that fall outside this range influence the growth and yields of the crops. When the temperature is extremely low, the activity of the coffee crop reduces, and the photosynthetic performance is noticeably affected. The net photosynthetic activity ceases almost completely (Batista-Santos et al., 2011; Partelli et al., 2009). On the other hand, very high temperatures may cause a decrease in the net photosynthetic rates of the

leaves (Cannell, 1985). The ideal temperature interval produces a high crop yield over the years. The temperature outside the optimal range results in reduced crop yield. Therefore, it is important to determine the mean air temperature and the extreme temperatures (maximum and minimum). Furthermore, considering the characteristics of the relief and location of the Minas Gerais State, the accurate estimation of extreme temperatures is important because the state exhibits topographic conditions that allow the formation of frosts on an annual basis in the southern region. The maximum temperatures (40–42 °C) are recorded in the northern regions of the state.

The mean, maximum, and minimum air temperatures can be monitored on a daily basis in weather stations. However, in the Minas Gerais region, the coverage of the official network of surface weather stations is limited. Besides, interruptions and errors in the database generated by these stations are quite common. The errors can be attributed to reading errors, damaged devices, and other unintended observational problems (Dumedah; Coulibaly, 2011; Mwale; Adeloye; Rustum, 2012). These factors limit climatic studies, e.g., studies on the climatic characterization of the region and studies on meteorological elements that slow down the development of agriculture.

Considering the fact that the average monthly air temperature varies with geographic coordinates and altitude, several researchers working in different regions of Brazil have been trying to develop techniques and models for estimating the air temperature. The multiple linear regression (MLR) model considers the latitude, longitude, and altitude of the location as independent variables (Alvares et al., 2013; Cargnelutti Filho; Maluf; Matzenauer, 2008; Pezzopane et al., 2004; Sediyama; Melo Júnior, 1998). These estimates have been made with different levels of precision and accuracy. However, the development of new tools such as the

Artificial Neural Network and Random Forests technique can maximize the performance, precision, and accuracy of estimating the air temperature.

The new techniques have been developed with the aim of achieving higher accuracy during the estimation of variables. The Artificial Neural Network (ANN) is a promising and effective tool for non-linear modeling and complex time-series. It has been used in different fields of science such as medicine (Muhammad et al., 2019), hydrology (Asadi et al., 2019), and agriculture (De Oliveira Aparecido et al., 2020). The ANN model is a mathematical model in which the architecture is analogous to brain functioning. The interconnecting processing elements are arranged in several layers (Kumar; Raghuwanshi; Singh, 2011). The ANN method helps understand and generalize the relationships between complex datasets. This expands the scope of the application of the method (Wu; Dandy; Maier, 2014).

ANNs have been used for the estimation of meteorological variables with good accuracy. Estimation of reference evapotranspiration (Antonopoulos; Antonopoulos, 2017; Kumar; Raghuwanshi; Singh, 2011), solar radiation (Bou-Rabee et al., 2017), and air temperature (Moreira; Cecílio, 2016) have been carried out using this technique. It is important to conduct this study to verify the applicability of the ANN method for estimating the mean, maximum, and minimum air temperature. The efficiency of the technique has been investigated. Reports on the use of ANNs (used to estimate the temperature in the region under study) are scarce.

The Random Forest (RF) is non-parametric statistical data modeling methods (Breiman, 2001). The models have been used to analyze data in different fields of science, such as medicine (Xie et al., 2020), biology (Fabris et al., 2018), and geoprocessing (Vogels et al., 2017). According to James et al. (2013), decision trees detect non-linear relationships in the evaluated system when the use of linear relationships, e.g., linear regression analysis, is

restricted. According to Seyedhosseini and Tasdizen (2015), RF is a classification and regression technique used to grow ensemble decision trees such that the correlation between the trees remains as low as possible. This condition can be achieved by the method of bootstrap sampling. In this method, resamples are replaced by simulating a single random sample. It must represent samples taken from the original population. Data from previously conducted analytical experiments are required to enhance the predictive and generalization abilities (Hesterberg et al., 2002).

RF has also been adopted to predict meteorological variables such as solar radiation (Benali et al., 2019) and air temperature (Noi; Degener; Kappas, 2017). RF has been found to be a more efficient predicting tool compared to other tools like ANN (Benali et al., 2019; Zhou et al., 2016). The RF is still little applied, and the interest in this predictive tool is increasing as it exhibits a good practical performance (Scornet, 2016). Therefore, it is important to evaluate the RF potential for estimating air temperature and to compare it with different methods.

The objective of this study was to develop and compare the performances of multiple linear regression (MLR), Artificial Neural Networks (ANN), and Random Forests (RF) models for estimating the mean, maximum, and minimum monthly air temperatures using input variables such as geographical coordinates and altitude for different areas in the Minas Gerais State with climatic classification Cwa or Cwb (Köppen; Geiger, 1928).

MATERIAL AND METHODS

STUDY AREA AND DATA SOURCES

The present study was developed for municipalities in the Minas Gerais state that are within the regions classified as Cwa (humid temperate climate with dry winter and hot

summer) and Cwb (humid temperate climate with dry winter and moderately hot summer). This classification was proposed by Köppen and Geiger (1928) (Figure 1). This Climatic Classification Systems (CMS) was developed by Köppen in 1918, and its most popular version was published in 1928 in collaboration with Rudolf Oskar Robert Williams Geiger. The Köppen and Geiger (1928) CMS a simple and comprehensive system, and hence it is widely used. The mean annual rainfall recorded in the region under study is 1379 mm (Brasil, 1992). The study was limited to the areas classified as Cwa and Cwb. The aim was to determine the maximum efficiency of the models tested. Highly accurate data were obtained when the models were used in regions exhibiting similar climatic characteristics.

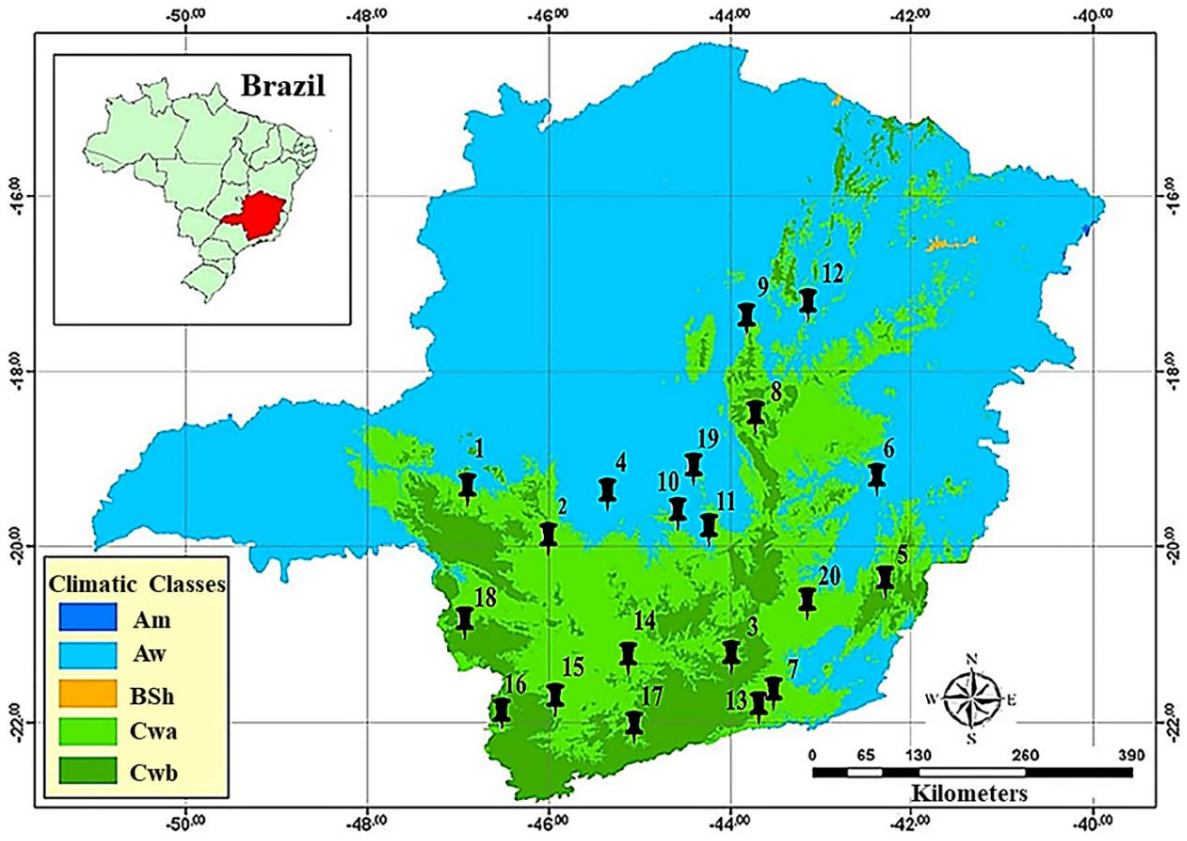


Figure 1 - Climate zoning in the state of Minas Gerais. Zoned according to the Köppen and Geiger (1928) climatic classification. Codes of the climatological stations of the National Institute of Meteorology. Source: Adapted from De Sá Júnior et al. (2012).

According to De Sá Júnior et al. (2012), the regions classified as Cwa and Cwb represent 21% and 11% of the area of the Minas Gerais state, respectively. There are 20 climatological stations located in the region under study. The regions fall under the realm of the national network of climatological stations (National Institute of Meteorology (INMET)). The respective geographical coordinates and climatic classification have been presented in Table 1. The average monthly data (mean (Tmean), maximum (Tmax), and minimum (Tmin) air temperature) over a period of 30 years, from 1987 to 2017, of each conventional station were used for the studies. The data were extracted from the Meteorological Database for Teaching and Research - BDMEP of INMET. Although some locations do not have a record of 30 years of data (Table 1), all stations presented more than 90% of the consistent data.

Table 1 - Principal climatological station of the INMET used to estimate the mean, maximum, and minimum air temperature.

ID	Climatological stations location	Latitude (S°)	Longitude (W°)	Altitude (m)	Climatic classes	Period (years)
1	Araxá	19.56	46.93	1004	Cwa	1887 - 2017
2	Bambuí	20.00	45.98	661	Cwa	1887 - 2017
3	Barbacena	21.25	43.76	1126	Cwb	1890 - 2017
4	Bom Despacho	19.68	45.36	695	Cwa	1887 - 2017
5	Caparaó	20.51	41.86	843	Cwb	1890 - 2017
6	Caratinga	19.80	42.15	609	Cwa	1887 - 2017
7	Cal. Pacheco	21.58	43.25	453	Cwa	1887 - 2009
8	C. Mato Dentro	19.03	43.43	652	Cwa	1887 - 2017
9	Diamantina	18.25	43.60	1296	Cwb	1887 - 2017
10	Florestal	19.88	44.41	760	Cwa	1887 - 2017
11	Ibirité	20.01	44.05	815	Cwa	1887 - 2015
12	Itamarandiba	17.85	42.85	1097	Cwb	1887 - 2017
13	Juiz de Fora	21.76	43.35	940	Cwa	1887 - 2017
14	Lavras	21.23	45.00	919	Cwa	1888 - 2017
15	Machado	21.66	45.91	874	Cwa	1891 - 2017
16	Poços de Caldas	21.91	46.38	1150	Cwb	1892 - 2015
17	São Lourenço	22.10	45.01	900	Cwa	1887 - 2017
18	S. Seb. do Paraíso	20.91	47.11	820	Cwb	1887 - 2013
19	Sete Lagoas	19.46	44.25	732	Cwa	1892 - 2015
20	Viçosa	20.75	42.85	690	Cwa	1890 - 2017

Source: Adapted from De Sá Júnior et al. (2012).

MULTIPLE LINEAR REGRESSION (MLR) METHOD

Based on the independent variables (geographic coordinates and altitude), MLR was developed to estimate the mean, maximum, and minimum average temperature of each month of the year for each location. The average temperatures were calculated as follows (Equation 1):

$$Y_i = \beta_0 + \beta_1 ALT + \beta_2 LAT + \beta_3 LON. \quad (1)$$

where Yi is Tmean, Tmax, or Tmin in °C and is the dependent variable. ALT represents the altitude in m, LAT represents the latitude in degrees, and LON represents the longitude in degrees, which are independent variables. β0, β1, β2, and β3, are the regression coefficients. MLR was implemented using the data analysis tool in Microsoft Excel®. Contrary to the methodology applied for ANN and RF, the month was not used as an input variable. Therefore, the data for Tmean, Tmax, and Tmin were classified based on the month. Subsequently, the MLRs were adjusted. Each month had a characteristic equation generating a specific statistical result. The methodology reported by Sediyama and Melo Júnior (1998) were used for the studies. This methodology increases the predictive capacity of MLR and facilitates the analysis of each independent variable in the month. The influence of each variable on the result can also be analyzed.

ARTIFICIAL NEURAL NETWORKS (ANNS) MODEL DEVELOPMENT

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 algorithm used for ANN was the Multilayer Perceptron (MLP) algorithm (Fausett, 1994). The architecture consisted of the input layer, hidden layers (where the data are processed), and output layer (where the results of processing are compiled) (Figure 2).

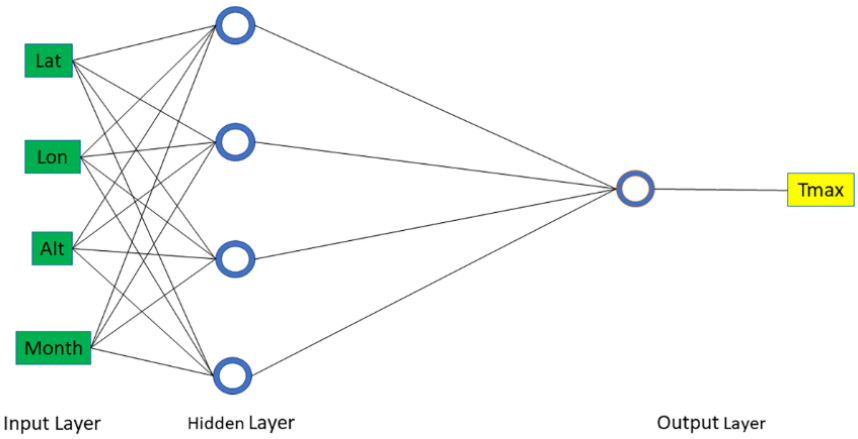


Figure 2 - Network structure scheme consisting of five neurons in the hidden layer built by WEKA (ANN2) to estimate Tmax (Source: The Authors).

The input data consisted of the month, latitude, longitude, and altitude of each evaluated location. Each ANN setting estimated the Tmean, Tmax, or Tmin for all the months. There are good reasons behind using these variables for these studies. The temporal variable consists of the cumulative month component, which is required to execute the projections. The latitude and longitude are the variables related to the position. The temperature changes with the position as the position changes from the Poles to the Equator Line. The temperature gradually increases from the poles to the equator. The altitude variable is regarded as the surface component. It can be stated that the higher the altitude, the lower the temperature. The ANN follows a mathematical structure connecting the processing nodes (neurons). The output of a neuron is the input of the subsequently combined neurons. The final model is built based on various assumptions on activation function (Equations 2 – 8). The equations are as follows:

$$\overrightarrow{N_1} = L\,\mathrm{n}(1 + e^{\left(w_{1,1}\cdot LAT + w_{2,1}\cdot LON + w_{3,1}\cdot ALT + w_{4,1}\cdot MONTH + w_{5,1}\cdot B\right)}, \qquad (2)$$

$$\overrightarrow{N_2} = L\,\mathrm{n}(1 + e^{\left(w_{1,2}\cdot LAT + w_{2,2}\cdot LON + w_{3,2}\cdot ALT + w_{4,2}\cdot MONTH + w_{5,2}\cdot B\right)}, \quad (3)$$

$$\overrightarrow{N_3} = L\,\mathrm{n}(1 + e^{\left(w_{1,3}\cdot LAT + w_{2,3}\cdot LON + w_{3,3}\cdot ALT + w_{4,3}\cdot MONTH + w_{5,3}\cdot B\right)}, \quad (4)$$

$$\overrightarrow{N_4} = L\,\mathrm{n}(1 + e^{\left(w_{1,4}\cdot LAT + w_{2,4}\cdot LON + w_{3,4}\cdot ALT + w_{4,4}\cdot MONTH + w_{5,4}\cdot B\right)}, \quad (5)$$

$$\overrightarrow{T_{\max}} = L\,\mathrm{n}(1 + e^{\left(w_{1,5}\cdot N_1 + w_{2,5}\cdot N_2 + w_{3,5}\cdot N_3 + w_{4,5}\cdot N_4 + w_{5,5}\cdot B\right)}, \quad (6)$$

$$\overrightarrow{T_{\min}} = L\,\mathrm{n}(1 + e^{\left(w_{1,5}\cdot N_1 + w_{2,5}\cdot N_2 + w_{3,5}\cdot N_3 + w_{4,5}\cdot N_4 + w_{5,5}\cdot B\right)}, \quad (7)$$

$$\overrightarrow{T_{mean}} = L\,\mathrm{n}(1 + e^{\left(w_{1,5}\cdot N_1 + w_{2,5}\cdot N_2 + w_{3,5}\cdot N_3 + w_{4,5}\cdot N_4 + w_{5,5}\cdot B\right)}. \quad (8)$$

Equations 2–5 represent the mathematical abstraction of the ANN built in Figure 2 extracting the neurons equations. Equations 6–8 are the estimate vectors of each output. $W_{i,j}$ represents the weights estimated using the backpropagation algorithm during ANN processing. The value of $B_{i,j}$ represents the bias associated with each measurer. The activation function applied was sigmoidal with non-linear output.

All adjustments were cross-assessed. Twenty folds of the sample set were used for the assessment for training to compensate for the reduced number of instances. Two different configurations were evaluated (Table 2). Results from the preliminary tests indicated that changes in the number of training epochs and the number of neurons present in the hidden layer interfered with the performance of the models. However, changes in the other parameters did not significantly influence the model performances.

Table 2. WEKA configuration in the ANN implementation.

	Tmean		Tmax		Tmin	
	ANN1	ANN2	ANN1	ANN2	ANN1	ANN2
Learning rate	0.3	0.3	0.3	0.3	0.3	0.3
Momentum	0.2	0.2	0.2	0.2	0.2	0.2
Number of training epochs	500	1000	500	500	500	1000
Number of hidden layers	1	1	1	1	1	1
Number of neurons into the hidden layer	5	6	6	5	6	6

DEVELOPMENT OF THE RANDOM FOREST (RF) MODEL

The implementation of RF in WEKA has its basis on a previously reported study (Breiman, 2001). Two configurations of RF were used, with the input variables being month, latitude, longitude, and altitude of each evaluated location. Thus, each RF setting could be used to estimate Tmean, Tmax, or Tmin for all the months under study. The steps followed has been presented in Figure 3.

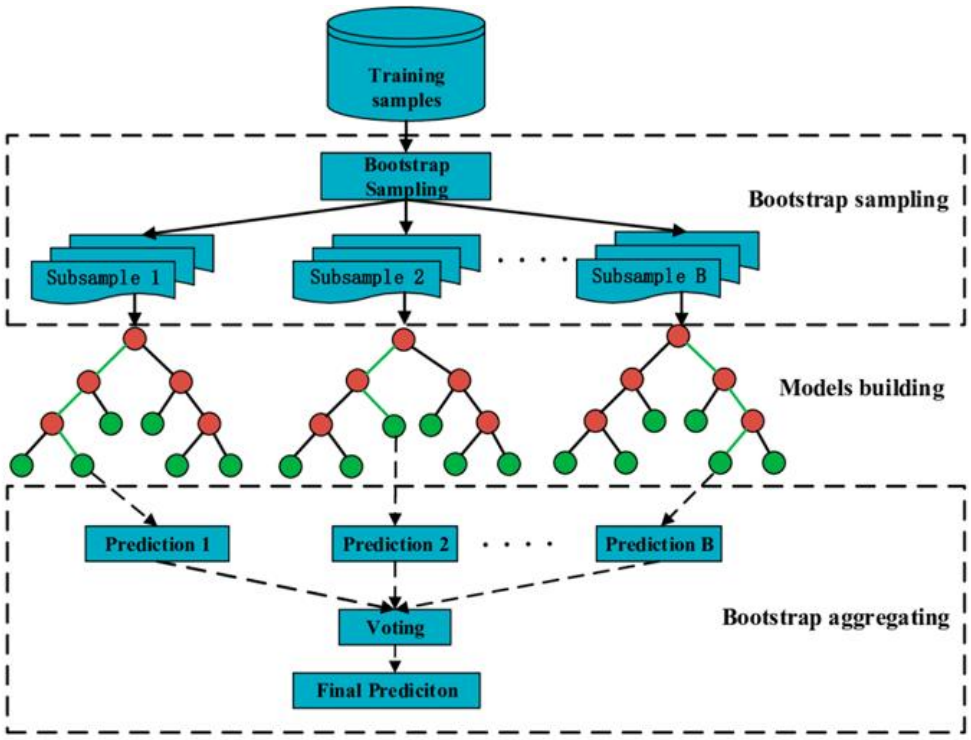


Figure 3–Schematic representation of the steps used in the RF model following the resampling strategy (Source: Wang et al., 2019).

In this study, preliminary examinations were conducted for several configurations. The configurations with 100 and 500 interactions exhibited better performance compared to other values obtained in the preliminary analysis. The preliminary tests revealed that the changes in the other parameters did not positively influence the model performance. The tests exhibited two distinct configurations for better results (Table 3).

Table 3: WEKA configuration in the RF implementation process.

	Tmean		Tmax		Tmin	
	RF1	RF2	RF1	RF2	RF1	RF2
Break ties randomly when several attributes look equally good		x		x	x	x
Size of each bag, as percentage of the training set size	100	100	100	100	100	100
Number of iterations	500	500	500	500	100	500
Minimum number of instances per leaf	1	1	1	1	1	1
Maximum depth of the tree	unl.	unl.	unl.	unl.	unl.	unl.

STATISTICAL TESTS

Various statistical indices were used to assess the predictive quality of each technique in terms of variation, precision, accuracy, and performance. The mean absolute error (MAE) and root mean square error (RMSE) indicates revealed how close the predicted values were to the observed value. Thus, the accuracy of each model could be predicted. The variation was quantified by the determination coefficient (R^2), which represents the percentage of the variation of the dependent variable explained by the independent variable. The best model should produce an R^2 value close to unity. The precision of the models was quantified based on Pearson's correlation coefficient (r), which indicates the degree of dispersion of the data obtained in terms of the mean. Accuracy was quantified using Willmott's index of agreement (d) and the performance index (c) (Camargo; Sentelhas, 1997). The performance index was calculated using the equation c = r. d. This equation was also used to quantify the performance of the model. The performances were classified as: Excellent (1 – 0.85), Very

good (0.85 – 0.76), Good (0.76 – 0.66), Average (0.66 – 0.61), Poor (0.61 – 0.51), Bad (0.51 – 0.41), and Terrible (less than 0.41).

Weka provides a tool to compare different combinations and different algorithms called WEKA Experiment Environment (Figure 4). This tool was used to compare the performance of each algorithm and configuration used in the present study conducted using the cross-validation technique. According to Noi, Degener, and Kappas, (2017), cross-validation is one of the most popular validation methods used to compare different combinations and different algorithms. In the cross-validation method, the dataset is divided into k groups (k-fold) of approximately the same size. Due to the number of observations, a 20-fold cross-validation method was used. The algorithms were applied for each fold, generating statistical performance values. Later, these average performance values were compared by Tukey's test at 5% probability. The statistical software Sisvar (Ferreira, 2019) was used for analysis. The MLR method was not implemented in WEKA. The approach was different from that was used in the ANN and RF methods. Hence, it was not possible to compare the MLR method with the other techniques using Tukey's test. The comparison between MLR and other techniques was made by comparing the statistical performance indicators.

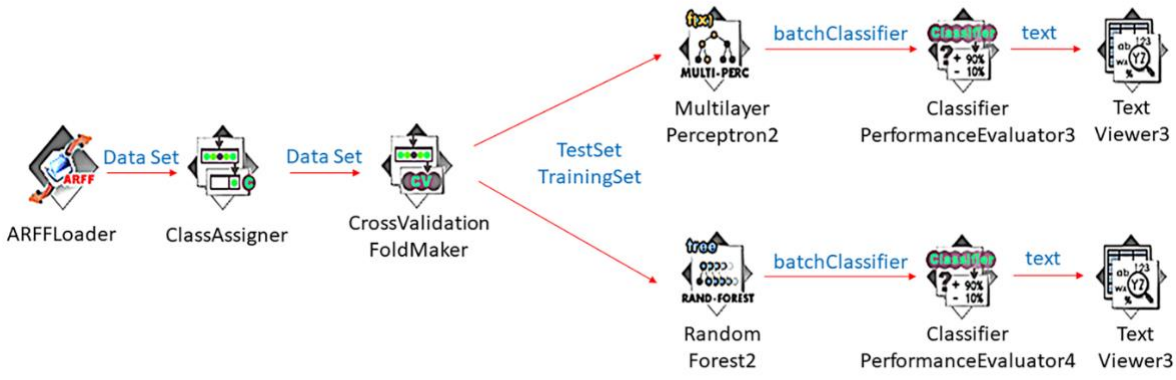


Figure 4–WEKA Experiment Environment workflow of the experiment (Source: The Authors).

RESULTS AND DISCUSSION

The MLR method coefficients were adjusted to estimate the Tmean, Tmax, and Tmin monthly air temperatures. The respective mean absolute errors (MAE), root mean square errors (RMSE), determination coefficient (R^2), Pearson's correlation coefficient (r), Willmott's index of agreement (d), and the consistency index (c) are shown in Table 4.

Table 4. Coefficients of the monthly air temperature models and statistical performance indicators.

	MLR method coefficients				MAE	RMSE	R^2	r	d	c
Month	(β_0)	Alt (β_1)	Lat (β_2)	Lon (β_3)	(°C)	(°C)				
				Tmean						
Jan	26.18	0.2980*	-0.144*	0.004*	0.25	0.33	0.88	0.94	0.97	0.90
Feb	26.41	-0.0030*	-0.346*	0.143^{ns}	0.33	0.41	0.75	0.86	0.92	0.80
Mar	27.05	-0.0040*	-0.311*	0.114^{ns}	0.28	0.34	0.85	0.92	0.96	0.89
Apr	22.87	-0.0046*	-0.363*	0.214*	0.38	0.45	0.82	0.90	0.95	0.86
May	28.84	-0.0026*	-0.760*	0.162^{ns}	0.47	0.56	0.75	0.87	0.92	0.80
Jun	33.19	-0.0011^{ns}	-0.917*	0.080^{ns}	0.62	0.72	0.65	0.80	0.88	0.71
Jul	25.62	-0.0011^{ns}	-0.718*	0.156^{ns}	0.86	0.99	0.38	0.62	0.73	0.45
Aug	21.73	-0.0014^{ns}	-0.779*	0.307^{ns}	0.75	0.87	0.51	0.71	0.81	0.58
Sep	17.94	-0.0028*	-0.828*	0.491*	0.63	0.72	0.71	0.84	0.91	0.77
Oct	18.72	-0.0046*	-0.613*	0.442*	0.44	0.49	0.85	0.92	0.96	0.88
Nov	18.29	-0.0061*	-0.409*	0.387*	0.27	0.32	0.93	0.97	0.98	0.95
Dec	21.63	-0.0053*	-0.173*	0.205*	0.23	0.30	0.91	0.95	0.98	0.93
				Tmax						
Jan	35.07	-0.0051*	-0.468*	0.179^{ns}	0.52	0.61	0.76	0.87	0.93	0.81
Feb	36.87	-0.0037*	-0.483*	0.141^{ns}	0.60	0.70	0.63	0.80	0.88	0.70
Mar	31.98	-0.0058*	-0.332*	0.199^{ns}	0.46	0.51	0.83	0.91	0.95	0.87
Apr	25.54	-0.0061*	-0.511*	0.407*	0.50	0.58	0.83	0.91	0.95	0.87
May	25.88	-0.0055*	-0.885*	0.509*	0.49	0.58	0.86	0.93	0.96	0.89
Jun	23.46	-0.0068*	-0.753*	0.514*	0.62	0.69	0.83	0.91	0.95	0.87
Jul	20.14	-0.0059*	-0.747*	0.577*	0.55	0.66	0.82	0.91	0.95	0.86
Aug	13.30	-0.0067*	-0.704*	0.762*	0.53	0.64	0.86	0.93	0.96	0.89
Sep	15.71	-0.0054*	-0.992*	0.855*	0.64	0.74	0.84	0.91	0.95	0.87
Oct	15.63	-0.0069*	-0.690*	0.753*	0.61	0.69	0.84	0.92	0.95	0.88
Nov	17.02	-0.0071*	-0.480*	0.609*	0.52	0.60	0.85	0.92	0.96	0.88
Dec	22.13	-0.0071*	-0.101*	0.336*	0.45	0.51	0.84	0.92	0.96	0.88
				Tmin						
Jan	19.08	-0.0038*	-0.071^{ns}	0.090^{ns}	0.35	0.42	0.71	0.84	0.91	0.77

Fev	22.79	-0.0025*	-0.188^{ns}	0.031^{ns}	0.49	0.56	0.47	0.69	0.80	0.55
Mar	25.67	-0.0023^{ns}	-0.241^{ns}	0.019^{ns}	0.51	0.66	0.43	0.66	0.78	0.51
Abr	30.60	-0.0032^{ns}	-0.222^{ns}	0.164^{ns}	0.77	0.97	0.41	0.64	0.78	0.50
Maio	39.19	-0.0005^{ns}	-0.690^{ns}	0.262^{ns}	0.92	1.16	0.37	0.60	0.73	0.44
Jun	40.56	0.0004^{ns}	-0.653^{ns}	0.362^{ns}	1.17	1.49	0.25	0.50	0.62	0.31
Jul	33.88	0.0015^{ns}	-0.577^{ns}	0.287^{ns}	1.52	1.92	0.10	0.32	0.40	0.13
Ago	30.86	0.0001^{ns}	-0.578^{ns}	0.173^{ns}	1.43	1.71	0.12	0.35	0.47	0.17
Set	30.06	-0.0002^{ns}	-0.733^{ns}	0.021^{ns}	1.20	1.37	0.22	0.47	0.61	0.29
Out	26.21	-0.0028^{ns}	-0.506^{ns}	0.067^{ns}	0.66	0.81	0.50	0.71	0.81	0.57
Nov	23.23	-0.0049*	-0.344^{ns}	0.126^{ns}	0.31	0.40	0.86	0.93	0.96	0.89
Dez	22.09	-0.0039*	-0.133^{ns}	0.054^{ns}	0.29	0.36	0.82	0.91	0.94	0.86

^{ns} no significant. *significant at 5% probability by F-test.

The models used to estimate Tmean (Table 4) reveal that R^2 values were in the range of 0.38 – 0.93 and the r valued ranged from 0.62 to 0.97. The models for estimating the data for the months of July and August exhibited a “bad” and “poor” performance (Camargo; Sentelhas, 1997), respectively. For these months, these models are not recommended to estimate the Tmean values. The model performances were “Good” when the other months were analyzed. The linear coefficients altitude (β1) and latitude (β2) were significant. A negative correlation was observed between altitude and Tmean and between latitude and Tmean, exhibiting a decrease in Tmean values with increasing altitude and latitude. These results were expected and in accordance with the vertical thermal gradient in the troposphere. Cargnelutti Filho, Maluf and Matzenauer (2008) and Gomes et al. (2014) reported a negative correlation between altitude and Tmean (Rio de Janeiro state and the Rio Grande do Sul state, respectively). However, there was no significant influence in latitude.

During the estimation of Tmax, RMSE was found to be in the range of 0.51 – 0.74. The R^2 values ranged between 0.63 and 0.86, and the r values ranged between 0.80 and 0.93 (Table 4). The model for February exhibited the lowest statistical indicators, and the model’s performance was “Good” (Camargo; Sentelhas, 1997). The linear coefficient of altitude (β1) was significant in all models. There was no significant influence of the linear coefficients

longitude (β3) on the months of January, February, and March. In the other months, a significant influence of β2, β3, and β4 was observed. Gomes et al. (2014) analyzed the models to estimate the maximum monthly air temperature of Rio de Janeiro. R^2 values were found to be in the range of 0.51 – 0.71. A significant influence of the altitude and latitude was observed. However, the linear coefficient of longitude did not significantly affect the data of most months. This difference can be explained by the small longitudinal difference between the meteorological stations in Rio de Janeiro state compared to the region evaluated in this study. The meteorological stations under consideration are at a sufficient longitudinal distance to be influenced by the continentality effect.

While estimating Tmin, it was observed that the r values ranged between 0.32 and 0.93. The R^2 values ranged between 0.10 and 0.86, and the RMSE values ranged between 0.36 – 1.92 (Table 4). The models used for estimating the Tmin values for the months between February and October exhibited a “Poor”, “Bad”, or “Terrible” performance index (Camargo; Sentelhas, 1997), reflecting the low precision and degree of accuracy. Furthermore, significant β1, β2, and β3 values were not recorded when these models were used to study the data corresponding to the abovementioned months.

The Tmin, corresponding to these months, varied due to the variation in other factors, such as wind, ocean currents, local topographic conditions, rain, cloudiness, and passage of the cold front (Aguado; Burt, 2010). According to Silveira et al. (2019), in addition to the statistical factors (vegetation, maritime, continentality, geographic coordinates, etc.), climatic conditions are influenced by dynamic atmospheric systems such as cold fronts. After the passage of the cold front, under conditions of clear skies and low atmospheric humidity, the heat loss by irradiation during the night is very high. This results in a drop in temperature,

mainly during winter, autumn, and spring. In some cases, this facilitates the occurrence of radioactive frosts (Escobar, 2007).

Therefore, the Tmin values could not be estimated with high precision using these models. In the other months (November, December, and January), the models performed well, and a significant influence of altitude was observed. Medeiros et al. (2005) (the Northeast region of Brazil) Cargnelutti Filho et al. (2006) (the Rio Grande do Sul state), and Gomes et al. (2014) (the Rio de Janeiro state), observed similar results. The altitude influenced the Tmin values the most.

The ANN and RF statistical performance indicators for estimating Tmean, Tmax, and Tmin in the regions classified as Cwa and Cwb (Minas Gerais state) are shown in Table 5. Contrary to the MLR model, which used separate equations for each month, the architectures chosen for the ANN and RF models could be used to estimate the Tmean, Tmax, and Tmin of all months together. Thus, to estimate the Tmean, Tmax, or Tmin of a given location, latitude, longitude, altitude, and the month were used as the input data. Moreover, the statistics for each configuration (Table 5) refer to all the months of the year. The model performance indices for each month need not be distinguished (unlike the MLR model).

Table 5. Summary of the statistical tests conducted using the ANN and RF models.

	Tmean					
	MAE	RMSE	R^2	r	d	c
RF1	0.47 a	0.61 a	0.94	0.97 a	0.98	0.95
RF2	0.43 a	0.57 a	0.95	0.98 a	0.99	0.96
ANN1	0.67 b	0.80 b	0.90	0.96 a	0.97	0.92
ANN2	0.64 b	0.78 b	0.91	0.96 a	0.98	0.93
	Tmax					
RF1	0.46 a	0.55 a	0.94	0.97 a	0.98	0.95
RF2	0.44 a	0.56 a	0.94	0.97 a	0.98	0.95
ANN1	0.73 b	0.86 b	0.85	0.93 b	0.96	0.88
ANN2	0.65 b	0.79 b	0.87	0.94 b	0.97	0.90
	Tmin					
RF1	0.59 a	0.77 a	0.94	0.96 a	0.98	0.95
RF2	0.58 a	0.76 a	0.94	0.97 a	0.98	0.95
ANN1	0.88 b	1.07 b	0.89	0.94 a	0.97	0.91
ANN2	0.87 b	1.03 b	0.89	0.94 a	0.97	0.92

Mean values followed by the same letter in a column do not differ significantly (Tukey's test ($p \leq 0.05$)).

The lower RMSE and MAE were observed when the RF technique was used (compared to the case when ANN was used). A significant difference was observed in the results obtained using these techniques (ANN and RF). There was no significant difference between the different configurations tested within each technique. The RMSE and MAE were higher estimating Tmin values compared to the Tmax and Tmean values, suggesting more variation within the Tmin estimates. The r values, calculated using the RF method, were higher than those calculated using the ANN method during the calculation of the Tmean, Tmax, and Tmin values. The values of the coefficient r did not differ significantly when these two techniques (and different configurations of the techniques) were used to determine the Tmean and Tmin values. Nevertheless, a significant difference was observed in the Tmax values when these two techniques were used. The other indices indicate that the RF model was superior to the ANN model. However, both the techniques could be used to estimate the Tmean, Tmax, and Tmin values with very high accuracy (Table 5). The fit quality of both

models can be confirmed by the high values of the performance index (c). These values were "Excellent" according to the evaluation criteria proposed by Camargo and Sentelhas (1997).

There was no significant difference between the RF configurations. However, the use of the concept of *Break ties randomly when several attributes look equally good*, a WEKA solution, increased the predictive capacity of the model. That option gets triggered when the output reaches a local optimum. When this condition becomes true, the algorithm initializes a random process to escape from a local optimal spot to reach the bests solutions. This procedure has been explained in detail by Breiman (2001). The previous studies suggest the execution of 100 interactions; however, 500 interactions were required to improve the RF performance. Were et al. (2015) reported more stable results using a higher number of interactions.

The changes made to the ANN parameters did not significantly influence the Tmean, Tmax, and Tmin values. However, increasing the *Number of Training Epochs* from 500 to 1000 improved the Tmean and Tmin predictive capacity of ANN. This *Number of Training Epochs* is a hyperparameter that defines the number of times the learning algorithm works through the entire training dataset. The best results were obtained when six neurons were integrated into one hidden layer during the estimation of Tmean and Tmin. However, the best result was obtained when five neurons were integrated into the hidden layer during the estimation of Tmax. The choice of the size of the hidden layer is very important because underestimated numbers of neurons can lead to poor approximation and generalization capabilities, while the use of excessive neurons can potentially result in overfitting. This can eventually make the search for the global optimum more difficult (Lee; Lam, 1995).

Although the MLR model could be used to estimate the Tmean, Tmax, and Tmin for some months of the year, in general, the RF and ANN models exhibited superior predictive

abilities (for all the analyzed statistical indices) than the MLR models. The RF model was found to be superior to the ANN model. Moreover, the low MLR predictive capacity (Tmin estimation) can cause problems for producers who need this information because the regions categorized as Cwa and Cwb are more suitable for the development of agricultural activities that require lower temperatures and average temperatures during the winter (below 20 °C; De Sá Júnior et al., 2012). Therefore, RF and ANN methods are more suitable for this region.

Several literature reports (reporting various applications) have indicated the superiority of the RF model in the regression estimation (Benali et al., 2019; Noi; Degener; Kappas, 2017; Rodríguez-Lado et al., 2015). The superiority of the RF model can be attributed to the advantages of the method, which include not making distributive assumptions about the predictors. The importance of each variable can be determined using this model, and the method is less sensitive to noise or overfitting (Armitage; Ober, 2010; Ismail; Mutanga, 2010). Even though RF is superior to ANN, the ANN method can be used to determine the Tmean, Tmax, and Tmin values with high accuracy. This has also been reported by Hasni et al. (2012). They concluded that the ANN technique could be reliably used for determining the temperatures.

The plot, shown in Figure 5, indicates the importance of each input attribute in the response variable of the evaluated algorithms. The most important contribution toward the estimation of the Tmean value was for the month. This was followed by the effect of the altitude (for all the evaluated models). In the estimate of Tmax by RF1 and RF2, the altitude exerted the maximum effect. However, when the ANN1 and ANN2 methods were used, the month was found to exert the maximum effect on the results. This was followed by the contribution of the altitude. The trend was similar to the trend observed when the MLR method was used. A significant influence of the altitude was observed for all months when the

MLR model was used for the calculations. The month attribute had the largest contribution to the Tmin estimate. This contribution was the maximum. These results can potentially explain the low capacity of the MLR model toward the estimation of Tmin as the month is not considered a variable in this model.

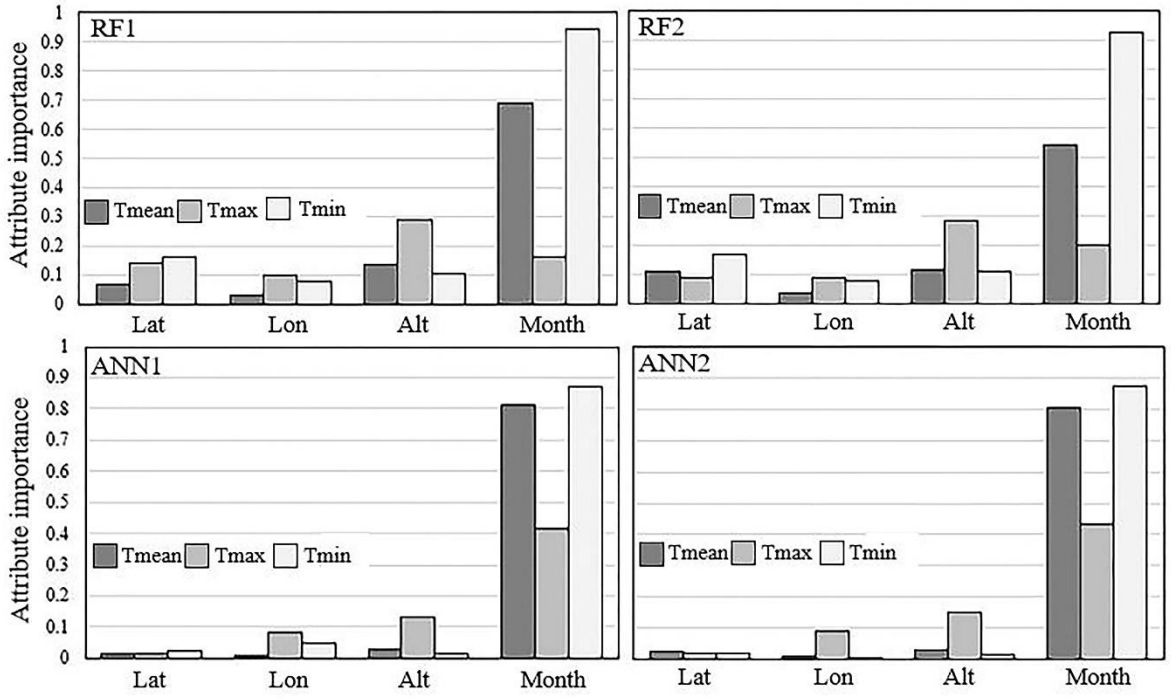


Figure 5. Attribute importance plots for the RF1, RF2, ANN1, and ANN2 models (Source: The Authors).

The results revealed that, for locations where it is difficult to collect data from weather stations (due to lack of infrastructure, reading errors, or use of damaged devices), the use of RF and ANN models is recommended for estimating the Tmean, Tmax, and Tmin values. In addition, researchers and producers can use such methods to create a risk zoning of pests and diseases, develop works related to plant growth, and develop crop varieties based on the temperature of the region.

An estimation of the Tmin values can help prevent the formation of frost in all the locations under study. This is because the region under analysis is susceptible to the occurrence of this phenomenon. According to Pimenta, Angélico and Chalfoun (2018),

adverse weather conditions (such as the formation of frost) can harm the production of the coffee fruit, affecting productivity and thereby changing the market value of the product. It is important to develop an efficient technique to determine the Tmax and Tmin values to develop a more accurate agricultural zoning of climatic risk. This can assist the producers in the choice of sowing time and harvest planning. Extreme weather conditions, especially in less developed regions, can be avoided. However, no statistical method can produce results that are exactly the same as the observed and/or recorded data. Hence, it is important that the weather stations function continuously (Alves et al., 2020). Furthermore, it is important to have computational knowledge to implement the RF and ANN models, therefore, mobile applications are needed to facilitate the use of these techniques. Further studies in the area are needed, and the results of the present study may support future forecasts.

CONCLUSION

The results of this study can help farmers, researchers, technicians, and local government officials in urban planning. Urbanization is characterized by surface alterations. Vegetated areas are replaced with impervious surfaces and buildings. This surface change alters the energy balance, increasing absorption and heat transfer between the earth's surface and the lower atmosphere, resulting in increased surface air temperatures (Song; Wu, 2016). Accelerated urban growth has been observed in the region under study. An effective tool for estimating the air temperature can assist in the application of new technologies that can potentially reduce the surface heating process.

The RF model exhibited a greater predictive performance compared to the ANN and MLR models for estimating the Tmean, Tmax, and Tmin values. The RF model explains at least 94% of the variability of the variables estimated using the independent dataset, i.e., only 6% of the response variable could not be predicted by the model. The RF is the most suitable

technique for estimating the air temperature. The input attributes were sufficient for the estimation. Therefore, this model is recommended for conducting studies in this specific region.

ACKNOWLEDGEMENTS

The authors express their gratitude to CAPES for scholarships that enabled the development of this research, and Brazilian National Institute of Meteorology (INMET) for making the series of meteorological data available.

REFERENCES

AGUADO. E.; BURT. J. E. **Understanding weather and climate**. 5th ed. New Jersey: Prentice Hall. 2010. 505p.

ALVARES. C. A. et al. Modeling monthly mean air temperature for Brazil. **Theoretical and Applied Climatology**, 113(3-4):407-427, 2013.

ALVES. M. P. A. et al. Reconstrução de dados e deteção de ondas de calor e de frio no Porto e concelhos vizinhos-Portugal. **Territorium**, 27(II):49-66, 2020.

ANTONOPOULOS. V. Z.; ANTONOPOULOS. A. V. Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. **Computers and Electronics in Agriculture**, 132:86-96, 2017.

ARMITAGE. D. W.; OBER. H. K. A comparison of supervised learning techniques in the classification of bat echolocation calls. **Ecological Informatics**, 5(6):465-473, 2010.

ASADI. H. et al. Rainfall-runoff modelling using hydrological connectivity index and artificial neural network approach. **Water**, 11(2):1-20, 2019.

BATISTA-SANTOS. P. et al. The impact of cold on photosynthesis in genotypes of Coffea spp.-Photosystem sensitivity. photoprotective mechanisms and gene expression. **Journal of Plant Physiology**, 168(8):792-806, 2011.

BENALI. L. et al. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam. horizontal diffuse and global components. **Renewable Energy**, 132:871-884, 2019.

BENAVIDES. R. et al. Geostatistical modelling of air temperature in a mountainous region of Northern Spain. **Agricultural and Forest Meteorology**, 146(3-4):173-188, 2007.

BENLLOCH-GONZÁLEZ. M. et al. Effect of moderate high temperature on the vegetative

growth and potassium allocation in olive plants. **Journal of Plant Physiology,** 207:22-29, 2016.

BOU-RABEE. M. et al. Using artificial neural networks to estimate solar radiation in Kuwait. **Renewable and Sustainable Energy Reviews**, 72:434-438, 2017.

BRASIL. Ministério da Agricultura e Reforma Agrária. Secretaria Nacional de Irrigação. Departamento Nacional de Meteorologia. **Normais climatológicas (1961-1990)**. Brasília: 1992. 84p.

BREIMAN. L. Random forests. **Machine Learning**, 45(1):5-32, 2001.

CAMARGO. A. P. DE.; SENTELHAS. P. C. Performance evaluation of different potential evapotranspiration estimating methods in the State of São Paulo. **Brazil. Revista Brasileira de Agrometeorologia**, 5(1):89-97, 1997.

CANNELL. M. G. R. Physiology of the coffee crop. In: CLIFFORD, M. N.; WILLSON, K. C. (eds). **Coffee**: Boston. MA: Springer, 1985. p.108-134.

CARDOSO. M. R. D. et al. Caracterização da temperatura do ar no estado de Goiás e no Distrito Federal. **Revista Brasileira de Climatologia**. 11:119-134, 2012.

CARDOSO. M. R. D. et al. Classificação climática de Köppen-Geiger para o estado de Goiás e o Distrito Federal. **Acta geográfica**, 8(16):40-55, 2015.

CARGNELUTTI FILHO. A.; MALUF. J. R. T.; MATZENAUER. R. Coordenadas geográficas na estimativa das temperaturas máxima e média decendiais do ar no Estado do Rio Grande do Sul. **Ciencia Rural,** 38(9):2448-2456, 2008.

CARGNELUTTI FILHO. A. et al. Altitude e coordenadas geográficas na estimativa da temperatura mínima média decendial do ar no Estado do Rio Grande do Sul. **Pesquisa Agropecuaria Brasileira**, 41(6):893-901, 2006.

COMPAHIA BRASILEIRA DE ABASTECIMENTO - CONAB. **Acompanhamento da safra brasileira de café**. Safra 2020 - Primeiro Levantamento. 6:1-62. 2020. Available in: <https://www.conab.gov.br/info-agro/safras/cafe>. Access in: April, 28, 2020.

COSTA. H. C. et al. Espacialização e sazonalidade da precipitação pluviométrica do estado de Goiás e Distrito Federal. **Revista Brasileira de Geografia Física**, 1:87-100, 2012.

DAMATTA. F. M. et al. Physiological and agronomic performance of the coffee crop in the context of climate change and global warming: A review. **Journal of Agricultural and Food Chemistry**, 66(21):5264-5274, 2018.

DE OLIVEIRA APARECIDO. L. E. et al. Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases. **International Journal of Biometeorology**, 64:671-688. 2020.

DE SÁ JÚNIOR. A. et al. Application of the Köppen classification for climatic zoning in the

state of Minas Gerais. Brazil. **Theoretical and Applied Climatology,** 108(1-2):1-7, 2012.

DUMEDAH. G.; COULIBALY. P. Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. **Journal of Hydrology**, 400(1-2):95-102, 2011.

ESCOBAR. G. C. J. Padrões sinóticos associados a ondas de frio na cidade de São Paulo. **Revista Brasileira de Meteorologia**, 22(2):241-254, 2007.

FABRIS. F. et al. A new approach for interpreting random forest models and its application to the biology of ageing. **Bioinformatics**, 34(14):2449-2456, 2018.

FAUSETT. L. **Fundamentals of neural networks:** Architectures. algorithms. and applications. Prentice Hall. Upper Saddle River. New Jersey. 1994. 461p

FERREIRA. D. F. SISVAR: A computer analysis system to fixed effects split plot type designs. **Revista Brasileira de Biometria**, 37(4):529-535, 2019.

GOMES. D. P.et al. Estimativa da temperatura do ar e da evapotranspiração de referência no estado do Rio de Janeiro. **Irriga**, 19(2):302-314, 2014.

HASNI. A. et al. Estimating global solar radiation using artificial neural network and climate data in the south-western region of Algeria. **Energy Procedia**, 18:531-537, 2012.

HESTERBERG. T. et al. Bootstrap methods and permutation tests. Ch18. In: MOORE, D.S; MCCABE, G.P; DUCKWORTH, W.M; SCLOVE, S.L. **The Practice of Business Statistics**. NY: W.H. Freeman and Co., 2002. 18:4-25.

ISMAIL. R.; MUTANGA. O. A comparison of regression tree ensembles: Predicting Sirex noctilio induced water stress in Pinus patula forests of KwaZulu-Natal. South Africa. **International Journal of Applied Earth Observation and Geoinformation**, 12(1):S45-S51, 2010.

JAMES. G. et al. **An introduction to statistical learning**. New York: Springer. 2013. v. 112. 18p.

KÖPPEN. W.; GEIGER. R. **Klimate der Erde**. Gotha: Verlag Justus Perthes. Wall-Map 150cmx200cm. 1928.

KUMAR. M.; RAGHUWANSHI. N. S.; SINGH. R. Artificial neural networks approach in evapotranspiration modeling: A review. **Irrigation Science**, 29(1):11-25, 2011.

LEE. K. W.; LAM. H. N. Optimal sizing of feedforward neural networks: Case studies. **Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems**, 79-82, 1995.

MEDEIROS. S. S. de. et al. Estimativa e espacialização das temperaturas do ar mínimas.

médias e máximas na Região Nordeste do Brasil. **Revista Brasileira de Engenharia Agrícola e Ambiental**, 9(2):247-255, 2005.

MOREIRA. M. C.; CECÍLIO. R. A. Software to estimate air temperature in the brazilian northeastern region using artificial neural networks. **Revista Engenharia na Agricultura-Reveng**, 24(2):164-171, 2016.

MUHAMMAD. W. et al. Pancreatic cancer prediction through an artificial neural network. **Frontiers in Artificial Intelligence**, 2:2, 2019.

MWALE. F. D.; ADELOYE. A. J.; RUSTUM. R. Infilling of missing rainfall and streamflow data in the Shire River basin, malawi - A self organizing map approach. **Physics and Chemistry of the Earth**, 50(52):34-43, 2012.

NOI. P. T.; DEGENER. J.; KAPPAS. M. Comparison of multiple linear regression. cubist regression. and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. **Remote Sensing**, 9(5):398, 2017.

PARTELLI. F. L. et al. Low temperature impact on photosynthetic parameters of coffee genotypes. **Pesquisa Agropecuária Brasileira**, 44(11):1404-1415, 2009.

PEZZOPANE. J. et al. Espacialização da temperatura do ar no Estado do Espírito Santo. **Revista Brsileira de Agrometeorologia**, 12(1):151-158, 2004

PIMENTA. C. J.; ANGÉLICO. C. L.; CHALFOUN. S. M. Challengs in coffee quality: Cultural. **Ciência e Agrotecnologia**, 42(4):337-349, 2018.

RODRÍGUEZ-LADO. L. et al. A pedotransfer function to map soil bulk density from limited data. **Procedia Environmental Sciences**, 27:45-48, 2015.

SCORNET. E. On the asymptotics of random forests. **Journal of Multivariate Analysis**, 146:72-83, 2016.

SEDIYAMA. G. C.; MELO JÚNIOR. J. C. F. Modelos para estimativa das temperaturas normais mensais médias. máximas. mínimas e anual no estado de Minas Gerais. **Engenharia Na Agricultura**, 6(1):57-61, 1998.

SEYEDHOSSEINI. M.; TASDIZEN. T. Disjunctive normal random forests. **Pattern Recognition**, 48(3):976-983, 2015.

SILVEIRA. R. B. et al. Ondas de calor nas capitais do Sul do Brasil e Montevidéu - Uruguai. **Revista Brasileira de Geografia Física**, 12(4):1259-1276, 2019.

SONG. Y.; WU. C. Examining the impact of urban biophysical composition and neighboring environment on surface urban heat island effect. **Advances in Space Research**, 57(1):96-109, 2016.

VOGELS. M. F. A. et al. Agricultural cropland mapping using black-and-white aerial

photography. object-based image analysis and random forests. **International Journal of Applied Earth Observation and Geoinformation**, 54:114-123, 2017.

WAHID. A. et al. Heat tolerance in plants: An overview. **Environmental and Experimental Botany**, 61(3):199-223, 2007.

WANG. H. et al. Intelligent identification of maceral components of coal based on image segmentation and classification. **Applied Sciences (Switzerland)**, 9(16):1-15, 2019.

WERE. K. et al. A comparative assessment of support vector regression. artificial neural networks. and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators**, 52:394-403, 2015.

WU. W.; DANDY. G. C.; MAIER. H. R. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. **Environmental Modelling & Software**, 54:108-127, 2014.

XIE. Z. et al. Artificial intelligence for rapid identification of the coronavirus disease 2019 (COVID-19). **medRxiv**, e20062661, 2020.

ZHOU. X. et al. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. **The Crop Journal**, 4(3):212-219, 2016.

ARTICLE 2 - EVALUATION OF MONTHLY MEAN REFERENCE EVAPOTRANSPIRATION ESTIMATION TECHNIQUES IN THE MINAS GERAIS STATE, BRAZIL

EVALUATION OF MONTHLY MEAN REFERENCE EVAPOTRANSPIRATION ESTIMATION TECHNIQUES IN THE MINAS GERAIS STATE, BRAZIL.

AVALIAÇÃO DE TÉCNICAS DE ESTIMATIVAS DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA MÉDIA MENSAL NO ESTADO DE MINAS GERAIS.

Abstract: Reference evapotranspiration (ET_0) is one important agrometeorological parameter for hydrological studies and agricultural water management. The ET_0 esteemed by the Penman-Monteith - FAO method requires several input data. However, in the Minas Gerais region, the meteorological data are limited. The aim of this study was to evaluates the performance of Artificial Neural Network (ANN), Random Forest (RF), Support Vector Machine (SVM) and Multiple Linear Regression (MLR) to estimate the monthly mean ET_0 with different input data combinations and in three scenarios: (Scenario I - SI) at the state level, where all climatological stations were used; and at regional level, where the Minas Gerais state was divided according to the climatic classification of each climatological stations. The climatic classifications proposed by Thornthwaite (Scenario II - SII) and by Köppen (Scenario III - SIII) were used. ANN and RF performed better in SI, SII and SIII with the I_8 (latitude; longitude; altitude; month; mean, maximum and minimum temperature; and relative humidity) or I_6 (latitude; longitude; altitude; month; mean temperature; and relative humidity) input data. The SVM and MLR performed better in all scenarios when only two input variables were used (I_2 - mean temperature and relative humidity). Although dividing into scenarios results in less input data for models training, SII and SIII showed a slightly better result in the southern areas of the Minas Gerais state.

Keywords: Artificial Neural Network, Random Forest, Support Vector Machine**,** Multiple Linear Regression,

Resumo: A evapotranspiração de referência (ET_0) é um parâmetro agrometeorológico importante para estudos hidrológicos e gestão de água na agricultura. A ET_0 estimado pelo método Penman-Monteith - FAO requer vários dados de entrada. Porém, na região de Minas Gerais, os dados meteorológicos são limitados. O objetivo deste estudo foi avaliar o desempenho de Redes Neurais Artificiais (RNA), Floresta Aleatória (FA), Máquina de Vetores de Suporte (MVS) e Regressão Linear Múltipla (RLM) na estimativa da ET_0 média mensal com diferentes combinações de dados de entrada e em três cenários: a nível estadual

(cenário I - SI), em que todas as estações climatológicas foram utilizadas; e a nível regional, em que o estado de Minas Gerais foi dividido de acordo com a classificação climática de cada estação climatológica. Foram utilizadas as classificações climáticas propostas por Thornthwaite (cenário II - SII) e por Köppen (cenário III - SIII). RNA e FA tiveram melhor desempenho nos SI, SII, SIII com a combinação de dados de entrada I_8 (latitude; longitude; altitude; mês; temperatura média, máxima e mínima; e umidade relativa) ou I_6 (latitude; longitude; altitude; mês; temperatura média; e umidade relativa). O MVS e o RLM tiveram melhor desempenho em todos os cenários quando apenas duas variáveis de entrada foram usadas (I_2 - temperatura média e umidade relativa). Embora a divisão em cenários resulte em menos dados de entrada para o treinamento de modelos, os cenários II e III mostraram um resultado ligeiramente melhor nas áreas mias ao Sul do estado de Minas Gerais.

Palavras-chave: Rede Neural Artificial, Floresta Aleatória, Máquina de vetor de suporte, Regressão Linear Múltipla.

INTRODUCTION

Evapotranspiration (ET) is the process of water transportation from the Earth's surface to the atmosphere including the evaporation process and transpiration process. The ET is important in estimating crop water requirements, irrigation water requirements and control several hydrological processes (MATTAR, 2016; WEN et al., 2015; YASSIN; ALAZBA). ET is an agrometeorological parameter that can be measured using the lysimeter or water balance approach. These methods for measuring ET are not always possible to use. The lysimeter and water balance approach are a time-consuming method and needs precisely and carefully planned experiments (KUMAR et al., 2002). Therefore, it uses estimation methods from climatological data.

ET can be found in the literature as crop evapotranspiration (ET_c) or reference crop evapotranspiration (ET_0). Both concepts measure the transfer rate of water from the soil plant system to the atmosphere. However, ET_c measures ET for any crop, while ET_0 is the ET rate from a reference crop surface. One of the ways to obtain the ET_c is through the ET_0 and the crop coefficient (Kc) (CARVALHO et al., 2011; SALAM et al., 2020).

In this study, we focus on the use of ET_0. The ET_0 can be used on a large area, e.g., climatic classification of a region (ALMOROX; QUEJ; MARTÍ, 2015; YANG et al., 2017), or small areas, e.g., obtaining crop water requirements or crop evapotranspiration (ETc) (EWAID; ABED; AL-ANSARI, 2019; XIANG et al., 2020). ET_0 originated from the of Penman (1948) equation. This equation was perfected by Monteith (1965). Monteith introduced a surface conductance term that accounted for the response of leaf stomata to its hydrologic environment to the Penman equation, originating the equation called Penman-Monteith evapotranspiration model. Later, the Food and Agriculture Organization (FAO) of the United Nations adopted standard parameters for the culture data, creating a reference crop published in FAO Irrigation and Drainage Paper N 56 (ALLEN et al., 1998). This process gave rise to the FAO standard reference crop evapotranspiration model (ET_{PM}). More information can be obtained in Xiang et al. (2020).

The FAO Penman-Monteith method is a nonlinear and complex. This method is considered more realistic physically, but it requires some additional meteorological variables when compared to other methods (YANG et al., 2017). This dependence on several meteorological variables associated with the limited of weather stations network and interruptions and errors in the database makes it difficult to measure ET_0. Thus, some models are used to estimate ET_0. These models seek less dependence on many weathers inputs e high predictive power.

Artificial Neural Network (ANN), Random Forest (RF), Support Vector Machine (SVM), and multiple linear regression (MLR), are models that showed a different levels predictive capacity of different meteorological variables and in other fields of science. The ANN, RF, and SVM models can capture complex relationships between input and output data, which makes them powerful models in modeling. These machine learning models have been successfully used to estimate ET_0 with fewer input meteorological data (FERREIRA et al., 2019; FERREIRA; DA CUNHA, 2020; YIN et al., 2017;). Although the inability of MLR to handle non-linear relationships between dependent and independent variables is evident in some studies, the MLR have been successfully used to estimate ET_0 (MALIK et al., 2019; MARTÍ; GONZÁLEZ-ALTOZANO; GASQUE, 2011).

The ANN is a promising and effective tool for non-linear modeling and complex time-series. The ANNs are parallel distributed systems, which are composed of simple processing

units that calculate some mathematical functions. Its customary architecture is composed of three layers: input, hidden, and output layers and each layer include an array of processing elements (FERREIRA et al., 2019; KUMAR et al., 2002; YIN et al., 2017). According to Yin et al., 2017, experimental works have shown that a single hidden layer is sufficient for ANNs to approximate any complex nonlinear function. However, different architectures must be tested. Several papers showed the excellent predictive capacity of the ANN model with different architectures in studies with ET_0 (FERREIRA; DA CUNHA, 2020; NOURANI; ELKIRAN; ABDULLAHI, 2019; SATTARI et al., 2021).

The RF is non-parametric statistical data modeling methods that is decision tree-based and uses the Breiman's "bagging" idea to ensemble many decision trees into a single but strong model (BREIMAN, 2001). This model generates original training samples (N), and from these samples generates new training random sample sets (k). During the overall selecting process, some samples may be collected more than once. Trees alone are considered a weak learner. However, the ensemble trees result in a model with high predictive capacity (HUANG et al., 2019). RF is a classification and regression technique that also has been adopted to predict agrometeorological parameter such ET_0 (FERREIRA; DA CUNHA, 2020; FENG et al., 2017; WANG et al., 2019b). RF has been found to be a more efficient predicting tool compared to other tools like ANN (BENALI et al., 2019; ZHOU et al., 2016).

SVM is a supervised machine learning algorithm developed by (VAPNIK, 2013). The SVM is used for regression, classification, pattern recognition and forecasting. SVM based on a statistical learning theory and concept of the structural risk minimization principle, which reduces the upper bound generalization error rather than the local training error (SHIRI et al., 2014; FENG; WEN; LI, 2015). This model has been used in the meteorological variables estimation and shown a high predictive power. The same training conditions and in some locations, the SVM has proven superior performance in the ET_0 estimate over other methodologies as ANN and GEP (Gene expression programming) (MOHAMMADREZAPOUR; PIRI; KISI, 2019; SHIRI et al., 2014).

The MLR aim at explaining the collinearity between a dependent variable and independent by means of a linear combination of predictors independent variables (more than one). This regression technique has been adopted in several fields of science, including climatology, hydrology, and irrigation, with different performances. According to Martí,

González-Altozano and Gasque (2011), using the same input data, the studied ANNs present very similar accuracy indicators and performance trends as the multiple linear regression models. However, the complexity of the input data can change this performance.

Specifically, this paper focuses on ET_0 estimation in the Minas Gerais state, Brazil. Agriculture is an important source of income in the Minas Gerais state. Coffee crop science is the main cultivated culture in the state (COMPAHIA BRASILEIRA DE ABASTECIMENTO - CONAB, 2020). In addition, the Minas Gerais state brings together water springs of important rivers as São Francisco River. The ET_0 is a fundamental variable in the study of water springs and in the agricultural development, because through this variable it is possible to measure the volume of water transferred from a hydrological basin or from a cultivated area to the atmosphere. The presence of gaps or discontinuities in the data series can delay the state development.

The 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 objective of this study was to develop, evaluate and compare the performance of ANN, RF, SVM and MLR models in estimating ET_0 (average monthly) with four different combinations of input data (I_8, I_6, I_3 and I_2) in three scenarios: at the state level (SI), in which all climatological stations are used in the models build; at regional level, in which the climatological stations were divided into two areas, according to the Thornthwaite climate classification (SII) and Köppen climate classification (SIII). The models were developed for each area in SII and SIII. The three distinct scenarios build seeks to achieve the maximum predictive capacity of each model

MATERIALS AND METHODS

STUDY AREA AND DATA SOURCES

The Minas Gerais state is the fourth largest in territorial extension with 586,513.993 km² (IBGE, 2020). Minas Gerais is in the southeastern region of Brazil, between the parallels of 14° 13' 58" and 22° 54' 00", of south latitude, and the meridians of 39° 51' 32" and 51° 02' 35" a west of Greenwich. Monthly data from 56 climatological stations of the Brazilian National Institute of Meteorology (INMET) were used. The respective geographical coordinates, altitude and climatic classification have been presented in Table 1.

Table 1 - Principal climatological station of the INMET used to estimate estimating ET_0.

ID	CS	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	-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
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	-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

CS: Climatological stations location; K: Köppen climatic classification; Tho: Thornthwaite climatic classification; Cwb, Cwa, Cfb, As and Aw are Humid subtropical with dry winter and temperate summer, Humid subtropical with dry winter and hot summer, Humid subtropical with oceanic climate, without dry season and temperate summer, Tropical with dry summer, Tropical with dry winter, respectively. A, B4, B3, B2, B1, C2, C3, and D are perhumid, humid, humid, humid, humid, moist subhumid, dry subhumid and semiarid, respectively.

Source: The Authors (2021)

The average monthly data mean, maximum, and minimum air temperatures (Tmean, Tmax, Tmin), relative humidity (RH), atmospheric pressure (P), wind speed (U_2), and insolation (n) of climatological stations with at least 10 years of flawless data (no missing or faulty data) from a period between 1989 to 2019 (30 years) were used for the studies. Wind speed, measured at 10 m height, was converted to 2 m (ALLEN et al., 1998). 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 out of the range 0 – 100 %; P was higher than 101.4 kPa and U_2 and n were negative.

The reasons for using these variables were: The latitude and longitude are the variables related to the position. The solar radiation intensity changes as the 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 the 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.

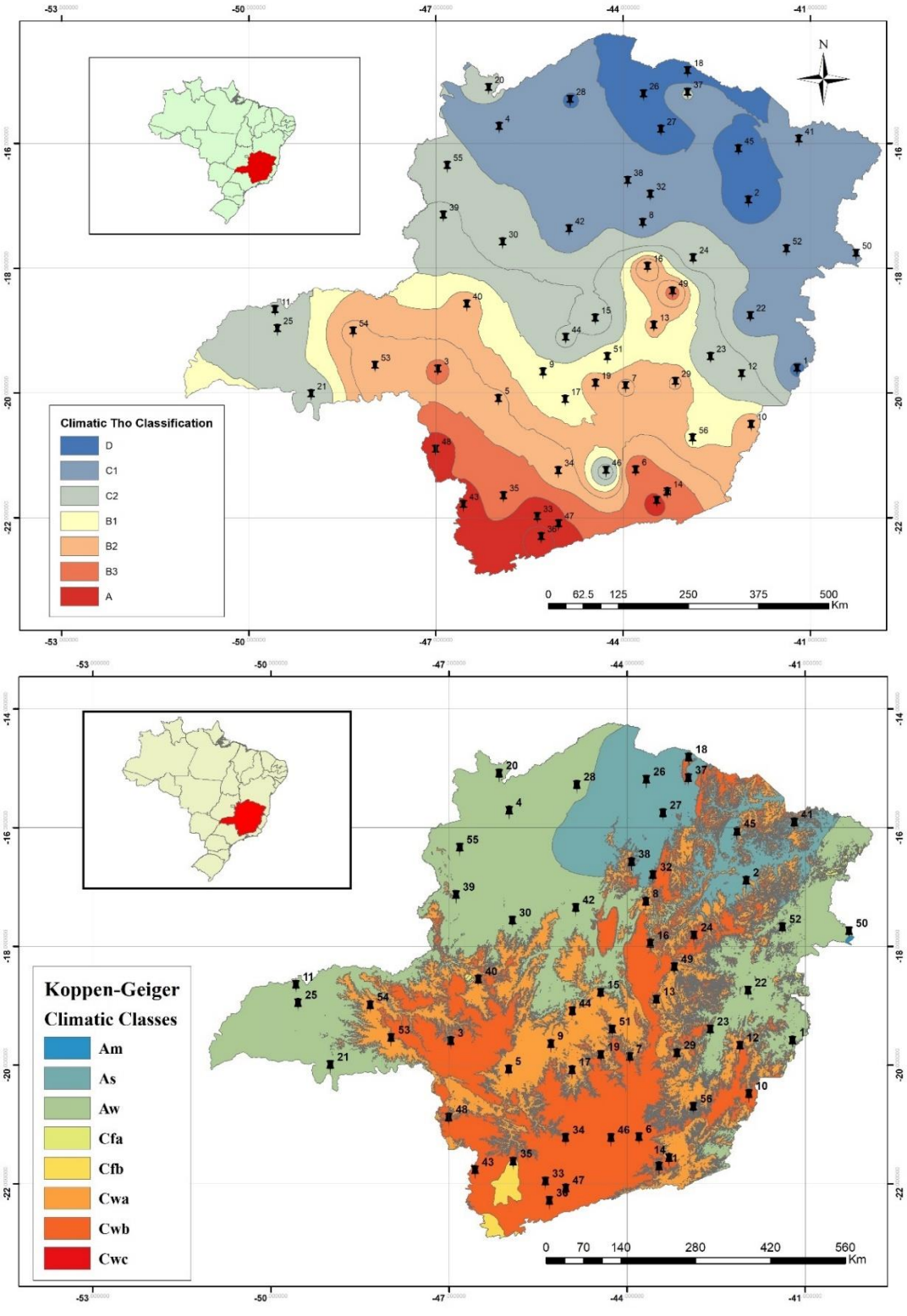
The models were developed in three different scenarios (SI, SII and SIII) build seeks to achieve the maximum predictive capacity of each model: SI - At State Level: the models

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. The region with climate classification A, B4, B3, B2 and B1 (Tho1 - 27 climatological stations), and the region with climate classification C2, C1 and D (Tho2 - 29 climatological stations) Climatic Classification Systems proposed by Thornthwaite (1948). The models were trained and tested with data from climatological stations of each climatic region (Figure 1a).

SIII - At Regional Level: The Minas Gerais state was divided into two regions. The region with climate classification Cwb, Cwa and Cfb (K1 - 35 climatological stations) and the region with climate classification Aw and As (K2 - 21 climatological stations) Climatic Classification Systems proposed by Köppen (1936). The models were trained and tested with data from climatological stations of each climatic region (Figure 1b).

Figure 1. Climate classification for Minas Gerais State, according to the Thornthwaite (1948) (Source: The Authors, 2021) (a), and Köppen (1936) (Source: Alvares et al., 2013) (b).



FAO PENMAN–MONTEITH FAO MODEL

The FAO Penman–Monteith equation (FPM) was used to estimate ET_0. This method is described by Allen et al. (1998). It is a common practice to use ET_0 estimated by the FPM equation as reference data. Climatological stations used in the study do not provide Rn data. The Rn data were obtained by means of insolation, latitude, day of the year and other variables. They are calculated using the equations detailed by Allen et al. (1998).

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 (I_n – *n* is the amount of input data) evaluated in this paper were: (I_8) latitude, longitude, altitude, month, Tmean, Tmax, Tmin, and UR; (I_6) latitude, longitude, altitude, month, Tmean and RH; (I_3) month, Tmean and UR; (I_2) Tmean and RH. The ANN, RF, SVM and MLR model were trained for each combination. The models were developed using data from each scenery. These combinations were compared to each other in each model.

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

ARTIFICIAL NEURAL NETWORKS (ANN)

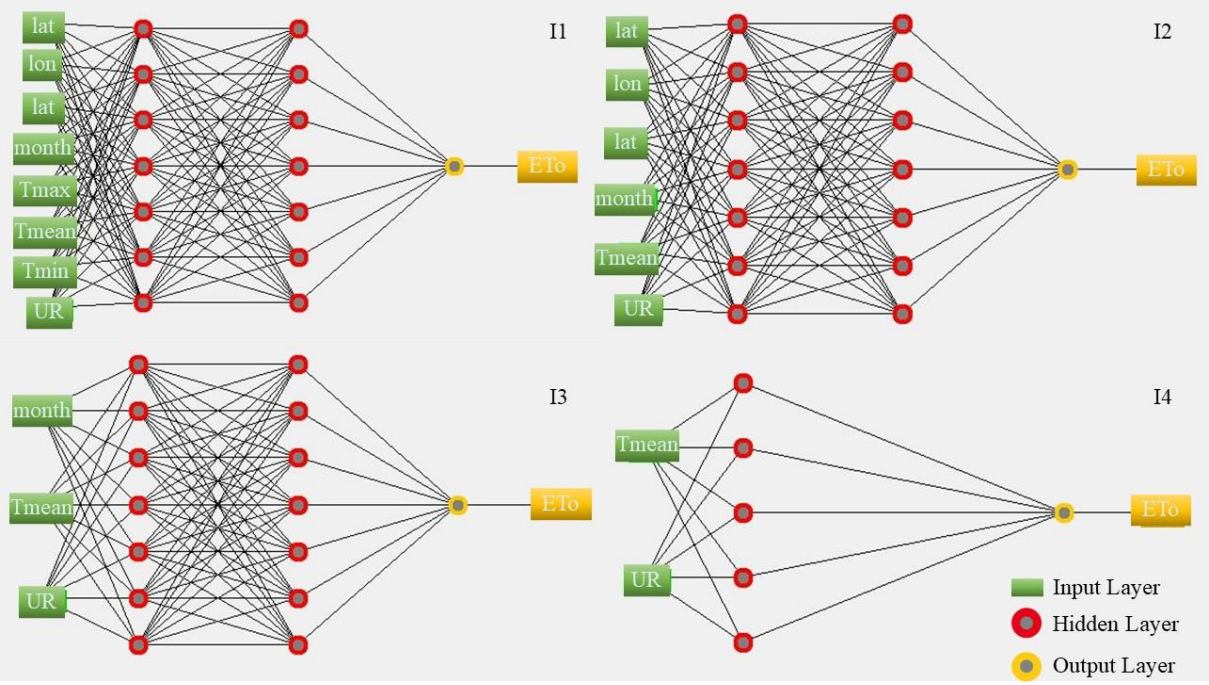
The ANN has performance characteristics resembling biology of the human brain. ANNs, in general, have an architecture with connections between nodes (neural networks) and methods to determine the connections weight. In this study, the ANN of the feed-forward multilayer perceptron (MLP) type was used (FAUSETT, 1994). The training in this ANN involves two phases. In the first phase or forward pass, the input sign spreads layer-by-layer forward. In the second phase or reverse pass, the sign is backpropagated for the 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 UR of each evaluated location, using ET_0 as the output variable.

All adjustments were by cross-validation. According to Sattari et al. (2021), the cross-validation approach enables successful results. This technique separates the data into two categories, where the first is used to train the model and the second part is processed as test data to determine the model's performance. After preliminary tests, the architecture with the best performance for each data input combination was obtained (Figure 2). The ANN different configurations and number of folds in cross-validation are shown in Table 2.

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



Source: The Authors (2021)

Table 2. WEKA configuration in the ANN implementation.

	ANN			
	I1	I2	I3	I4
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

Source: The Authors (2021).

SUPPORT VECTOR MACHINE (SVM)

In this study, it was applied SVM equations based on Vapnik's theory (VAPNIK, 2013). SVM are separated into two main categories: (1) the classifier model and (2) 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 (SATTARI et al., 2021). SVR can be translated into the following equation:

$$y = f(x) = w\varphi(x_i) + b \quad (1)$$

where *x* is the input data; φ(x) represents a function, which can transfer the *x* into the high-dimensional feature spaces; *ω* (weight vector) and b are coefficients which are estimated by minimizing the regularized risk function. The error function in SVM model (Equation 2) is minimized based on mentioned constraints in Equations 3. Further details on the application of SVM can be found in Chang & Lin (2012).

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

$$Subject\ to \begin{cases} y_i - b\,[w.\varphi(X_i)] \leq \varepsilon + \xi_1 \\ [w.\varphi(X_i)] + b - y_i \leq \varepsilon + \xi_1^* \\ \xi_1\xi_1^* \geq 0 \end{cases} \quad (3)$$

where, C is the capacity or penalty parameter, y_i is the estimated output by SVM, ξ_i and ξ^*_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 have been tested. The linear Kernel

function proved to be more efficient in estimating ET_0. The linear kernel function is as follows:

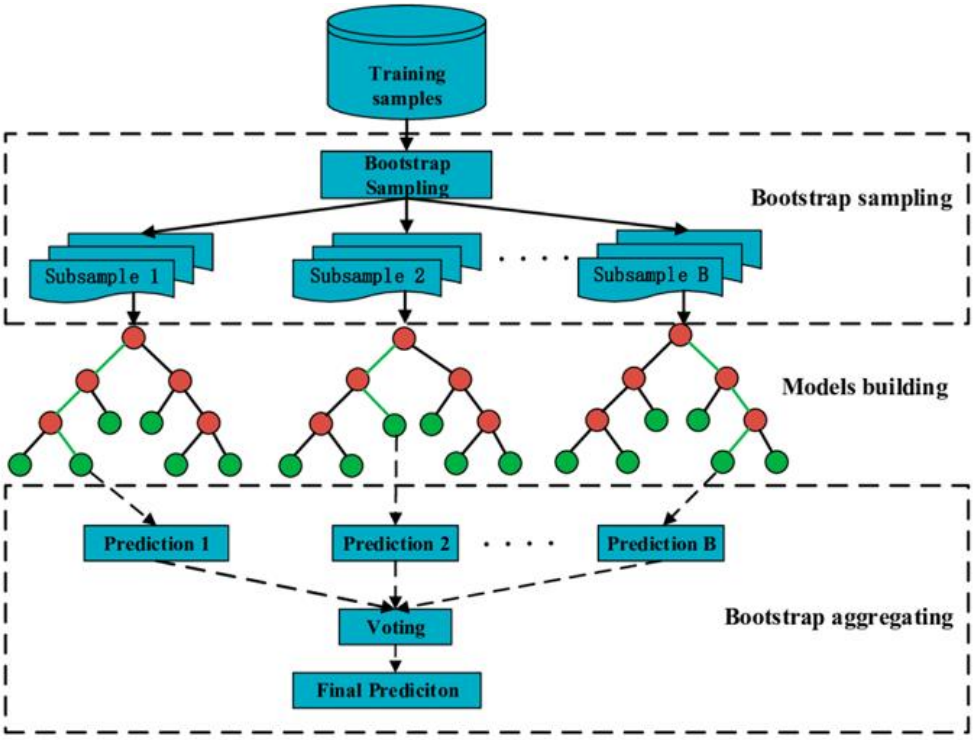
$$K(x_i, y_i) = x_i \cdot y_i \quad (4)$$

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 (I_8, I_6, I_4 and I_2) and were evaluated in the three different scenarios. WEKA configuration parameters in the SVM implementation were: SVM Type, ϵ-SVR and cost parameter C, 0.01. The other WEKA configuration parameters were kept as standard. Eighteen folds of the sample set were used in cross-assessed. The same WEKA configuration parameters were used in all input data combinations and in all scenarios.

RANDOM FOREST (RF)

RF is an ensemble learning technique based on a collection of tree predictors (XU; KNUDBY; HO, 2014). It is a combination of many predictor trees (forest), in which each tree is generated from a random vector, sampled independently and with the same distribution for all trees in the forest. The operating steps of the RF model were presented in Figure 3. According to Wang et al. (2019b), 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; (iii) and predict new data by aggregating the predictions of the *n*. Further details can be found in Wang et al. (2019a) and Wang et al. (2019b).

Figure 3 - Schematic representation of the steps used in the RF model following the resampling strategy.



Source: Wang et al. (2019a).

RF was implemented by the WEKA. The WEKA configuration that resulted in the greatest predictive capacity was: 100 is the size bag, as percentage of the training set size and 500 iterations. The other WEKA configuration parameters were kept as standard. All adjustments were by cross-validation and twenty folds of the sample set were used. The same WEKA configuration parameters were used in all input data combinations and in all scenarios.

MULTIPLE LINEAR REGRESSION (MLR)

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

$$Y_i = \beta_0 + \beta_1 lat + \beta_2 lon + \beta_3 alt + \beta_4 month + \beta_5 Tmax + \beta_6 Tmean + \beta_7 Tmin + \beta_8 UR. \quad (5)$$

where Y_i is the dependent variable (ET_0); lat, lon, alt, month, Tmean, Tmax, Tmin and UR are independent variables; and β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7 and β_8 are the regression coefficients.

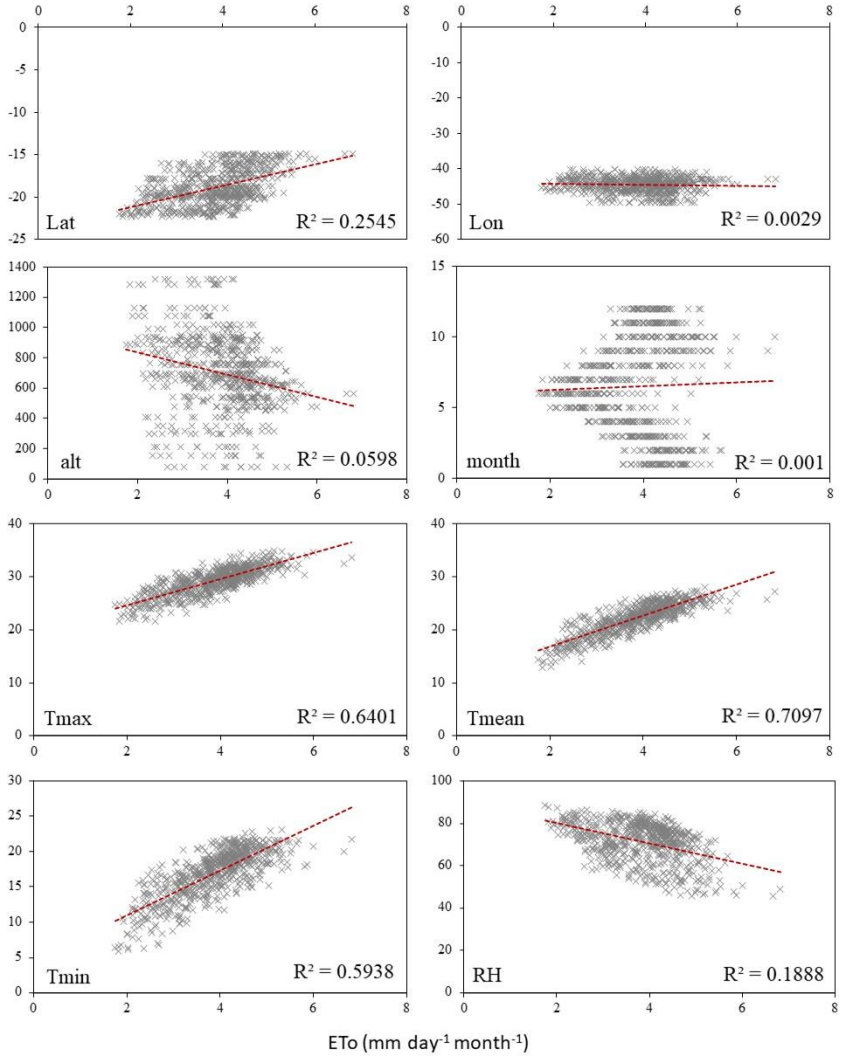
MLR was implemented using the WEKA. The attribute selection method used in the WEKA configuration was M5 method. This method initially builds the MLR model with all

independent variables. Then, independent variables with the smallest standardized coefficient are step-wisely removed 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, by adding the more parameters, the AIC of the model may present inadequate performance (SAMADIANFARD et al., 2018). The other WEKA configuration parameters were kept as standard. Eighteen folds of the sample set were used in cross-assessed. The same WEKA configuration parameters were used in all input data combinations and in all scenarios.

RESULTS AND DISCUSSIONS

The linear correlation between the input data and the ET_0 are shown in Figure 4. The variables Tmean, Tmax, and Tmin showed the best correlation. The other variables have low (lat, alt and RH) or no (lon and month) correlation with ET_0. The inversely proportional behavior with ET_0 was observed in the lat, alt and RH variable. High latitudes tend to be cooler regions, with less energy available for the ET_0 process. Increasing the altitude decreases the temperature according to the vertical thermal gradient in the troposphere. The increase in RH increases the potential gradient, increasing the water transfer rate from the soil-plant system to the atmosphere. However, a proportional behavior was observed between the Tmean, Tmax and Tmin variables with ET_0. The increase in Tmean, Tmax and Tmin results in more energy available for ET_0. Sattari et al. (2021) observed the same behavior of the variables Tmen, 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.

Figure 4. Scatter plots of the relationship between ET_0 and each input data



Source: The Authors (2021)

The ability of machine learning approaches in different conditions and scenarios was investigated. 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 -) is presented in Table 3. All the models developed with the I_8 and I_6 input combination exhibited better performances than their versions developed with the I_3 and I_2. The lowest predictive capacity was observed when the RF model was used with the I_8 input combination. The greatest predictive capacity, in SI, was observed when the RF and ANN models was used with the I_6 and I_8 input combination, respectively. The models SVM and MLR exhibited better performances than ANN and RF when only Tmean and HRmean (I_2) were used as input data.

When comparing the combination I_8 with I_6, the average r, MAE and RMSE of all models does not show high variations. The removal of the geographic coordinate (I_6 to I_3) resulted in a more expressive performance reduction of the SVM and MLR models. The highest impact on performance was observed in the ANN and RF when the month variable was removed (I_3 to I_2). The 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.

Table 3. Performances 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*	*0.21*	*0.278*	0.860	0.332	0.429
RF	0.955	0.191	0.25	***0.966***	***0.166***	***0.22***	0.934	0.22	0.296	0.859	0.335	0.426
SVM	**0.933**	**0.23**	**0.29**	0.927	0.242	0.31	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*

Value in bold indicates the best result within each model; value in italics indicates the best result within input data combination

Source: The Authors (2021)

The statistical performance indicators of the models used in the ET_0 estimation in SII is showed in Table 4. The Minas Gerais state was divided into two areas (Tho1 and Tho2) for application of model. Tho1 and Tho2 had 48.2% and 51.8, respectively, of the data available as input data. The highest predictive capacity in Tho1 and Tho2 area was observed when the ANN model was used with the I_8 input combination and RF model was used with the I_6 input combination, respectively. The removal of the Tmax and Tmin input data (I_6) did not increase the model's predictive capacity in Th1 area, except for the RF model. This behavior is similar to that observed in SI. However, all models performed better when the I_6 input combination in Tho2 area was used (better results).

The removal of the month variable (I_3 to I_2) resulted in the highest impact on the ANN and RF models quality. When comparing the combination I_8 with I_3, the average r of the ANN and RF models decreased by 7.2% and 5.7%, respectively. The MAE of the ANN and RF models increased by 36.4% and 31.6%, respectively. However, no expressive variation was observed in the performance of SVM and MLR models.

Table 4. Performances 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	0.19	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***	***0.21***	***0.26***	*0.895*	*0.269*	*0.353*	0.818	0.377	0.462
RF	0.925	0.227	0.302	**0.943**	**0.2**	**0.267**	0.879	0.29	0.374	0.817	0.35	0.453
SVM	0.893	0.276	0.353	**0.898**	**0.271**	**0.346**	0.840	0.342	0.427	*0.840*	*0.34*	*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

Value in bold indicates the best result within each model; value in italics indicates the best result within input data combination

Source: The Authors (2021)

The statistical performance indicators of models in SIII are showed in Table 5. This scenario, the Minas Gerais state was divided K1 and K2 area. 62.5% and 37.5% of the climatological stations are distributed in areas K1 and K2, respectively. In general, the ANN and RF models were superior to the SVM and RLM models with the input combinations I_8, I_6 and I_3. When the I_2 combination is used, the SVM and RLM models were superior. The model with highest predictive capacity in K1 area was the ANN with the I_8 input combination. The RF model with the I_6 input combination showed highest predictive capacity in K2 area.

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

Table 5. Performances 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	0.17	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	0.36	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	**0.89**	**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*

Value in bold indicates the best result within each model; value in italics indicates the best result within input data combination

Source: The Authors (2021)

The ANN and RF models showed greater predictive capacity in all scenarios when compared to the SVM and MLR models. This high capacity is achieved with the data input combination I_8 or I_6. Both models had similar performances, but on average the RF showed slight superiority. Ferreira et al. (2019) and Ferreira & Da Cunha (2020) conducted studies aimed at evaluating the performance of different machine learning in the ET_0 estimate for all states in Brazil and for Minas Gerais state, respectively. 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). Nevertheless, in some studies the RF model performed slightly better than other models (i.e., Generalized regression neural networks - GRNN) to estimating ET_0 (FENG et al., 2017; WANG et al., 2019b). There are papers suggesting better performance than other machine learning models in different situations and regions (MEHDIZADEH; BEHMANESH; KHALILI, 2017; SHIRI et al., 2014). Therefore, there is a need for studies that address more than one models.

The SVM and MLR models showed similar statistical indices and behavior in all scenarios. These results can be explained by the use of the linear Kernel function by SVM that probably presented behavior similar to an MLR. Tests with the nonlinear Kernel function did

not result in improvements in prediction. According to Pisner and Schnyer (2020), this SVM is used to recognize patterns in complex databases. Possibly the data used does not present a complexity that justifies the use of SVM.

The SVM and MLR models showed a greater predictive capacity in all scenarios when the input data limited to only Tmean and HR (I_6). 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 withdrawal of Tmax and Tmin gave the best result. Sattari et al. (2021) 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 the Tmax and Tmin showed a good correlation with ET_0 (Figure 4), the weight of Tmax and Tmin is diluted in the calculation of the Tmean used in the calculation of ET_0. Thus, adding Tmax and Tmin can make the ET_0 estimate 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 processes, improving the understanding of the system (BOWDEN; DANDY; MAIER, 2005; MAIER; DANDY, 2000). However, this result shows that only linear regression is not enough to decide which input data should be removed in order to increase the predictive performance.

When the independent variables lat, lon and alt were removed (I_3), 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. According to Mehdizadeh, Behmanesh and Khalili (2017), temperature and solar radiation are one of the main impact data on ET_0. Several studies have indicated the influence of lat, lon and alt variables on temperature and solar radiation (ALVARES et al., 2013b; OZGOREN; BILGILI; SAHIN, 2012) Thus, variations in lat, lon and alt may indirectly impact ET_0. This can explain these observed results.

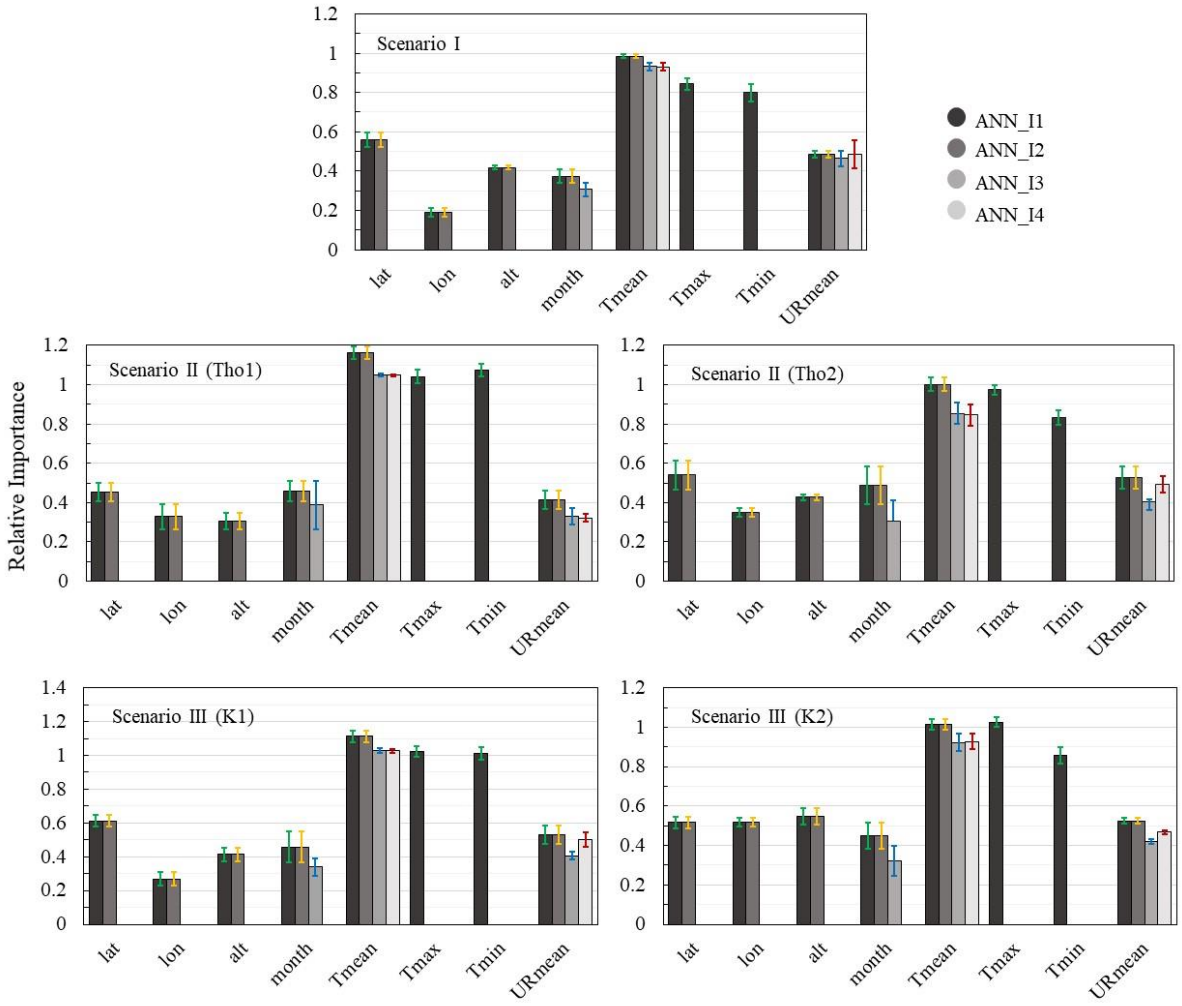
The division of the input data into two areas with climatic similarity aimed to increase the performance of the models. The division presented in SII and SIII managed to slightly increase the capacities of the models in relation to SI. However, this increase was only

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 between input data; however, these models have a greater ability to generalize (SHIRI et al., 2014). According to Ferreira et al. (2019), although the models developed locally perform better, these models may have low predictive capacity when used in other regions, since they can be highly specific to the location.

The plot, show in Figure 6, 7 and 8, indicates the importance of each input variable in the response variable of the evaluated algorithms. The WEKA was used to select attributes. Attributes were selected using the "*ClassifierAttributeEval*" tool associated with "*Ranker*" method. These tools rank attributes by their individual evaluations. The correlation coefficient was the measure used to evaluate the performance of attribute combinations in the *Ranker* configuration. Using this the ranking method of WEKA in a similar way, Yadav, Malik and Chandel (2014) verified the importance of each input variable in the solar radiation prediction. The rank of each input variable helped to build more efficient models. Wang et al. (2019b) also verified the importance of the meteorological data, but by different methodologic show in present study. As reported by Yin, Wu and Dai, (2010), it is necessary to analyze the relative importance of meteorological variables is to be understanding of the impact of global climate change on evapotranspiration, and for water resource management.

Different ANN settings were used for each input data (Table 2). These ANN settings resulted in different weights for each input attribute (Figure 6). However, a similar behavior was observed between the different configurations. In all scenarios, Tmean, Tmax and Tmin had a greater weight in the estimate. In SIII K2, the relative importance of Tmax surpassed Tmed (Figure 5e). This result may explain the decrease in ANN's performance in this scenario when Tmax and Tmin are removed (Table 5). 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 the ANN performance when compared to the removal of the variables lal, lon and alt.

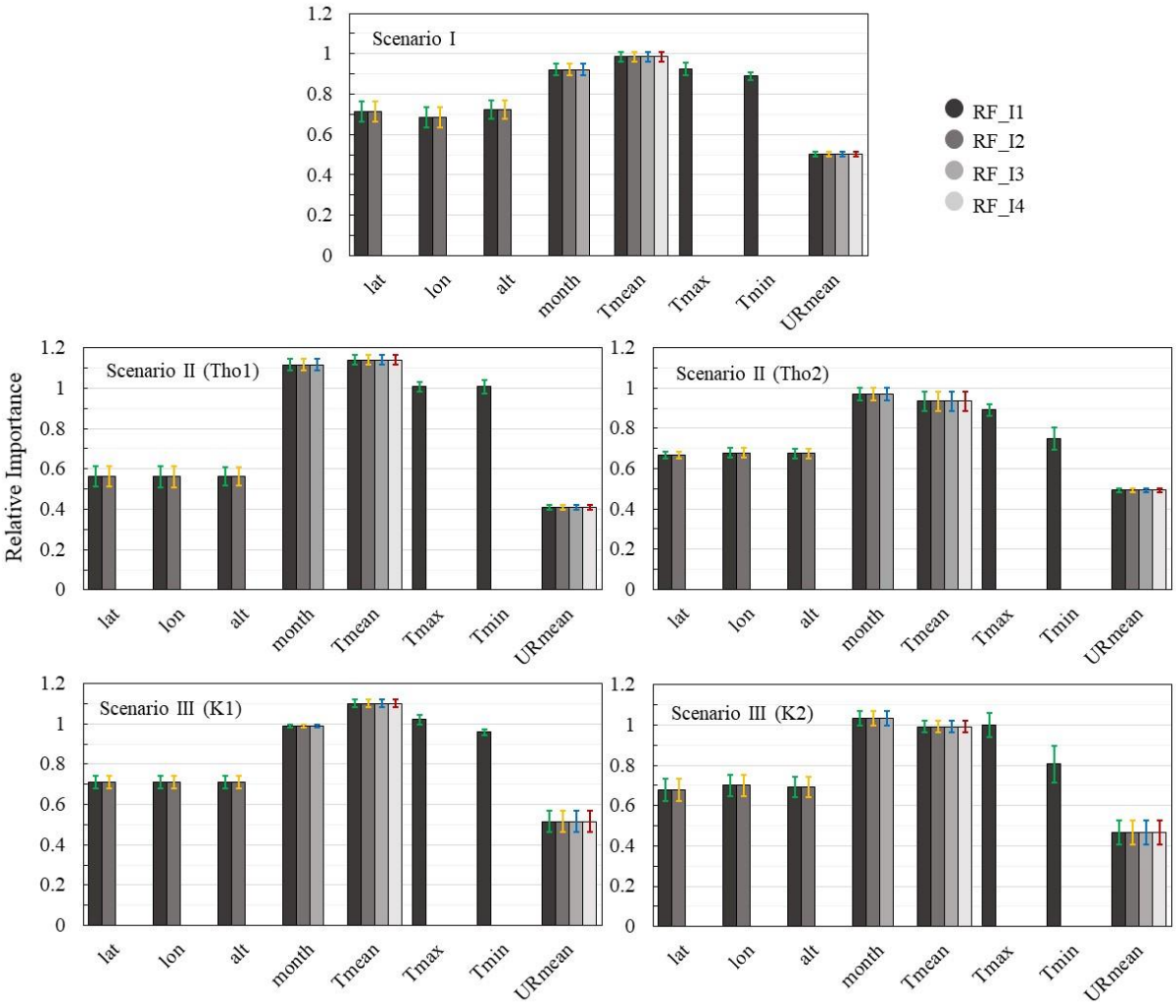
Figure 5. Importance of the input variable of ANN models



Source: The Authors (2021)

Ranked value of each input variable of RF is show in Figure 6. The Tmean and month variable had a higher weight in the ET_0 estimate. In SII Tho1 and SIII K2 (Figure 6c and 6e), the month variable was more important than Tmean variable. This result may explain the drop in the RF model 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 (Table 3, 4 and 5) and discussed previously.

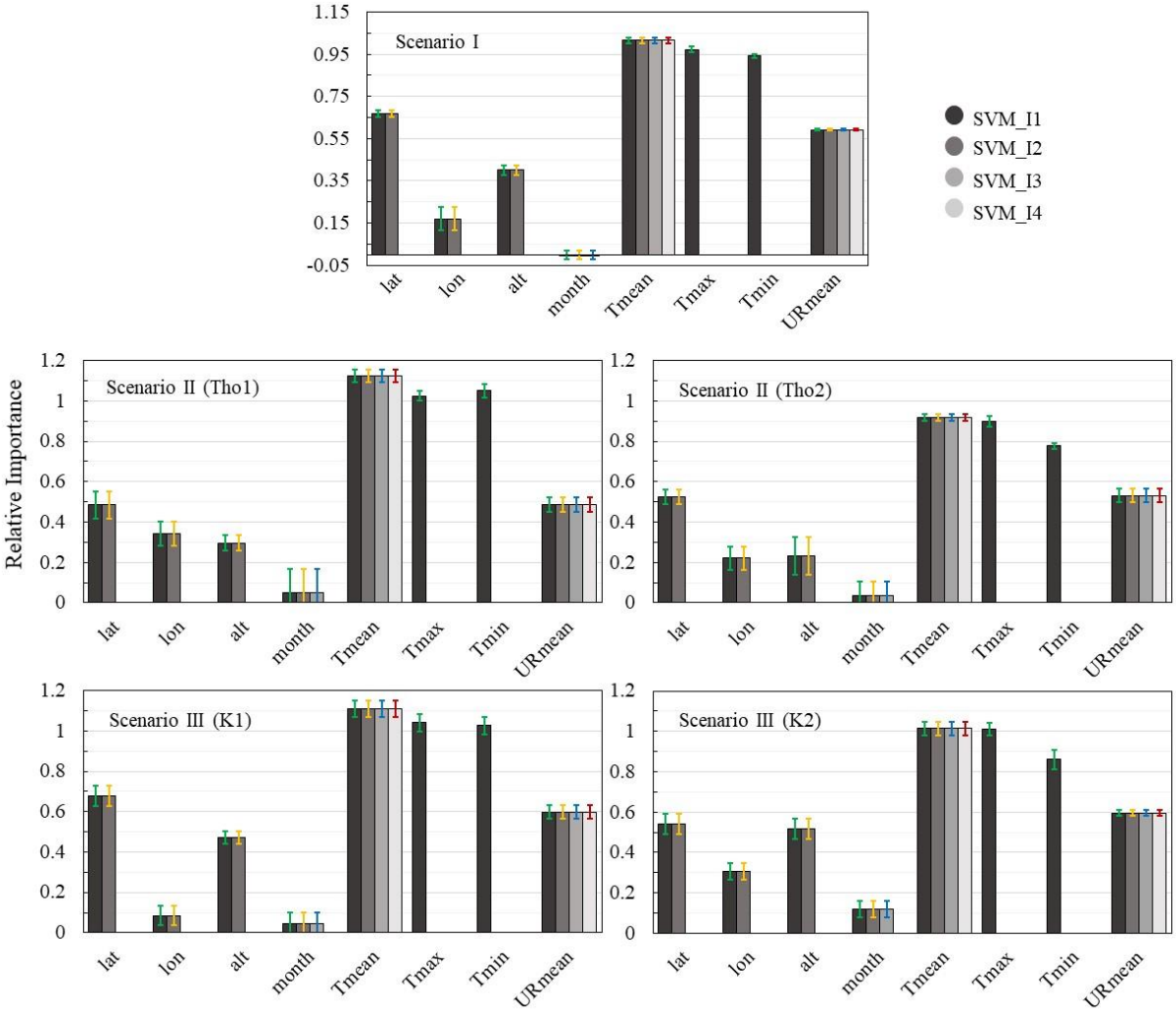
Figure 6. Importance of the input variable of RF models



Source: The Authors (2021)

The relative importance of each input variable of SVM is show in Figure 7. Tmean, Tmax and Tmin variable 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 of the SVM model (Table 3, 4 and 5), there was no significant variation in performance when the month variable is removed. Both results make it possible to state that, for this region, the month variable does not contribute to the performance of the SVM model.

Figure 7. Importance of the input variable of SVM models



Source: The Authors (2021)

Although each model has a different pattern in the ranking of the input variables (Figure 5, 6 and 7), air temperature was the most important attribute. The observed correlation between temperature and ET_0 (Figure 4) may explain the importance of temperature in the estimate. This behavior was not observed in SIII K2 and SII Tho2. However, in these scenarios, no significant difference was observed between the month and Tmean variables. Wang et al. (2019b) observed the rank of importance of meteorological variables based on the RF method. The three most important variables were: insolation (n), Tmax and RH. 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 experienced. In this way, new research or applications can base on these results and choose the best method that suits the conditions of the input data. However, it is recommended that the models be previously experimented with 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, the removal of the variables Tmax and Tmin increased the predictive capacity, although these variables have shown high relative importance.

It is important to note that the month variable was highly important in estimating through the RF. However, a low importance was observed when the SVM model was used. This variable was not correlated with ET_0 (Figure 4). These results reinforce the need for more techniques to select the meteorological variables used in the modeling. Linear regression alone is not sufficient to identify the relevance of the input data. Furthermore, different models may present different behaviors regarding the classification of the importance of the input variable and still present satisfactory results.

The adjusted MLR method coefficients are shown in Table 4. Different from the attribute importance assessment of ANN, RF and SVM, the attribute selection method was applied (M5 method), which indicates the importance of each input attribute in the generated model. It is observed that in some models the method used (M5 method) excluded the month variable. This behavior indicates a low importance of this variable in the MLR estimate. This result is similar to that observed in the analysis of the importance of the input variables in the SVM. The exclusion of lat and Tmax was also observed in some cases.

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	
		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_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.

Source: The Authors (2021)

The results presented revealed that, for locations in the Minas Gerais state, the 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 a 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 presents less computational effort. The models, although they have a

high predictive capacity, cannot be perfect. Other meteorological variables not considered as input data (e.g., radiation and wind speed) and other factors (e.g., data recorded in error) contributed to the decrease in the predictive capacity of these models.

No statistical method or machine learning 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 weather stations function continuously (ALVES et al., 2020). These 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 (BRUGNARO; BACHA, 2006) and the Minas Gerais state has the third largest Gross Domestic Product in Brazil of 2018 (IBGE, 2020). The results of these models assist in the management of irrigation, in climatic zoning, in 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.

CONCLUSION

The combination of input data I_6 (alt, lat, lon, month, Tmean and RH), in general, provided the best results in the ET_0 estimate between the evaluated models, so this combination is the recommended one. The RF and ANN models presented the highest predictive ability in the ETo estimate. Both models in best-case scenarios with input data I_6 or I_8 explains more than 96% of the variability of the variables estimated using the independent dataset. However, the RF model presented a small superiority when compared to the ANN model. Temperature was the input meteorological variable that presented the greatest relative importance. The month variable presented the greatest variation of importance in relation to the model used. The month variable presented median importance (ANN), high importance (RF) and low importance (SVM). Therefore, it is concluded that although the temperature is fundamental for the estimation of ET_0, other variables can present different levels of importance in the prediction of ET_0.

ACKNOWLEDGEMENTS

The authors express their gratitude to CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil) for their financial support and scholarships. This study was

financed in part by the CAPES – Finance Code 001; and Brazilian National Institute of Meteorology (INMET) for making the series of meteorological data available.

REFERENCE

ALLEN, R. G. et al. **Crop evapotranspiration -** guidelines for computing crop water requirements**,** Rome: FAO, 1998. 297p. (FAO Irrigation and drainage paper 56).

ALMOROX, J.; QUEJ, V. H.; MARTÍ, P. Global performance ranking of temperature-based approaches for evapotranspiration estimation considering Köppen climate classes. **Journal of Hydrology**, v. 528, p. 514–522, 2015.

ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013a.

ALVARES, C. A. et al. Modeling monthly mean air temperature for Brazil. **Theoretical and Applied Climatology**, v. 113, n. 3–4, p. 407–427, 2013b.

ALVES, M. P. A. et al. Reconstrução de dados e deteção de ondas de calor e de frio no Porto e concelhos vizinhos–Portugal. **Territorium**, n. 27 (II), p. 49–66, 2020.

BENALI, L. et al. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. **Renewable Energy**, v. 132, p. 871–884, 2019.

BOWDEN, G. J.; DANDY, G. C.; MAIER, H. R. Input determination for neural network models in water resources applications. Part 1—background and methodology. **Journal of Hydrology**, v. 301, n. 1–4, p. 75–92, 2005.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

BRUGNARO, R.; BACHA, C. J. C. Analysis of increased participation of agriculture in the Brazilian GDP from 1994 a 2004. In: CONGRESS OF THE EUROPEAN REGIONAL SCIENCE ASSOCIATION, **Anais**...2006.

CHANG, C.-C.; LIN, C.-J. LIBSVM: a library for support vector machines," 2001. Software available at http://www. csie. ntu. edu. tw/~ cjlin/libsvm. 2012.

COMPAHIA BRASILEIRA DE ABASTECIMENTO - CONAB. **Acompanhamento da safra brasileira de café**. Safra 2020 - Primeiro Levantamento. 6:1-62. 2020. Available in: <https://www.conab.gov.br/info-agro/safras/cafe>. Access in: April, 28, 2020.

EWAID, S. H.; ABED, S. A.; AL-ANSARI, N. Crop water requirements and irrigation schedules for some major crops in Southern Iraq. **Water**, v. 11, n. 4, p. 756, 2019.

FAUSETT, L. **Fundamentals of neural networks: architectures, algorithms, and**

applications. [s.l.] Prentice-Hall, Inc., 1994.

FENG, Q.; WEN, X.; LI, J. Wavelet analysis-support vector machine coupled models for monthly rainfall forecasting in arid regions. **Water resources management**, v. 29, n. 4, p. 1049–1065, 2015.

FENG, Y. et al. Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. **Agricultural Water Management**, v. 193, p. 163–173, 2017.

FERREIRA, L. B. et al. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM–A new approach. **Journal of hydrology**, v. 572, p. 556–570, 2019.

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. **Agricultural Water Management**, v. 234, p. 106113, 2020.

HUANG, G. et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. **Journal of Hydrology**, v. 574, p. 1029–1041, 2019.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Panorama.** 2020a. Available in:< https://cidades.ibge.gov.br/brasil/mg > Access in: April, 28, 2021.

KOPPEN, W. das. Das geographische system der klimat. **Handbuch der klimatologie**, p. 46, 1936.

KUMAR, M. et al. Estimating evapotranspiration using artificial neural network. **Journal of Irrigation and Drainage Engineering**, v. 128, n. 4, p. 224–233, 2002.

MAIER, H. R.; DANDY, G. C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. **Environmental modelling & software**, v. 15, n. 1, p. 101–124, 2000.

MALIK, A. et al. The viability of co-active fuzzy inference system model for monthly reference evapotranspiration estimation: case study of Uttarakhand State. **Hydrology Research**, v. 50, n. 6, p. 1623–1644, 2019.

MARTÍ, P.; GONZÁLEZ-ALTOZANO, P.; GASQUE, M. Reference evapotranspiration estimation without local climatic data. **Irrigation Science**, v. 29, n. 6, p. 479–495, 2011.

MEHDIZADEH, S.; BEHMANESH, J.; KHALILI, K. Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. **Computers and electronics in agriculture**, v. 139, p. 103–114, 2017.

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**, v. 19, n. 2, p. 392–403, 2019.

MONTEITH, J. L. Evaporation and environment. In: Symposia of the society for experimental biology, **Anais**...Cambridge University Press (CUP) Cambridge, 1965.

NOURANI, V.; ELKIRAN, G.; ABDULLAHI, J. Multi-station artificial intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements. **Journal of Hydrology**, v. 577, p. 123958, 2019.

OZGOREN, M.; BILGILI, M.; SAHIN, B. Estimation of global solar radiation using ANN over Turkey. **Expert Systems with Applications**, v. 39, n. 5, p. 5043–5051, 2012
.
PENMAN, H. L. Natural evaporation from open water, bare soil and grass. **Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences**, v. 193, n. 1032, p. 120–145, 1948.

PISNER, D. A.; SCHNYER, D. M. Support vector machine. In: **Machine Learning**. [s.l.] Elsevier, 2020. p. 101–121.

SAMADIANFARD, S. et al. Wavelet neural networks and gene expression programming models to predict short-term soil temperature at different depths. **Soil and Tillage Research**, v. 175, p. 37–50, 2018.

SATTARI, M. T. et al. Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. **Hydrology and Earth System Sciences**, v. 25, n. 2, p. 603–618, 2021.

SHIRI, J. et al. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. **Computers and Electronics in Agriculture**, v. 108, p. 230–241, 2014.

THORNTHWAITE, C. W. An approach toward a rational classification of climate. **Geographical review**, v. 38, n. 1, p. 55–94, 1948.

VAPNIK, V. **The nature of statistical learning theory**. [s.l.] Springer science & business media,p. 188, 2013.

WANG, H. et al. Intelligent identification of maceral components of coal based on image segmentation and classification. **Applied Sciences (Switzerland)**, v. 9, n. 16, p. 1–15, 2019a.

WANG, S. et al. Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China. **Agricultural Water Management**, v. 221, p. 220–230, 2019b.

WEN, X. et al. Support-vector-machine-based models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions. **Water Resources Management**, v. 29, n. 9, p. 3195–3209, 2015.

XIANG, K. et al. Similarity and difference of potential evapotranspiration and reference crop

evapotranspiration–a review. **Agricultural Water Management**, v. 232, p. 106043, 2020.

XU, Y.; KNUDBY, A.; HO, H. C. Estimating daily maximum air temperature from MODIS in British Columbia, Canada. **International Journal of Remote Sensing**, v. 35, n. 24, p. 8108–8121, 2014.

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. **Renewable and Sustainable Energy Reviews**, v. 31, p. 509–519, 2014.

YANG, Q. et al. Sensitivity of potential evapotranspiration estimation to the Thornthwaite and Penman–Monteith methods in the study of global drylands. **Advances in Atmospheric Sciences**, v. 34, n. 12, p. 1381–1394, 2017.

YASSIN, M. A.; ALAZBA, A. A.; MATTAR, M. A. Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. **Agricultural Water Management**, v. 163, p. 110–124, 2016.

YIN, Y.; WU, S.; DAI, E. Determining factors in potential evapotranspiration changes over China in the period 1971–2008. **Chinese Science Bulletin**, v. 55, n. 29, p. 3329–3337, 2010.

YIN, Z. et al. Integrating genetic algorithm and support vector machine for modeling daily reference evapotranspiration in a semi-arid mountain area. **Hydrology Research**, v. 48, n. 5, p. 1177–1191, 2017.

ZHOU, X. et al. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. **The Crop Journal**, v. 4, n. 3, p. 212–219, 2016.