Ambiente & Água - An Interdisciplinary Journal of Applied Science
ISSN 1980-993X – doi:10.4136/1980-993X
www.ambi-agua.net
E-mail: ambi.agua@gmail.com

Climate-related variables may not improve monthly scale rainfall predictions by artificial neural networks for the metropolitan region of Belo Horizonte, Brazil

ARTICLES doi:10.4136/ambi-agua.2879

Received: 02 Aug. 2022; Accepted: 05 Nov. 2022

**Mateus Alexandre da Silva[1*]; Marina Neves Merlo[1];
Michael Silveira Thebaldi[1]; Danton Diego Ferreira[2];
Felipe Schwerz[3]; Fábio Ponciano de Deus[1]**

[1]Departamento de Recursos Hídricos. Universidade Federal de Lavras (UFLA), Trevo Rotatório Professor Edmir Sá Santos, s/n, CEP: 37203-202, Lavras, MG, Brazil. E-mail: marinanevesmerlo@gmail.com, michael.thebaldi@ufla.br, fabio.ponciano@ufla.br
[2]Departamento de Automática. Universidade Federal de Lavras (UFLA), Trevo Rotatório Professor Edmir Sá Santos, s/n, CEP: 37203-202, Lavras, MG, Brazil. E-mail: danton@ufla.br
[3]Departamento de Engenharia Agrícola. Universidade Federal de Lavras (UFLA), Trevo Rotatório Professor Edmir Sá Santos, s/n, CEP: 37203-202, Lavras, MG, Brazil. E-mail: felipe.schwerz@ufla.br
[*]Corresponding author. E-mail: mateus4lexandre@outlook.com

ABSTRACT

Artificial neural networks (ANNs) may experience problems due to insufficient or uninformative predictors, and these problems are common for complex predictions such as those for rainfall. However, some studies point to the use of climate variables and anomalies as predictors to make the forecast more accurate. This research aimed to predict the monthly rainfall, one month in advance, in four municipalities of the metropolitan region of Belo Horizonte using an ANN trained with different climate variables; additionally, it aimed to indicate the suitability of such variables as inputs to these models. The models were developed using the MATLAB® software Version R2011a using the NNTOOL toolbox. The ANNs were trained by the multilayer perceptron architecture and the feedforward and backpropagation algorithm using two combinations of input data, with two and six variables, and one combination of input data with the three most correlated variables to observed rainfall from 1970 to 1999 to predict the rainfall from 2000 to 2009. The climate variable most correlated with the rainfall of the following month was the average compensated temperature. Even when using the variables most correlated with precipitation as predictors ($0.66 \leq n_t$ index ≤ 1.26), there was no notable improvement in the predictive capacity of the models when compared to those that did not use climate variables as predictors ($0.55 \leq n_t$ index ≤ 0.80).

Keywords: artificial intelligence, ENSO, hydrological modelling.

Variáveis relacionadas ao clima podem não melhorar previsões de precipitação pluvial em escala mensal realizadas por redes neurais artificiais para a região metropolitana de Belo Horizonte, Brasil

RESUMO

As redes neurais artificiais (RNAs) podem apresentar problemas devido a preditores insuficientes ou não informativos, o que é comum para previsões complexas, como as de precipitação pluvial. No entanto, alguns estudos apontam para o uso de variáveis e anomalias climáticas como preditores para tornar a previsão mais precisa. Esta pesquisa teve como objetivo prever a precipitação mensal, com um mês de antecedência, em quatro municípios da região metropolitana de Belo Horizonte utilizando uma RNA treinada com diferentes variáveis climáticas; além disso, buscou indicar a adequação de tais variáveis como entrada para esses modelos. Os modelos foram desenvolvidos por meio do software MATLAB® versão R2011a utilizando a *toolbox* NNTOOL. As RNAs foram treinadas pela arquitetura *multilayer perceptron* e pelo algoritmo *feedforward* e *backpropagation* usando duas combinações de dados de entrada, com duas e seis variáveis, e uma combinação de dados de entrada com as três variáveis mais correlacionadas com precipitação observada de 1970 a 1999 para prever a precipitação de 2000 a 2009. A variável climática mais correlacionada com a precipitação do mês seguinte foi a temperatura média compensada. Mesmo utilizando as variáveis mais correlacionadas com a precipitação como preditores ($0,66 \leq$ índice $n_t \leq 1,26$), não houve melhora significativa na capacidade preditiva dos modelos quando comparado aos que não utilizaram variáveis climáticas como preditores ($0,55 \leq$ índice $n_t \leq 0,80$).

Palavras-chave: ENSO, inteligência artificial, modelagem hidrológica.

1. INTRODUCTION

The city of Belo Horizonte experienced great destruction caused by rainfall events in January 2020, which was the wettest month since the beginning of climatological measurement in the city. Specifically, 966.6 mm of rainfall was accumulated, reaching 101.6 mm in a period of three hours in certain places (G1 MINAS, 2020; INMET, 2022).

To try to avoid or even mitigate these types of damage, rainfall forecasting is a very important tool that can save lives and property and ensure economic activities (Lee *et al.*, 2018). However, in the hydrological cycle, rainfall is one of the most complex variables to understand and model due to its high temporal and spatial variability (Tian *et al.*, 2017; Tauro *et al.*, 2018). Nevertheless, this phenomenon is influenced by several factors, such as climate variables (air temperature, relative humidity, insolation, wind speed, among others) and climate anomalies (Mawonike and Mandonga, 2017; Peres and Maier, 2022).

Modelling that aims to forecast rainfall using only rainfall data itself is beneficial only when climate variables such as air temperature, wind speed, and relative humidity are not available or when a simple model is seeked in relation to the input data, and still, many researchers accept climate anomalies, such as the El Nino Southern Oscillation (ENSO), as a good predictor for time-series events, such as rainfall (Aksoy and Dahamsheh, 2009; Hossain *et al.*, 2018). Therefore, the development of models that allow the addition of variables that are related to rainfall behavior may be one way to circumvent the lack of forecast accuracy. Despite the complexity involved in predicting rainfall, artificial neural networks (ANNs) have proven to be able to predict time series (Torres *et al.*, 2021), such as rainfall.

ANNs are based on the functioning of the human brain, possessing the ability to acquire learning through input data. An ANN is formed by one or more layers, which are composed of one or more interconnected neurons, in which the processing of input signals (data) is performed using the weights of the connections between neurons. Each neuron communicates with the neurons of the next layer until the output layer is reached and a response is issued by the model. After that, the weight values are modified so that the calculated output matches the actual output as accurately as possible (Gonzales-Fernandes *et al.*, 2019).

Some researchers have tried to predict rainfall using ANNs trained by climate variables, such as Esteves *et al.* (2019), who used the rainfall itself and the mean air temperature to train the model, aiming to predict the rainfall occurrence in the central-south region of Brazil. These authors reported that it was possible to reach an average accuracy, but the process may be different in other Brazilian regions due to continental dimensions with contrasting climates. Information about climate anomalies, such as ENSO, has also been used as predictors by some researchers, as in Hossain *et al.* (2020), where information about ENSO was used in rainfall prediction for Australia. The aforementioned authors explained that an acceptable prediction was not reached because rainfall is the result not only of regional factors, such as ENSO, but also of local ones, and they recommended the addition of such factors for future research. However, training an ANN is not just about adding a large number of predictors, as explained by May *et al.* (2011), who cited that the use of variables that have little or no predictive power affects the complexity of the model and hinders the learning of the ANN. Thus, according to the aforementioned authors, a careful preselection of variables to compose the models' input is indicated.

Given the need for forecasting, the complexity of rainfall and the potential of climate variables and anomalies as predictors, this study aimed to analyze the suitability of climate variables and information about the occurrence of the ENSO climate anomaly as rainfall predictors for four municipalities located in the metropolitan region of Belo Horizonte, Minas Gerais, Brazil, using an ANN.

2. MATERIALS AND METHODS

Because precipitation varies both temporally and spatially, the analysis was performed on a regional scale (Hossain *et al.*, 2020) and comprised four municipalities in the metropolitan mesoregion of Belo Horizonte in the state of Minas Gerais, Brazil. For municipality selection, the following criteria were established: existence of monthly data of total rainfall, mean compensated temperature, mean relative humidity, and mean wind speed in the period between 1970 and 2009, with a maximum data gap percentage of 30%; belonging to the same mesoregion; and a maximum altitude difference of 300 m between them.

The choice of time interval used in the study was made by analyzing the time series of rainfall in the municipality of Belo Horizonte and opting for a period of 40 years with the least amount of data gaps. It should be stressed that the procedures adopted in this study can be reproduced for time intervals different from the one used. The identification of the municipalities, as well as some of their characteristics, are indicated in Table 1.

Table 1. Identification and characteristics of the municipalities covered in the study.

Climatological station (city)	Latitude	Longitude	Köppen climate classification - (Martins *et al.*, 2018)	Altitude (m)	Average annual rainfall (mm)[a]
Belo Horizonte	-19.934382	-43.952292	Aw	915.17	1544.64
C. do Mato Dentro	-19.020355	-43.433948	Aw	663.02	1283.50
Florestal	-19.885422	-44.416889	Cwa	753.51	1399.29
Sete Lagoas	-19.48454	-44.173798	Aw	753.68	1325.69

a = Period of the training data (1970 – 1999); Aw = tropical with winter drought; Cwa = subtropical with dry winter and hot summer.

Figure 1 shows the geographical location of the municipalities covered by the study.

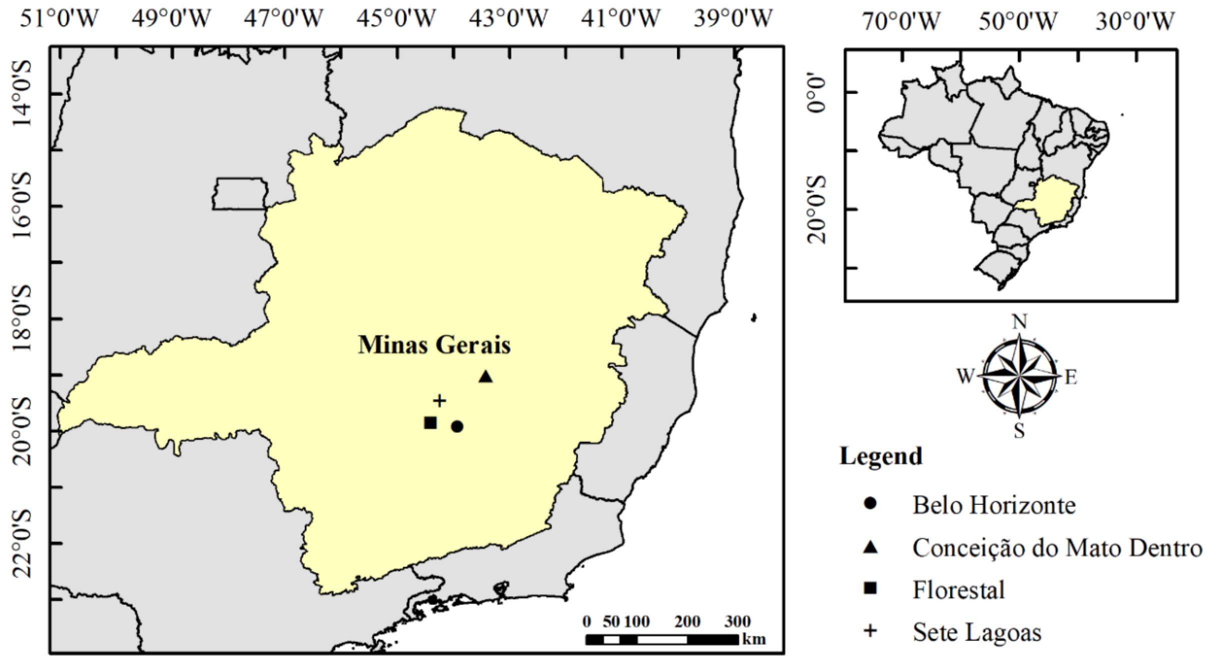

Figure 1. Geographic location of the municipalities covered by the study.

2.1. Data used in training and validation of the ANN

Aiming to represent the seasonality inherent to the rainfall time series, the sequence number corresponding to the month (one to 12) was used to compose the database for training and validation of the ANN as were the data for the historical monthly total rainfall, average compensated temperature, average relative humidity and average wind speed for the years 1970 to 2009 obtained from the Database of the National Institute of Meteorology – BDMEP (INMET, 2022).

To provide data on the occurrence of the climate phenomena El Niño and La Niña, the multivariate ENSO index (MEI) was used for the years 1970 to 2009, with bimonthly records obtained from the National Oceanic and Atmospheric Administration platform (NOAA, 2020). In this case, the lowest value (one) indicates stronger cases of La Niña, while the highest value (69) indicates stronger cases of El Niño.

2.2. Data pre-processing

To achieve a single monthly value for the MEI, a weighted average between overlapping months (December - January; January - February; (...); November - December) was calculated using the number of days in each month.

The database was divided into two intervals: training (1970-1999) and validation (2000-2009), the latter being variable due to the availability of data from each climatological station. To ensure that the characteristics of the rainfall time series did not change, no procedure for gap filling was conducted.

To verify the homogeneity and consistency of the used monthly rainfall data, a double mass curve was elaborated. This procedure was adopted to verify if the rainfall values for the training period were well measured, since errors can occur, for example, due to changing of the device installation location. This verification was performed so that the monthly rainfall observed at each station was validated (y-axis), while the average rainfall of the other stations was considered as the reference of observed monthly rainfall (x-axis).

To verify the existence of a tendency in the time series of total monthly rainfall for the training period (1970 to 1999), the Mann-Kendall test was performed, admitting a significance level of 5% (p-value < 0.05). For the ANN training, rainfall at time "t+1" was set as the target, and three combinations of input data were also set as follows:

- C1 - sequential number corresponding to the month and total rainfall, both at time "t".
- C2 - sequential number corresponding to the month, total rainfall, compensated average temperature, average relative humidity, average wind speed, and MEI at time "t".
- C3 - three variables whose time series obtained the highest Pearson's linear correlation coefficient in relation to the target's time series at time "t".

The values of monthly data for the ANN training varied with its availability from each climatological station for the training period. For the cities of Belo Horizonte, Conceição do Mato Dentro, Florestal, and Sete Lagoas, respectively, were used 310-, 304-, 332-, and 255-months data for combination "C1", and 307-, 266-, 315-, and 248-months data for combination "C2" and "C3".

It should be stressed that rainfall at time “t” was the current observed rainfall. Due to the different measurement units inherent in the input data, the values were normalized using Equation 1.

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

Where:

z = the normalized value;

x = the value to be normalized;

max(x) = the maximum value among the values to be normalized; and

min(x) = the minimum value among the values to be normalized.

2.3. ANN training and rainfall prediction for the validation period

The ANN was developed in MATLAB® software Version R2011a using the NNTOOL toolbox. To train the ANN with different combinations of data, the multilayer perceptron architecture was used with the feedforward backpropagation algorithm widely cited in the literature due to its excellent results in predicting series of monthly rainfall (Amiri *et al.*, 2018; Mohammadpour *et al.*, 2018), and the Levenberg–Marquardt training function was also used (Levenberg, 1944). The configurations of the ANN were experimentally defined by "trial and error", and the best performing configuration was selected. The number of middle layers ranged from two to four and the number of neurons, in each, from two to 10. The number of learning cycles was fixed at 1000.

For ANN training that used the input data combinations "C1" and "C3", we used two hidden layers with four neurons, with the sigmoid tangent hyperbolic and log-sigmoid transfer functions. An output layer with one neuron and a linear transfer function was used as well. For the ANN training that used the "C2" input data combination, two hidden layers with six neurons each were used with the log-sigmoid and sigmoid-tangent hyperbolic transfer functions. In addition, an output layer with one neuron and the linear transfer function was used for C2.

During the ANN training stage, each model was trained 10 times with different initial training weights, keeping the best result and discarding the others. The predictions that presented a value less than 0 were converged to 0 since the rainfall value cannot be negative.

2.4. Validation of forecasted rainfall

For the validation of the rainfall predicted by ANN, the following statistical indicators were calculated as in Silva *et al.* (2021): Pearson's linear correlation coefficient (r), mean absolute error (MAE) and bias (b). In order to identify the season in which the largest errors occurred, besides their tendency to underestimate or overestimate, the mean absolute error and bias were calculated separately for the rainy season (October to March) and dry season (April to September).

The amount of monthly data for validation varied with the availability of data from each climatological station in the validation period, being for the municipalities of Belo Horizonte, Conceição do Mato Dentro, Florestal, and Sete Lagoas, respectively, for data combination "C1", 120-, 199-, 109-, and 120-months data, and for data combination "C2" and "C3", 120-, 118-, 29-, and 120-months data.

It is possible to classify the degree of correlation between two variables into three classes by the Pearson's linear correlation coefficient (Schober *et al.*, 2018): $0 \leq r < 0.10$ negligible correlation; $0.10 \leq r < 0.40$ weak correlation; $0.40 \leq r < 0.70$ moderate correlation; $0.70 \leq r < 0.90$ strong correlation; and $0.90 \leq r < 1.00$ very strong correlation.

To evaluate the model performance, the n_t index (Equation 2) proposed by Ritter and Muñoz-Carpena (2013) was used. The n_t index suggests that model efficiency should be considered satisfactory when the error is "small", taking into account the width of the data range covered by the calculated values. This index ensures that a model that predicts rainfall with a small error value within a small data range is not considered better than another model that predicts with a larger error value, but within a larger data range.

$$n_t = \frac{SD}{RMSE} - 1 \tag{2}$$

In this way, the efficiency of the model is evaluated depending on the number of times (n_t) that the variability of the observations is greater than the mean error of the model. To this end, the mean error is represented by the square root of the root mean square error (RMSE) and the variability of the observed data by the population standard deviation (SD).

Ritter and Muñoz-Carpena (2013) also defined four performance classes based on the n_t index, with $n_t < 0.7$ unsatisfactory, $0.7 \leq n_t < 1.2$ acceptable, $1.2 \leq n_t < 2.2$ good and $2.2 \leq n_t$ very good.

In order to verify possible significant differences between the observed and predicted time series for the best performing combination, the Mann-Whitney test was performed, assuming a significance level of 5% (p-value < 0.05).

3. RESULTS AND DISCUSSION

3.1. Characterization of monthly rainfall data

The hydrological homogeneity of the region where the climatological stations were inserted, as well as the consistency of the data from the time series of monthly rainfall, were proven by the high coefficient of determination values (R^2 = 0.9976-0.9995), which were obtained by fitting the linear trend line to the double mass curve. Thus, it was affirmed that the hydrological behavior for the analyzed climatological stations was similar and eliminated the occurrence of errors of transcription of the data observed in the field or alteration of the angular coefficient of the straight line.

Table 2 shows the p-values and "τ" obtained by applying the Mann-Kendall test at the 5% significance level to the time series of total monthly rainfall of the climatological stations addressed in this study during the training period.

Table 2. p-values and "τ" obtained for the Mann-Kendall test applied to the time series of total monthly rainfall of studied locations in the training period.

Time Series (month)	p-value (τ)			
	Climatological stations			
	Belo Horizonte	C. do Mato Dentro	Florestal	Sete Lagoas
January	0.066 (0.223)	0.196 (0.152)	0.133 (0.174)	0.064 (0.241)
February	0.443 (0.092)	0.657 (0.055)	0.454 (0.089)	0.669 (0.057)
March	0.798 (0.032)	0.766 (0.037)	0.910 (0.015)	0.199 (0.168)
April	0.744 (-0.040)	0.755 (0.039)	0.989 (0.003)	1.000 (0.002)
May	0.306 (-0.123)	0.109 (-0.192)	0.744 (-0.040)	0.301 (-0.136)
June	0.274 (-0.135)	0.926 (0.013)	0.527 (-0.080)	0.427 (-0.108)
July	0.155 (-0.177)	0.766 (-0.038)	0.160 (-0.179)	0.229 (-0.163)
August	0.382 (0.108)	0.431 (0.098)	0.945 (-0.010)	0.971 (0.007)
September	0.196 (0.155)	0.966 (-0.007)	0.754 (-0.038)	1.000 (0.000)
October	0.125 (-0.183)	0.002 (-0.378)[a]	0.007 (-0.317)[a]	0.019 (-0.299)[a]
November	0.532 (-0.076)	0.260 (-0.137)	0.264 (-0.132)	0.110 (-0.204)
December	0.132 (0.180)	0.001 (0.401)[a]	0.969 (0.006)	0.062 (0.239)

a = significant trend by the Mann-Kendall test using a significance level of 5%.

Table 2 shows that there was a significant trend (p-value < 0.05) of decrease for the total rainfall values in October (τ < 0) for the time series of the Conceição do Mato Dentro, Florestal and Sete Lagoas climatological stations. In contrast, there was only a significant trend (p-value > 0) of increase in total rainfall values (τ > 0) in December for the time series of the Conceição do Mato Dentro climatological station. For the other months of the time series of the climatological stations analyzed, there was no significant trend of increase or decrease in the total rainfall value.

3.2. Model performance using "C1" and "C2" input data combinations in training

Table 3 shows the values of the statistical indicators calculated for the validation of the rainfall series predicted by means of the ANN, using the input data combination "C1" in the training.

Table 3. Pearson's linear correlation coefficient (r), n_t index, mean absolute error in the dry (MAE_d) and rainy (MAE_r) seasons and bias in the dry (b_d) and rainy (b_r) seasons, calculated for the validation of the rainfall series predicted by means of the ANN used for training the input data combination "C1" and "C2" (between parentheses).

Climatological station	r	MAE_d (mm)	MAE_r (mm)	b_d (mm)	b_r (mm)	n_t
Belo Horizonte	0.84 (0.83)	23.85 (25.57)	86.23 (90.75)	2.94 (-3.28)	-23.50 (-14.61)	0.80 (0.80)
C. do Mato Dentro	0.77 (0.75)	25.90 (25.44)	92.33 (96.27)	10.04 (9.34)	-17.81 (-3.11)	0.55 (0.51)
Florestal	0.78 (0.84)	23.64 (23.85)	76.79 (56.10)	13.87 (21.32)	4.65 (-31.00)	0.60 (0.83)
Sete Lagoas	0.85 (0.78)	21.64 (40.52)	76.43 (72.87)	11.66 (26.37)	-24.66 (-17.12)	0.80 (0.60)

According to the classification proposed by Schober *et al.* (2018), all the predicted rainfall time series obtained a strong correlation (Table 3) with the observed rainfall time series. Such classification indicates linearity between the increase in values of the observed and predicted time series, suggesting that if the observed data are above average, the predicted data will also be above average (Martins, 2014). Thus, it was noted that the models were able to successfully predict the seasonality present in the time series data using both data combinations, identifying the months with higher and lower rainfall rates. However, with the exception of the climatological station in the Florestal municipality, the correlation values obtained using the "C2" input data combination decreased in relation to the values obtained for training the ANN using the "C1" input data combination.

The highest mean absolute error and bias values for the dry season were obtained using the "C2" input data combination (Table 3). Comparing the mean absolute error values to the mean rainfall of the dry season, the values could be regarded as high, but there was a sharp reduction in the value of the mean rainfall caused by the months with low or no rainfall. Thus, the errors of the dry season were less relevant than the rainfall values of each month of the dry season. The bias for the same period showed that with the exception of the model developed for the Belo Horizonte climatological station using the "C2" input dataset, the models overestimated rainfall, reinforcing the idea that the values obtained for the mean absolute error for the dry season were influenced by the drier months of the dry season.

The mean absolute error values for the wet season were in the range of 76.43 to 92.33 mm using the "C1" input data combination and 56.10 to 96.27 mm using the "C2" input data combination, and there were high values even in the wettest months. Additionally, the highest mean absolute error values for the wet season were presented by the combination of input data "C2". Among the bias values, there was only one positive value obtained for the climatological station in the municipality of Florestal when using the combination of input data "C1" in training the ANN. This fact indicated that the models underestimated the rainy season rainfall, and the lack of accuracy of the ANN in predicting the wettest months was a possible factor for the increased error and bias.

Following the n_t index values using the input data combination "C1", for the climatological stations in the Conceição do Mato Dentro and Florestal municipalities, the model was classified as unsatisfactory, and for the climatological stations in the Belo Horizonte and Sete Lagoas municipalities, it was classified as acceptable (Ritter and Muñoz-Carpena, 2013). Using the combination of input data "C2", the values of the n_t index were for the climatological stations in the Conceição do Mato Dentro and Sete Lagoas municipalities, the model was classified as unsatisfactory and for the climatological stations in the Belo Horizonte and Florestal municipalities, the model was acceptable (Ritter and Muñoz-Carpena, 2013).

Similar to the result obtained by Pearson's linear correlation coefficient, the n_t index indicated a regression in the model performance of the climatological stations of Conceição do Mato Dentro and Sete Lagoas using the combination of input data "C2" in relation to the use of the combination of input data "C1" for training the ANN. For the climate stations in the municipalities of Florestal and Belo Horizonte, there was an increase in the index and a stable index, respectively.

Through the joint analysis of the results shown in Table 3, it is noted, in general, that the addition of the climatological variables did not change notably or impair the performance of the models. Although the component variables of the "C2" input data combination were used by meteorologists to feed models that predicted the climate and its variability (NOAA, 2011), they could present changes in their behavior in short periods of time and were more useful for forecasting on smaller temporal scales, such as the hourly and daily scales.

As an example, Martins *et al.* (2019) explained that the relative humidity reached higher percentages at night. Thus, on scales larger than the hourly scale, such as the monthly scale, these values were reduced because they were measured as the average of the period.

Such behavior of the models was also explained by May *et al.* (2011). The authors presented that the ANN experienced problems of under-specification due to the choice of insufficient or uninformative input variables, or even by over-specification due to the use of uninformative or even redundant variables. These problems can affect the model complexity, learning difficulty, and ANN performance. The authors added that to train an ANN, it is necessary to select the input variables, and one of the widely used methods is to classify the variable based on Pearson's linear correlation coefficient, performing the selection in descending order of classification. Other methods can also be used to detect different types of relationships between predictor and predicted variables, aiming to exclude less relevant ones, such as: forward selection, stepwise regression, minimum entropy and partial mutual information (May *et al.*, 2011).

3.3. Variable selection used to compose the "C3" input data combination

To select the three variables most correlated with rainfall to compose the "C3" input data combination, Table 4 shows Pearson's linear correlation coefficient values obtained between the input variables used in input data combination "C2" and the target.

Table 4. Pearson's linear correlation coefficient calculated between the sequence number corresponding to the month (N), total rainfall (R), average compensated temperature (T), average relative humidity (H), average wind speed (W) and MEI, and the target for the choice of variables for the input data combination "C3".

Climatological station	N	R	T	H	W	MEI
Belo Horizonte	0.33	0.47	0.45	0.21	0.12	0.03
C. do Mato Dentro	0.38	0.44	0.49	-0.08	0.37	0.01
Florestal	0.38	0.43	0.51	0.09	0.08	0.04
Sete Lagoas	0.39	0.47	0.48	0.17	0.26	0.01
Average	0.37	0.45	0.48	0.10	0.21	0.02

By Table 4 analysis, it is possible to verify that both for the average and individually, the three variables that obtained higher Pearson's linear correlation coefficient values were mean compensated temperature, total rainfall and number corresponding to the month, respectively. In contrast, the average relative humidity, wind speed and MEI, in general, had considerably lower correlation values.

It is known that wet air favours the formation of rainfall, as in the occurrence of convective rainfall, and that wind speed affects the behavior of evapotranspiration. However, in this study, these variables did not have great predictive power for the rainfall of the following month. This fact can be explained in part by Mawonike and Mandonga (2017), who presented that variability in relative humidity affects the occurrence of rainfall, but the maximization of this effect occurs when relative humidity is above 80%. These values are observed less frequently on a monthly scale since days with low values of relative humidity reduce the average values.

For the wind-speed variable, the possible explanation for the low correlation, according to Alencar *et al.* (2015), is that the variation in the average wind speed at the monthly scale is relatively low, i.e., while there was a large discrepancy between the monthly rainfall values, the monthly wind speed values remained with limited variability.

The lowest Pearson's linear correlation coefficient corresponded to the MEI. This fact was supported by Grimm and Ferraz (1998) and Mota *et al.* (2019), who reported that the Southeast region of Brazil, where the study area is inserted, has a transitional character. Thus, the anomalies (El Niño and La Niña) can move more to the north or south from one event to another, making it possible to change the effects in relation to the same event that occurred previously, which does not occur for the extreme south of Brazil, for example.

Thus, for the input data combination "C3", the three variables that obtained the highest

Pearson's linear correlation coefficient value with the target were used, i.e., the number corresponding to the month, average compensated temperature, and total rainfall.

3.4. Model performance using the "C3" input data combination in training

The statistical indicators for the validation of the predicted rainfall time series used in the training of the ANN and the input data combination "C3" are shown in Table 5.

Table 5. Pearson's linear correlation coefficient (r), n_t index, mean absolute error in the dry (MAE_d) and rainy (MAE_r) seasons, bias in the dry (b_d) and rainy (b_r) seasons, and p-value of the Mann–Whitney test, calculated for the validation of the rainfall time series predicted by means of the ANN using the "C3" input data combination for training.

Climatological station	r	MAE_d (mm)	MAE_r (mm)	b_d (mm)	b_r (mm)	n_t	p-value
Belo Horizonte	0.85	21.33	83.92	2.21	-26.72	0.80	0.433
C. do Mato Dentro	0.80	27.80	83.45	13.24	-8.93	0.66	0.027[a]
Florestal	0.91	17.79	50.66	11.88	21.48	1.26	0.316
Sete Lagoas	0.85	20.42	71.46	7.91	-26.06	0.84	0.402

a = significant difference by the Mann–Whitney test at 5% statistical probability.

As shown in Table 5, the values of Pearson's linear correlation coefficient were between 0.80 and 0.91, which, according to Schober *et al.* (2018), indicated a strong correlation between the observed and predicted rainfall time series for the municipalities of Belo Horizonte, Conceição do Mato Dentro and Sete Lagoas and a very strong correlation for the municipality of Florestal. Moreover, such values indicated linearity between the increase in predicted and observed values and a tendency of predicted values above the average when the observed values were above the average (Martins, 2014).

With the exception of the value obtained for the model of the Sete Lagoas climatological station using the combination of input data "C1", which remained constant, there was an increase in values compared to those obtained using the combinations of input data "C1" and "C2" for model training. The high Pearson's linear correlation coefficient values, in this case, indicated that the ANN was able to learn the seasonality of the rainfall time series and was able to identify the months with increased and decreased rainfall.

The mean absolute error values for the dry season were reduced in relation to those obtained for the same season using the input data combinations "C1" and "C2" for model training. As an exception, for the model of the climatological station of Conceição do Mato Dentro, the mean absolute error increased. Similar to the mean absolute error, there was a reduction in the bias values for most stations in relation to the values of bias obtained for the same season when the ANN used the combination of input data "C1" and "C2" for training, and the tendency to overestimate was maintained.

For the rainy season, by comparison, there was a reduction in the average absolute error value for all climatological stations in relation to the use of the input data combinations "C1" and "C2". The bias values for the same season were between -26.72 and 21.48 mm, showing no defined tendency to decrease or increase in relation to the values obtained for the same indicator in the same season using the data combinations "C1" and "C2" for the ANN training.

Using the combination of input data "C3" in training the ANN, the n_t index presented values between 0.66 and 1.26. Thus, the model was rated as unsatisfactory for the climatological station in the municipality of Conceição do Mato Dentro, acceptable for the climatological stations in the municipalities of Belo Horizonte and Sete Lagoas, and good for the climatological station in the municipality of Florestal. This result showed that the values of the n_t index increased for all climatological stations in relation to the values obtained for training the ANN when the combinations of input data "C1" and "C2" were used. When using the "C3" combination of input data for training, there were fewer models rated as unsatisfactory compared to the other combinations, and one model was rated as good, which had not occurred before.

Comparing the results obtained using different combinations of input data for training the ANN, the model performance using only the three variables that obtained higher correlation coefficients with the target (input data combination "C3") improved. Similarly, Lee *et al.* (2018), using the multilayer perceptron architecture in addition to the feedforward backpropagation algorithm, attempted to predict rainfall in South Korea by initially using data from 10 different climate indices. After the evaluation and selection of five indices that showed better results, the authors obtained a better model performance. Despite the performance improvement after the selection of variables, it was notable that it was not expressed, indicating that even the climate variables most correlated with the rainfall, such as the average compensated temperature, did not contribute as good predictors to the model.

Furthermore, according to the Mann–Whitney test, only the climatological station of the Conceição do Mato Dentro had a significant difference between the observed and predicted time series. It can be explained by the greater magnitude of the mean absolute error value for the dry season obtained for this model, which was approximately 40% higher than the average of the values obtained for the other climatological station models. Such an increase in values caused the median of the predicted rainfall series to also increase, becoming statistically significant. The significant difference between observed and predicted rainfall can also be explained by the significant trend of increasing rainfall, which was found only in December for the climatological station of Conceição do Mato Dentro (Table 2), thus increasing errors in the rainy season.

For the other climatological stations, there was not enough evidence to conclude that there was a significant difference between the observed and predicted time series, i.e., no clear differences were detected between the median value of the observed and estimated time series. The fact that there was no statistically significant difference between the median values for the other climatological stations, coupled with the high Pearson's linear correlation coefficient values, means that the predictors used were able to satisfactorily explain the rainfall phenomenon for the municipalities of Belo Horizonte, Florestal and Sete Lagoas.

Aiming at a visual analysis of the results obtained for training the ANN using the "C3" input data combination, the rainfall values observed and predicted by the ANN on a monthly scale are plotted in Figure 2.

By visual analysis of Figure 2, it is possible to note that there was no expressive difference between the observed and predicted data in the periods of lower rainfall. Additionally, although the models predicted the intervals with greater rainfall with good accuracy, they had difficulty in predicting values above 300 mm, which was probably the reason for the high values of mean absolute error and the tendency to underestimate in the rainy seasons. This fact indicated that the predictors used were not able to explain the occurrence of rainfall above 300 mm.

This might contribute to the model of the climatological station in the Conceição do Mato Dentro municipality being classified as unsatisfactory by the n_t index using a combination of input data “C3”. Consolidating what was indicated by the n_t index, it was noted that the only model classified as good, corresponding to the climatological station of the Florestal municipality, was the one that presented the best visual graphical agreement between series. The superior performance of this model in relation to the others can be explained by the absence of values of total monthly rainfall greater than 300 mm, as in the climatological stations of the other municipalities.

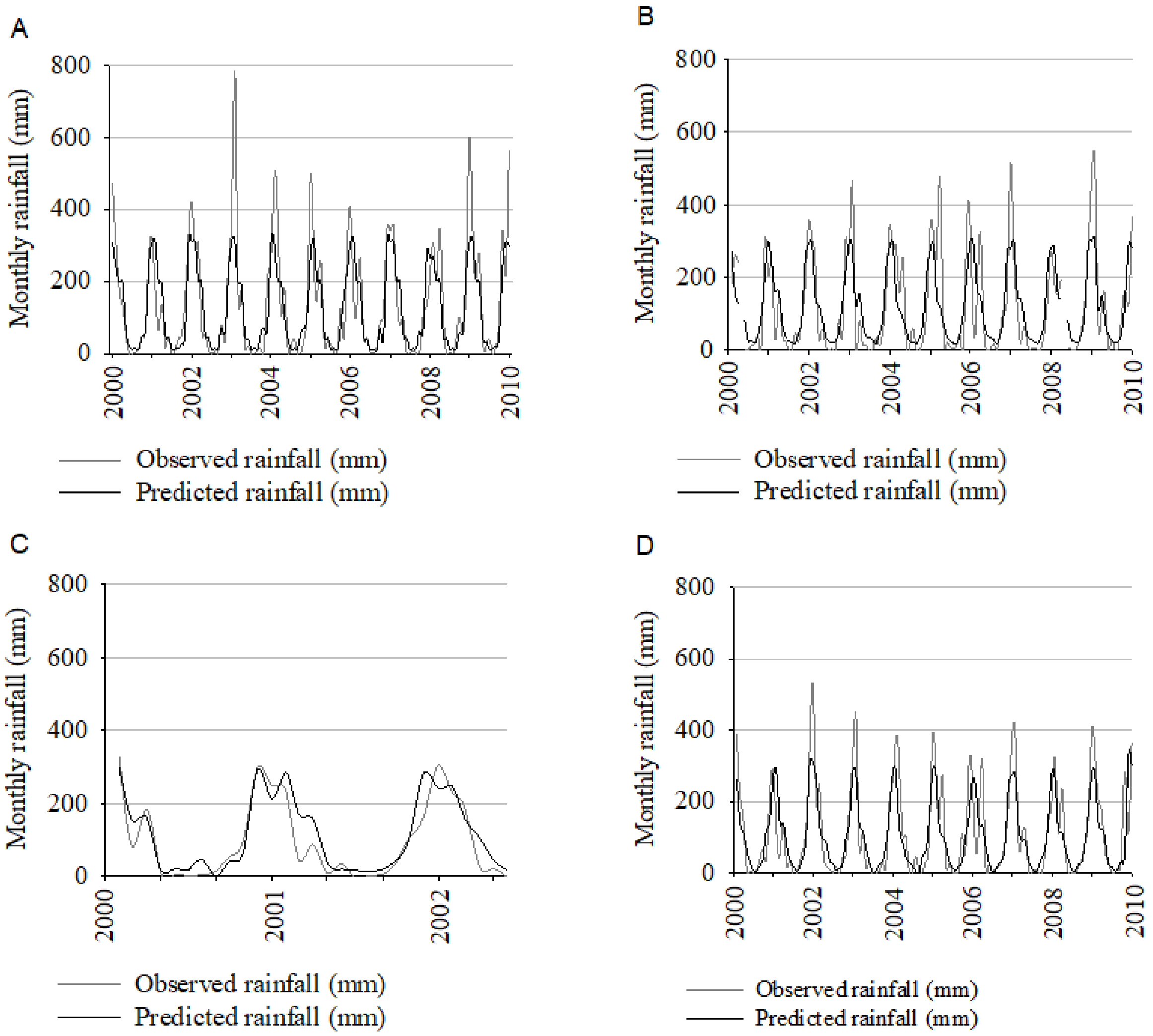

Figure 2. Observed and predicted rainfall by ANN using the "C3" input data combination for training for the climatological stations in the municipalities of Belo Horizonte (A), Conceição do Mato Dentro (B), Florestal (C) and Sete Lagoas (D).

For comparison purposes, one can analyze the prediction obtained for the climatological station of Belo Horizonte, where in 2003, there was the highest peak of observed rainfall among all series, reaching values close to 800 mm/month, and it also had a large error value. However, for the same station, in 2007, there was no rainfall above 300 mm, which led to good graphical agreement between the observed and predicted series. The same fact was verified for the climatological stations of the other municipalities, e.g., Conceição do Mato Dentro between the years 2008 and 2009 and Sete Lagoas between the years 2001 and 2002.

An analogous behavior of the predicted series was detected by Yadav and Sagar (2019) and Nguyen *et al.* (2021). The first authors aimed to predict monthly rainfall in India using multilayer perceptron architecture plus a feedforward backpropagation algorithm. In the ANN training, air temperature, relative humidity, and wind speed data were used. The second author aimed to predict daily and monthly rainfall in the United States of America using multilayer perceptron architecture plus a feedforward backpropagation algorithm. In the ANN training, air temperature, dew point, humidity, pressure, visibility, and wind speed data were used. Both authors graphically demonstrated that the largest values of errors presented by the model occurred in the rainfall peaks. This fact indicates that even after selection, climatic variables may not contribute satisfactorily to the prediction of extreme maximum events in different parts of the world.

Similarly, Moustris *et al.* (2011) used the multilayer perceptron architecture, the feedforward backpropagation algorithm and the input data of maximum, minimum, average and cumulative rainfall of the four previous months, as well as an index to indicate the seasonality of these and the four months to be predicted, to determine the maximum, minimum, average and cumulative rainfall for the upcoming four months in Greece. The model presented a limitation regarding the prediction of rainfall peaks. This fact indicates that rainfall alone is not a good predictor of rainfall peaks. Such studies found in the literature corroborate the results found for the climatological stations of the municipalities analyzed in this study.

According to Moustris *et al.* (2011), one of the explanations for the limitation in predicting periods with extreme values may be because there is not enough data for training. According to the authors, positive rainfall extremes occur with low frequency and high randomness, and if there are not enough records in the data used for training the ANN, they will not acquire the necessary experience needed for prediction.

To understand the error behavior, its magnitude on an annual scale was calculated by subtracting the value of the observed annual rainfall from the value of the predicted rainfall. The balance of the error on an annual scale between the rainfall predicted by means of ANN using the combination of input data "C3" in training and observed for the climatological stations of the municipalities analyzed in the study is indicated in Figure 3.

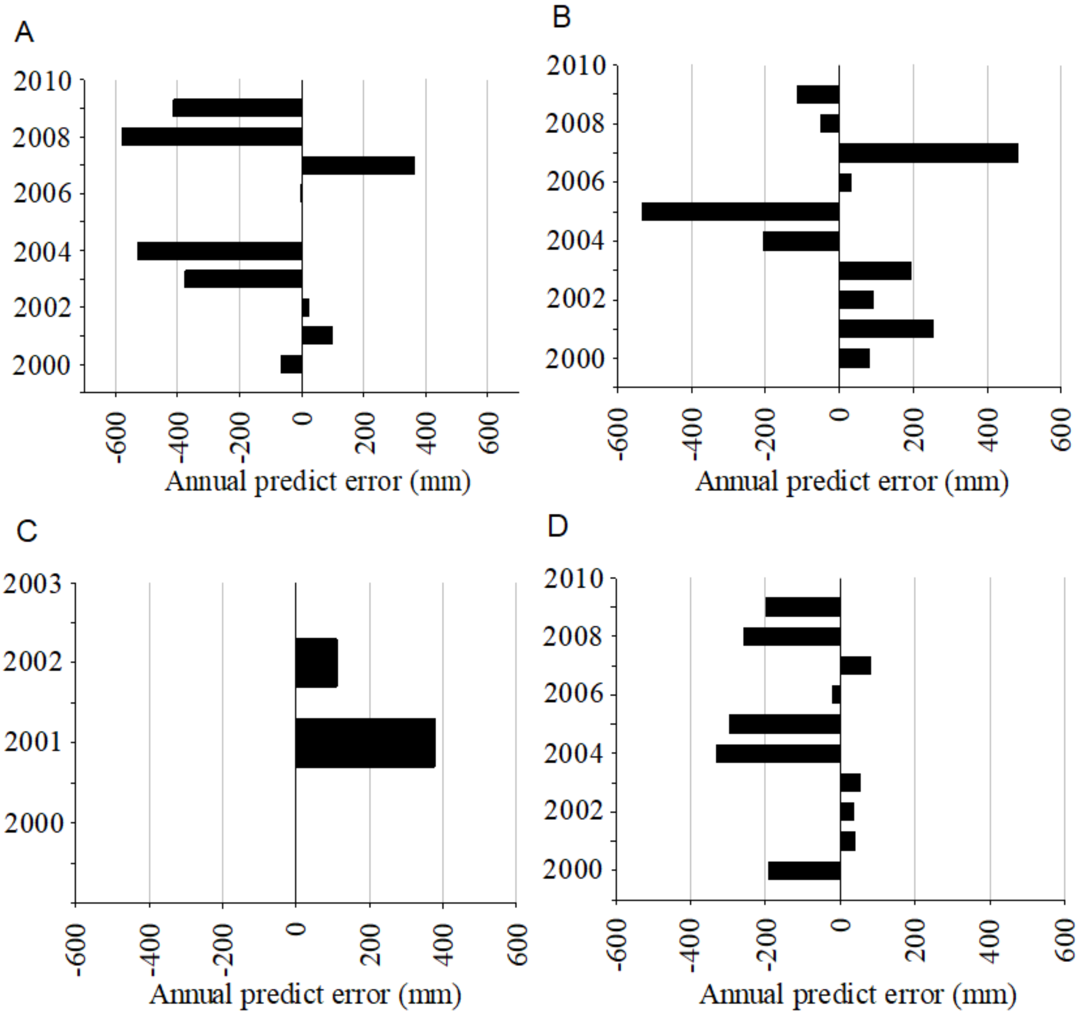

Figure 3. Balance of error on an annual scale between observed and predicted rainfall by the ANN using the "C3" input data combination for training for the climatological stations in the municipalities of Belo Horizonte (A), Conceição do Mato Dentro (B), Florestal (C), and Sete Lagoas (D).

As shown in Figure 3, the magnitude of the annual errors indicating underestimation of the balance was greater than the magnitude of those indicating overestimation, except for the climate station in the municipality of Florestal.

Comparing the annual rainfall to the error values, it was observed that the four years with the highest magnitudes of annual errors for the climatological stations of Belo Horizonte (-577.55; -528.31; -413.50 and -373.80 mm) and Sete Lagoas (-330.10; -296.05; -258.07 and -199.5 mm) corresponded to the four years with the largest observed rainfall. For the climatological station in the Conceição do Mato Dentro municipality, the year with the largest magnitude of error (-532.13 mm) corresponded to the year with the highest observed rainfall volume.

The municipality of Florestal presented only positive values in the balance of the annual error; however, there were only three years of data available for the validation of the predicted rainfall, and in 2000, the annual error was 0.47 mm, which can be disregarded. In comparing the average observed to the observed rainfall in the years 2001 and 2002, it was noticed that these were below average, which was probably the reason for the station presenting only annual overestimation errors. For comparison purposes, in the other climatological stations, with the exception of 2001 for the Sete Lagoas station, the years of overestimation coincided with years of below average observed rainfall.

Regarding the years with balance that indicated overestimation, the greatest magnitude was represented by the climatological station of Conceição do Mato Dentro in 2007. By comparing the average rainfall observed at this station to that year, the latter was below average and was the year with the lowest rainfall. Analyzing the annual rainfall of years without data gaps at Conceição do Mato Dentro stations during the training period in relation to the values observed in 2007, there were only two years with lower rainfall. This fact reinforces that, for the model to be able to perform a good forecast, the predictor-time series should include examples of months with extreme values in sufficient quantity for such unusual events to be understood by the ANN.

4. CONCLUSIONS

Of the climatic variables used as predictors in this study, the average compensated temperature presented, for the research area, high values of correlation to rainfall and the others had low values. Despite the ENSO phenomenon altering the dynamics of rainfall in several parts of the world, including Brazil, for the analyzed region, the information on its occurrence presented the worst performance as a predictor. Even using the variables most correlated with precipitation as predictors, there was no notable improvement in the predictive capacity of the models. The predictors used generally led the models to satisfactorily predict the rainfall, but there was a limitation in the prediction of extreme rainfall data. Rainfall forecasting through the developed models could provide information to help in decision-making regarding the potential damage caused by drought periods, e.g., problems with water supply, water quality and water flow in dams, and actions against damage caused by floods during high rainfall periods, especially in areas with a high rate of soil waterproofing, such as Belo Horizonte city.

5. ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

6. REFERENCES

AKSOY, H.; DAHAMSHEH A. Artificial neural network models for forecasting monthly precipitation in Jordan. **Stochastic Environmental Research and Risk Assessment**, v. 23, n. 7, p. 917-931, 2009. https://doi.org/10.1007/s00477-008-0267-x

ALENCAR, L. P. de; SEDIYAMA G. C.; MANTOVANI, E. C. Estimativa da evapotranspiração de referência (ETo padrão FAO), para Minas Gerais, na ausência de alguns dados climáticos. **Engenharia Agrícola**, v. 35, n. 1, p. 39-50, 2015. https://doi.org/10.1590/1809-4430-Eng.Agric.v35n1p39-50/2015

AMIRI, M. A.; CONOSCENTI, C.; MESGARI, M. S. Improving the accuracy of rainfall prediction using a regionalization approach and neural networks. **Kuwait Journal of Science**, v. 45, n. 4, p. 66-75, 2018.

ESTEVES, J. T.; ROLIM, G. de S.; FERRAUDO, A. S. Rainfall prediction methodology with binary multilayer perceptron neural networks. **Climate Dynamics**, v. 52, p. 2319-2331, 2019. https://doi.org/10.1007/s00382-018-4252-x

G1 MINAS. **Chuva destrói parte de BH; MG tem 55 mortos em 6 dias**. 2020. Available at: https://g1.globo.com/mg/minas-gerais/noticia/2020/01/29/apos-mais-um-temporal-com-enchentes-bh-e-regiao-metropolitana-contabilizam-mais-estragos.ghtml. Access April 06 2022.

GRIMM, A. M.; FERRAZ, S. E. T. Sudeste do Brasil: Uma região de transição no impacto de eventos extremos da Oscilação Sul parte 1: El Niño. *In*: CONGRESSO BRASILEIRO DE METEOROLOGIA, 10., 1998, Brasília. **Proceedings[…]** Brasília: SBMET, 1998. 1 CD-ROM.

GONZALEZ-FERNANDEZ, I.; IGLESIAS-OTERO, M. A.; ESTEKI, M.; MOLDES, O. A.; MEJUTO, J. C.; SIMAL-GANDARA. J. A critical review on the use of artificial neural networks in olive oil production, characterization and authentication. **Critical Reviews in Food Science and Nutrition**, v. 59, n. 12, p. 1913-1926, 2019. https://doi.org/10.1080/10408398.2018.1433628

HOSSAIN, I.; ESHA, R; IMTEAZ, M. A. An attempt to use non-linear regression modelling technique in long-term seasonal rainfall forecasting for Australian capital territory. **Geosciences**, v. 8, n. 8, p. 1-12, 2018. https://doi.org/10.3390/geosciences8080282

HOSSAIN, I.; RASEL, H. M. IMTEAZ, M. A.; MEKANIK, F. Long-term seasonal rainfall forecasting using linear and non-linear modelling approaches: a case study for Western Australia. **Meteorology and Atmospheric Physics**, v. 132, p. 131-141, 2020. https://doi.org/10.1007/s00703-019-00679-4

INMET. **Banco de Dados Meteorológicos Para Ensino e Pesquisa**. Available at: http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep. Access February 17 2022.

LEE, J.; KIM, C; LEE, J. E.; KIM, N. W.; KIM. H. Application of artificial neural networks to rainfall forecasting in the Geum River Basin, Korea. **Water**, v. 10, n. 10, 2018. https://doi.org/10.3390/w10101448

LEVENBERG, K. A method for the solution of certain non-linear problems in least squares. **Quarterly of Applied Mathematics**, v. 2, n. 2, p. 164–168, 1944.

MARTINS, F. B.; GONZAGA, G.; SANTOS, D. F. dos; REBOITA, M S. Classificação climática de Köppen e de Thornthwaite para Minas Gerais: Cenário atual e projeções futuras. **Revista Brasileira de Climatologia**, Special Edition, p. 129-146, 2018. http://dx.doi.org/10.5380/abclima.v1i0.60896

MARTINS, M. E. G. Coeficiente de correlação amostral. **Revista de Ciência Elementar**, v. 2, n. 2, 2014. http://doi.org/10.24927/rce2014.042

MARTINS, P. A. da S.; QUERINO, C. A. S.; MOURA, M. A. L.; QUERINO, J. K. A. da S.; MOURA, A. R. de M. Variabilidade espaço-temporal de variáveis climáticas na mesorregião sul do Amazonas. **Revista Ibero-Americana de Ciências Ambientais**, v. 10, n. 2, p. 169-184, 2019. https://doi.org/10.6008/CBPC2179-6858.2019.002.0015

MATLAB. **V. R2011a**. The MathWorks Inc, 2011.

MAWONIKE, R.; MANDONGA G. The effect of temperature and relative humidity on rainfall in Gokwe region, Zimbabwe: A factorial design perspective. **International Journal of Multidisciplinary Academic Research**, v. 5, n. 2, p. 36-46, 2017.

MAY, R.; DANDY, G.; MAIER, H. Review of input variable selection methods for artificial neural networks. Chap. 2. *In:* SUZUKI, K. (ed.). **Artificial Neural Networks**: Methodological Advances and Biomedical Applications. London: InTech, 2011. https://doi.org/10.5772/644

MOHAMMADPOUR, R.; ASAIE, Z. SHOJAEIAN, M. R.; SADEGHZADEH, M. A hybrid of ANN and CLA to predict rainfall. **Arabian Journal of Geosciences**, v. 11, p. 1-9, 2018. https://doi.org/10.1007/s12517-018-3804-z

MOUSTRIS, K. P.; LARISSI, I. K.; NASTOS, P. T.; PALIATSOS, A. G. Precipitation forecast using artificial neural networks in specific regions of Greece. **Water Resources Management**, v. 25, n. 8, p. 1979-1993, 2011. https://doi.org/10.1007/s11269-011-9790-5

MOTA, E. P. da; CUNHA, D. M.; CRUZ, F. M.; PANQUESTOR, E. K. Precipitações em Governador Valadares - MG e sua relação com o fenômeno ENOS nos períodos chuvosos de 2008 a 2017. **ForScience**, v. 7, n. 1, e003355, 2019. https://doi.org/10.29069/forscience.2019v7n1.e355

NOAA. **Multivariate ENSO Index (MEI)**. Available at: https://www.esrl.noaa.gov/psd/enso/mei.old/ Access February 13 2020.

NOAA. **Weather observations**. 2011. Available at: https://www.noaa.gov/education/resource-collections/weather-atmosphere/weather-observations Access March 8 2021.

NGUYEN, H. N.; NGUYEN, T.; LY, H.; TRAN, V. Q. NGUYEN, L. K.; NGUYEN, M. V. *et al.* Prediction of daily and monthly rainfall using a backpropagation neural network. **Computer Science and Information Engineering**, v. 24, n. 3, p. 367-379, 2021. http://dx.doi.org/10.6180/jase.202106_24(3).0012

PERES, T. C.; MAIER, E. L. B. Rainfall variability in Brazil between 1920 and 2010. **International Journal of Climatology**, v. 42, n. 13, 2022. https://doi.org/10.1002/joc.7622

RITTER, A.; MUÑOZ-CARPENA, R. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. **Journal of Hydrology**, v. 480, p. 33-45, 2013. https://doi.org/10.1016/j.jhydrol.2012.12.004

SILVA, M. A. da; MERLO, M. N.; THEBALDI, M. S.; VIOLA, M. R. Validação da precipitação estimada pelo produto 3B42 do satélite "Tropical Rainfall Measuring Mission" para a sub-bacia hidrográfica Paraíba do Sul, estado de São Paulo, Brasil. **Revista Brasileira de Climatologia**, v. 28, p. 585-601, 2021. http://dx.doi.org/10.5380/rbclima.v28i0.75926

SCHOBER, P.; BOER, C.; SCHWARTE, L. A. Correlation coefficients: appropriate use and interpretation. **Anesthesia & Analgesia**, v. 126, n. 5, p. 1763-1768, 2018. https://doi.org/10.1213/ANE.0000000000002864

TAURO, F.; SELKER, J.; VAN de GIESEN, N.; ABRATE, T.; UIJLENHOET, R.; PORFIRI, M. *et al.* Measurements and observations in the XXI century (MOXXI): innovation and multi-disciplinarity to sense the hydrological cycle. **Hydrological Sciences Journal**, v. 63, n. 2, p. 169-196, 2018. https://doi.org/10.1080/02626667.2017.1420191

TIAN, J.; LIU, J.; YAN, D.; LI, C.; CHU, Z.; YU, F. An assimilation test of Doppler radar reflectivity and radial velocity from different height layers in improving the WRF rainfall forecasts. **Atmospheric Research**, v. 198, n. 1, p. 132-144, 2017. https://doi.org/10.1016/j.atmosres.2017.08.004

TORRES, J. F.; HADJOUT, D.; SEBAA, A.; MARTINEZ-ÁLVAREZ, F.; TRONCOSO, A. Deep learning for time series forecasting: a survey. **Big Data**, v. 9, n. 1, p. 3-21, 2021. https://doi.org/10.1089/big.2020.0159

YADAV, P.; SAGAR, A. Rainfall prediction using artificial neural network (ANN) for tarai Region of Uttarakhand. **Current Journal of Applied Science and Technology**, v. 33, n. 5, p. 1-7, 2019. https://doi.org/10.9734/cjast/2019/v33i530096