Revista Brasileira de Computação Aplicada, April, 2022

DOI: 10.5335/rbca.v14i1.12964
Vol. 14, Nº 1, pp. 70–80
Homepage: seer.upf.br/index.php/rbca/index

RBCA ISSN 2176-6649

ORIGINAL PAPER

Hydrological forecast in Macaé River Basin with Neural Networks

Julia Godinho[1], Janaina Santanna Gomide Gomes[1], Rafael Malheiro[1], Laura Emmanuella Santana [2]

[1]Instituto Politécnico - UFRJ, [2]Escola Agrícola de Jundiaí - UFRN

*juliagodinho08@gmail.com; janainagomide@gmail.com; rafaelmalheiro@globo.com; lauraemmanuella@gmail.com

Received: 2021-09-14. **Revised:** 2022-04-17. **Accepted:** 2022-04-26.

Abstract

Background: Hydrological forecasting is a valuable tool for dealing with socio-environmental problems, and it can be used in natural disaster alert systems and as assistant aid in making public policies. This work presents an application of a hydrological model based on Artificial Neural Networks (ANN). The variable modeled was the flood stage of the fluviometric station Fazenda Airis, located in the Macaé River drainage basin. For this purpose, the datasets used are composed of daily records of flow and rainfall stations between 2010 to 2013, made available by the National Water Agency (ANA) and the INEA (Environment State Institute of Rio de Janeiro) Flood Alert System. The adopted methodology investigates the influence of the input variables and ANN architecture on the models' performance. **Results:** The results obtained were considered very satisfactory and support the proposition of the potential of Artificial Neural Networks for hydrological forecasting. It was found that of the 189 models created, 42.3 % had the coefficient of determination R^2 above 0.80. **Conclusions:** The best ANN developed received daily data from six rainfall stations and one fluviometric station, obtaining for metrics R^2 and MAE the values of 0.88, 7.03 cm, respectively. Finally, the results were compared with related works and are similar or superior even with shorter time series.

Keywords: Artificial Neural Networks; Drainage Basin of the Macaé River; Hydrological forecasting.

Resumo

Background: A previsão hidrológica é uma ferramenta valiosa no tratamento de problemas socioambientais, podendo ser utilizada em sistemas de alerta de desastres naturais e auxiliar na formulação de políticas públicas. Este trabalho apresenta a aplicação de um modelo hidrológico baseado em Redes Neurais Artificiais (RNA). A variável modelada foi o estágio de inundação da estação fluviométrica Fazenda Airis, localizada na bacia do rio Macaé. Para tanto, os conjuntos de dados utilizados são compostos por registros diários de vazões e estações pluviométricas entre 2010 a 2013, disponibilizados pela Agência Nacional de Águas (ANA) e pelo Sistema de Alerta de Inundação do INEA (Instituto Estadual do Meio Ambiente do Rio de Janeiro). A metodologia adotada investiga a influência das variáveis de entrada e da arquitetura da RNA no desempenho dos modelos. **Resultados:** Os resultados obtidos foram considerados muito satisfatórios e suportam a proposição do potencial das Redes Neurais Artificiais para previsões hidrológicas. Constatou-se que dos 189 modelos criados, 42,3% tinham o coeficiente de determinação R^2 acima de 0,80. **Conclusões:** A melhor RNA desenvolvida recebeu dados diários de seis estações pluviométricas e uma estação fluviométrica, obtendo para as métricas R^2 e MAE os valores de 0,88 e 7,03 cm, respectivamente. Por fim, os resultados foram comparados com trabalhos relacionados e são semelhantes ou superiores mesmo com séries temporais mais curtas.

Palavras-Chave: Redes Neurais Artificiais; Bacia do Rio Macaé; Previsão hidrológica.

1 Introduction

Floods are natural phenomena resulting from the rise in the water level in a river section, caused by extreme rainfall events and can be intensified by human interventions. Rio de Janeiro is among the six Brazilian states with the highest number of registered natural disasters, where 65% of all occurrences are associated with flooding

(Ikemoto et al., 2015).This high incidence is due, in part, to the urbanization process and the Straightening of rivers. The urbanization process promotes the increased waterproofing, surface runoff, land occupation and construction of rain conduits, thus it increases the maximum flow peaks during rainfall events (Tucci and Silveira, 2002).

Short-term hydrological forecasting is one of the tools used in disaster risk management due to its ability to predict future situations a few days or hours in advance. It is commonly used in the natural disaster warning system, in addition to assisting in decision-making processes in public policies (dos Santos Neto et al., 2020).

The National Center for Monitoring and Alerting of Natural Disasters (Cemaden), created in 2011 by the federal government in Brazil, is responsible for monitoring meteorological events and natural threats in Brazilian municipalities susceptible to the occurrence of natural disasters. Observing the potential of computational intelligence techniques for building hydrological models, Cemaden has modeling hydrometeorological data through Machine Learning techniques for the Grande River in the state of Rio de Janeiro (de Lima and Scofield, 2017). As a result, a model based on Artificial Neural Networks was developed in de Lima et al. (2016) capable of forecasting the river stage within 120 minutes in advance.

Works have applied the Artificial Neural Networks (ANN) technique for hydrological forecasting in different temporal bases using approaches to define the best architecture (Machado, 2005, da Silva et al., 2018), feature selection techniques and to improve generalization (Rodrigues et al., 2015). Furthermore, some authors compare the results with different machine learning algorithms (Freire et al., 2009, Gorodetskaya et al., 2018, dos Santos Sousa et al., 2018).

This work aims to evaluate models based on Artificial Neural Networks for hydrological forecasting on a daily basis in the Macaé river basin, located on the coast of the State of Rio de Janeiro, Brazil. In order to achieve this goal, the proposed methodology gathers different datasets which cover daily rainfall and flow data, proposes different features combinations and evaluates models based on ANN varying theirs hyper-parameters to predict the stage of the Macaé river. The code developed for this work is available in the git repository[1].

2 Related Work

Studies have evaluated the performance of models based on ANNs in hydrological modeling, observing the applicability of the technique in issuing alerts and managing water resources (de Lima and Scofield, 2021, Shamseldin, 2009).

In this context, in Debastiani et al. (2016) the authors evaluated twelve data treatments with combinations of precipitation, evapotranspiration and flow variables, as well as their temporal transformations and time step for hydrological forecasting on a daily scale. The input vectors were defined after an exploratory correlation analysis between the hydroclimatic variables.

Similarly, the authors in da Silva et al. (2016) investigated the best combination of input data with daily precipitation and flow data from a sub-basin located on the north coast of São Paulo, in the period 1985 to 1989. The simultaneous use of precipitation and flow data resulted in better ANNs, with Nash-Sutcliffe coefficients of 0.77. Thus, the authors highlighted the need to promote continuous monitoring of the two variables.

Some pieces of work use the ANNs technique for forecasting on a monthly scale, such as the work presented in Machado (2005), which proposes a hydrological model using precipitation, flow and potential evapotranspiration data to predict flow in the Jangada River, Paraná, Brazil. The best ANNs had a correlation coefficient equal to 96.9%. The good result obtained was associated with the adopted methodology, which considered the influence of the number of neurons in the input layer and in the intermediate layer, the number of epochs, the initialization and extension of the data series.

In Matos et al. (2014) the authors propose upstream control alternatives to the area of Ponte Mística. The dataset contains daily and hourly data of rainfall and flow from the Ponte Mística and flows of sub-basins located upstream of it.

The methodology adopted in Sousa and de Sousa (2010) investigates the ANN architecture, varying the neurons in the input and intermediate layers and the activation function, using data from a fluviometric station and five rainfall stations. It was found that an ANN with normalized input data, 15 neurons in the hidden layer and a logarithmic sigmoid activation function is able to predict with 77.0% efficiency the average flow of the Piancó river basin. Other authors adopted the approach of investigating the architecture of ANN and data input (Celeste et al., 2014, Cristaldo et al., 2018, Gorodetskaya et al., 2018, dos Santos Neto et al., 2020, da Silva et al., 2018, Alberton et al., 2021, Mendonça et al., 2021).

Some examples compare other Machine Learning techniques with ANNs. Batista (2009) implemented time series models and multi-layer perceptron ANNs to predict the flow of the Grande River, a tributary of Camargos hydropower plant reservoir. It was identified that the ANNs were adequate to the data in question and found results superior to the time series. The work presented in dos Santos Sousa et al. (2018) compared the algorithms of Decision Trees, Random Forest, Logistic Regression and ANNs to forecast the annual volume of rain in Manaus. After training and testing more than 4,000 models with different parameters, it was verified that the ANNs obtained the best performances.

Some studies have used deep neural networks for hydrological forecasting. In Barino and Bessa (2020), for example, the forecasting model is based on one-dimensional convolutional neural networks. The authors used as input a time series of flow and turbidity of the Madeira River in the Amazon in the period 2016 to 2019. The authors concluded that when using flow and turbidity as input, the predictive model becomes up to 5 times simpler than the model using only flow.

This work also proposes to use ANNs to perform the hydrological prediction. However, this is the first

[1]Link: https://github.com/JuliaGodinho08/RNA_Macae

Table 1: Rainfall stations selected from Macaé river basin.

Number	Station	Period	Latitude	Longitude	Municipality
2241005	Fazenda São João	1967-2020	-22.3903	-42.4947	Nova Friburgo
2242004	Galdinópolis	1950-2020	-22.3631	-42.3808	Nova Friburgo
2242003	Piller	1950-2020	-22.4047	-42.3392	Nova Friburgo
2242145	Ponte Baião	2010-2013	-22.383	-42.083	Macaé
2241032	Fazenda Airis	2011-2013	-22.3281	-41.9825	Macaé
2241030	Severina	2011-2013	-22.2956	-41.8786	Macaé

time, from our knowledge, that Machine Learning algorithms were used to predict the Macaé river stage. Another differential of this work is the study of features combinations that consider fluviometric and pluviometric stations to evaluate the ANN models and even with a smaller time series the results achieved are equal or better to the related work.

3 Methodology

The proposed methodology is composed of five steps: 1) definition of the study area, 2) data collection, 3) data preparation, 4) definition of input variables, 5) training of ANNs and 6) performance analysis.

3.1 Definition of the Study Area

The study area corresponds to the Macaé river drainage basin, located in the central-northern coastal strip of the State of Rio de Janeiro in Brazil (Fig. 1). The drainage area of this basin is approximately 1,765 km². The Macaé river is the main stream of the basin and originates in the Serra Macaé de Cima in Nova Friburgo. It travels for about 136 km, flowing into the Atlantic Ocean in the city of Macaé. This is the first work of our knowledge that makes the hydrological forecast for this region using neural networks.

3.2 Data collection

The datasets consist of daily flow data and daily precipitation data. All data were obtained through the website Hidroweb[2] from National Water Agency (ANA) or made available by the Flood Alert System of the Environment State Institute (INEA).

The historical rainfall series were obtained from Hidroweb website and 6 rainfall stations located within the study area were selected: Fazenda São João (2241005), Galdinópolis (2242004), Piller (2242003), Ponte Baião (2242145), Fazenda Airis (2241032) e Severina (2241030). The selection criteria considered the proximity to the target station and the length of the historical series. Table 1 shows the length of the historical series and the location of each rainfall station, which can also be seen in Fig. 1.

The historical series of discharge and stage were obtained from the Hidroweb portal and provided by the Flood Alert System at INEA. Datasets from six fluviometric stations were collected. These datasets have discharge or stage data of rivers in the Macaé river basin available. These stations are: Macaé de Cima (59120000), Galdinópolis (59125000), Piller (59135000), Barra do Sana (59134000), Fazenda Airis (59138800) e São Pedro (59143000). Table 2 shows for each monitoring station the location and the river it monitors. The location of the stations can be seen in Fig. 1.

The Fazenda Airis station (59138800) provides level data and will be the object of the forecast.The choice of Fazenda Airis as a target station for the forecast models was due to its location, close to the urbanized area of the municipality of Macaé. The other fluviometric stations will be used as input data in the training of the ANNs. These are located upstream from the Fazenda Airis (5913880), except for the São Pedro station (59143000), which monitors the river of the same name, one of the main tributaries of the Macaé River.

Fig. 2 shows the pluviograph chart and hydrograph of the Fazenda Airis (59138800) between 2010 and 2013, according to the collected data. Through this, it was possible to graphically verify the responses to precipitation events at the channel level.

3.3 Data preparation

Data preparation includes data selection, cleaning and transformation, in order to prepare the dataset that contains missing values and features with different scales.

The missing values of the historical series were filled by a linear regression algorithm as suggested in de Mello et al. (2017). After, the data were divided into training, validation and testing sets, with 60%, 20% and 20%, respectively. It is important to mention that when we use neural networks, the algorithm does not perform well when the numeric input attributes have very different scales. Thus, it is necessary to apply feature scaling. The features of the dataset are normalized rescaled so that they end up ranging from 0 to 1, agreeing with the limits of the ReLU activation function chosen for the experiments as suggested in Dawson and Wilby (2001).

Data manipulation and analysis were performed in Python language, using the Pandas library[3] and Scikit-Learn[4] for pre-processing and data evaluation.

[2]Hidroweb : https://www.snirh.gov.br/hidroweb/serieshistoricas

[3]Pandas: https://pandas.pydata.org/docs/index.html

[4]Scikit-learn: https://scikit-learn.org/stable/index.html

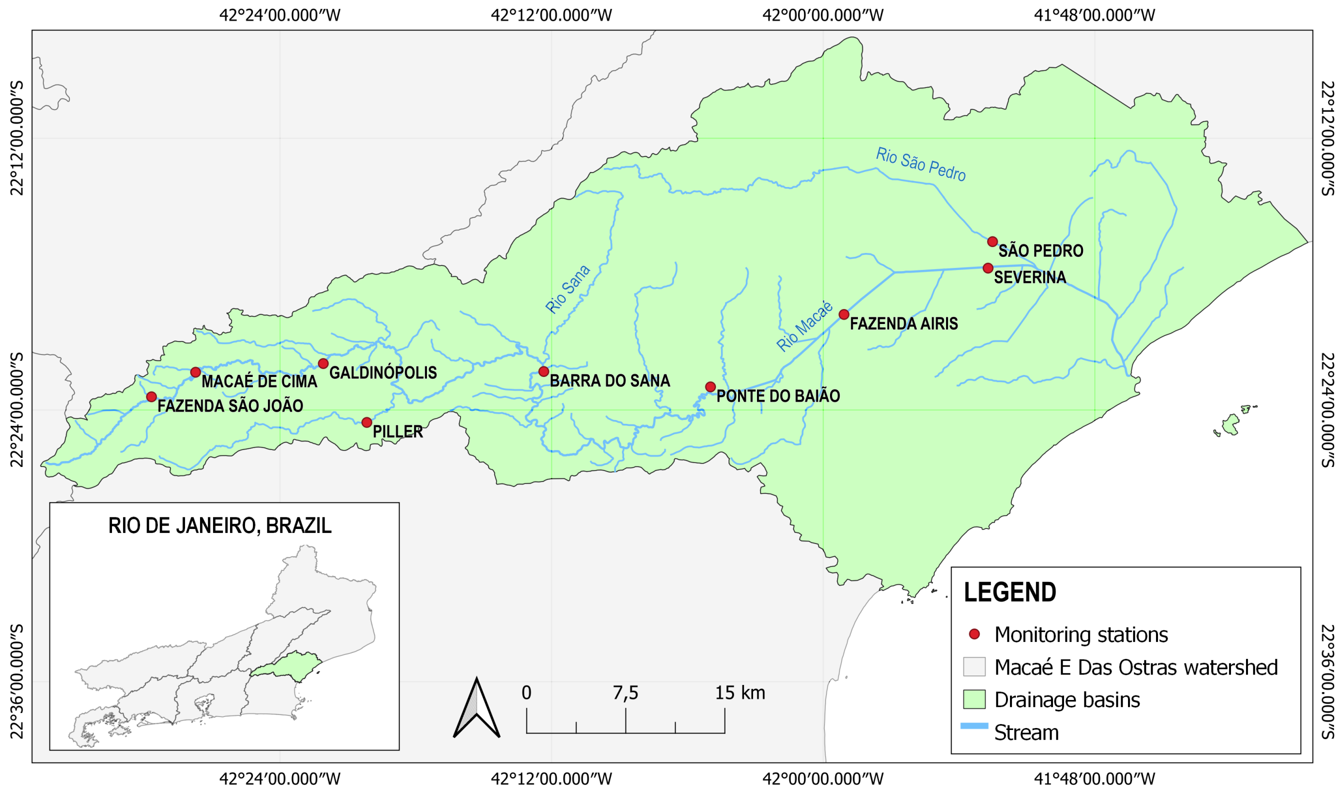


Figure 1: Macaé E Das Ostras Drainage Basin.

Table 2: Selected fluviometric stations in the Macaé river basin.

Number	Station	Period	Latitude	Longitude	Monitored river
59120000	Macaé de Cima	1967-2020	-22.3722	-42.4622	Rio Macaé de Cima
59125000	Galdinópolis	1950-2020	-22.3689	-42.3792	Rio Macaé
59135000	Piller	1950-2020	-22.4092	-42.3361	Rio Bonito
59134000	Barra do Sana	2011-2013	-22.3686	-42.2058	Rio Macaé
59138800	Fazenda Airis	2010-2013	-22.3281	-41.9825	Rio Macaé
59143000	São Pedro	2011-2013	-22.2761	-41.8753	Rio São Pedro

3.4 Definition of input variables

In order to investigate the influence of the features in predicting the river level, 9 combinations of features were proposed. Table 3 presents all combinations with the applied variables, the interval of the historical series and the number of records contained. The extension of the series was defined by the stations covered by each alternative. For all models, the output variable is at the daily level at the Fazenda Airis fluviometric station (59138800).

The feature combinations were chosen in order to enable the analysis of the contribution between rainfall and flow data for different response times. The response time of some alternatives was chosen experimentally. Below is a description of each set.

Alternative 1: for forecasting the level in time t, this input alternative receives daily rainfall, discharge and level data from all monitoring stations (Tables 1 and 2) with one day lag ($t-1$). Thus, each variable represents a neuron in the RNA input layer;

Alternative 2: similar to alternative 1. However, it receives the data of the last 2 days, being the dataset with the most records;

Alternative 3: uses daily rainfall data from the previous 4 days. Thus, it is expected to evaluate the effectiveness of a discharge prediction model that receives only precipitation data. This alternative is suitable for evaluating the performance considering areas where there are no fluviometric stations yet. As the rainfall monitoring network in Brazil is broader than fluviometric monitoring, as noted in Sarmento (2021), this could be an alternative for level predicting where only rainfall stations data is available.

Alternative 4: receives only daily data of average rainfall with a response time of 4 days. Values were estimated by arithmetic mean ($\overline{P}$) of rainfall stations from Table 1. This way, the ANNs trained with this alternative will have four neurons in the input layer, one neuron for each of the four days preceding the forecast in time t;

Alternative 5: combines daily data from fluviometric

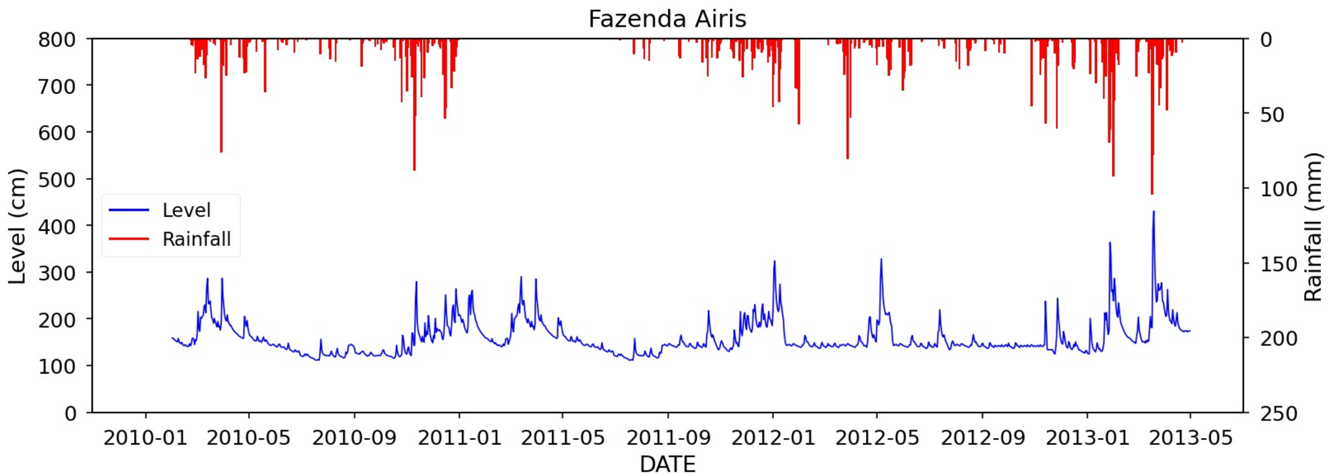


Figure 2: Pluviograph chart and hydrograph of Fazenda Airis between 2010 e 2013.

Table 3: Input data alternatives with the features used, data lag in days, historical series interval and number of records each model receives in training.

N	Variables	Data lag (days)	Period
1	$P1$, $P2$, $P3$, $P4$, $P5$, $P6$, $Q1$, $Q2$, $Q3$, $Q4$, $Q5$, $Q6$	1	02/05/11 a 30/04/13
2	$P1$, $P2$, $P3$, $P4$, $P5$, $P6$, $Q1$, $Q2$, $Q3$, $Q4$, $Q5$, $Q6$	2	03/05/11 a 30/04/13
3	$P1$, $P2$, $P3$, $P4$, $P5$, $P6$	4	05/05/11 a 30/04/13
4	$\overline{P}$	4	01/02/10 a 30/04/13
5	$\overline{P}$, $Q1$, $Q2$, $Q3$, $Q4$, $Q5$, $Q6$	1	02/01/11 a 30/04/13
6	$Q1$, $Q2$, $Q3$, $Q4$, $Q5$, $Q6$	1	02/01/11 a 30/04/13
7	$Q5$	1	02/02/10 a 30/04/13
8	$P1$, $P2$, $P3$, $P4$, $P5$, $P6$, $Q5$	1	02/05/11 a 30/04/13
9	$Q1$, $Q2$, $Q3$, $Q4$, $Q6$	1	04/01/11 a 30/04/13

Note: The variables $P1$, $P2$, $P3$, $P4$, $P5$ and $P6$ precipitation data from rainfall stations of Fazenda São João, Galdinópolis, Piller, Ponte do Baião, Fazenda Airis e Severina, respectively; $\overline{P}$ is the average rainfall of the cited monitoring stations. The variables $Q1$, $Q2$, $Q3$, $Q4$, $Q5$ and $Q6$ are the daily readings recorded by the fluviometric stations Macaé de Cima, Galdinópolis, Piller, Barra do Sana, Fazenda Airis and São Pedro, respectively.

stations and average rainfall ($\overline{P}$) with a one-day response time;

Alternative 6: receives daily data from 6 fluviometric stations in $(t-1)$. Thus, it is expected to analyze the feasibility of a forecasting model that does not receive rainfall attributes;

Alternative 7: uses only the daily stage data from the Fazenda Airis ($Q5$) in $(t-1)$, so the ANNs trained with this set will have only one neuron in the input layer. This alternative can be considered in basins with limited hydrological monitoring, as it requires monitoring of only one variable;

Alternative 8: combines rainfall data from all rainfall stations and daily level in Fazenda Airis ($Q5$) with a lag of one day;

Alternative 9: considers only flow data upstream and downstream of the Fazenda Airis. This combination evaluates the performance of a model that does not receive data from the target station as an attribute.

3.5 ANNs training

The hyperparameters that were varied in training are the number of hidden layers and the number of neurons. So that, for each input alternative (Table 3) ANNs will be created with 1, 2 and 4 hidden layers, also varying the number of hidden neurons between 5, 10, 20, 50 , 80, 100 and 120. Thus, each input alternative will be trained for 21 different architectures. At the end of the experiments, there are 189 models.

In order to improve the generalization of the ANNs used, two regularization criteria were implemented: regularization ℓ_2 and early stopping (Rodrigues et al., 2015). The early stopping technique ends the training when the validation error stops decreasing, avoiding model overfitting. While the regularization ℓ_2 reduces the magnitude of the weights of the ANNs.

In all models, the Adam optimization algorithm was implemented, used so that ANN adjusts its weights faster and more efficiently. Considering the randomness of the initializations of the ANN weights, each model was initialized and trained 5 times. The results will be presented by the mean obtained and with its respective standard deviation.

The process of creating, training and testing the ANNs was done using the *Keras* package, an API of *TensorTensorFlow* 2 used for ML solutions[5].

3.6 Performance Evaluation

To analyze the performance of the model, error and adjustment measures were combined as suggested in Dawson and Wilby (2001). The metrics used are: Mean Absolute Error (MAE), Mean Square Error (MSE) and Coefficient of Determination (R^2), see Eqs. (1) to (3), respectively;

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \quad (1)$$

$$MSE(\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

where $\hat{y}_i$ are the predicted values, y_i the observed values, $\bar{y}$ is the mean of the observed values, and n is the amount of data.

4 Results and Discussions

In this section we will analyze the results obtained in the experiments proposed in the previous section for the prediction of the daily level of the Fazenda Airis fluviometric station.

The first part of the discussion is directed to the architecture of the neural network, then the results will be discussed considering the input variables used.Each alternative input set (Table 3) was trained for 21 different architectures, as per the hyperparameters shown in the previous section.

All ANNs will be identified by the nomenclature that indicates the input set, the number of neurons and hidden layers. The first number after the letter "M" refers to the alternative input data, as per Table 3. The number after the letter "C" indicates the amount of hidden layers of the ANN, which can assume the values of 1, 2 or 4 layers.

[5]Keras API: https://keras.io/about/

Table 4: Best models for each input alternative, using as reference the coefficient of determination R^2. Where, M is input alternative (Table 3) and the data lag is given in days.

M	ANN	MSE	MAE	R^2
1	M1C4N10	180.78	6.34	0.88 (0.04)
2	M2C1N5	164.57	6.81	0.87 (0.06)
3	M3C1N20	746.53	17.19	0.49 (0.10)
4	M4C4N20	798.97	19.44	0.45 (0.07)
5	M5C4N5	182.60	6.81	0.86 (0.07)
6	M6C4N20	219.29	7.57	0.86 (0.02)
7	M7C1N180	190.12	6.94	0.84 (0.03)
8	M8C1N5	178.97	7.03	0.88 (0.01)
9	M9C4N100	715.91	13.06	0.53 (0.22)

The number after the letter "N" refers to the number of neurons present in each intermediate layer, which varies between 5, 10, 20, 50, 80, 100 and 120. Taking as an example, the nomenclature M3C2N100 refers to an ANN that received the input variables of alternative 3, having 2 hidden layers and 100 neurons in each hidden layer.

Of the 189 models generated, 80 obtained the coefficient R^2 above 80%. Where the best value of R^2 obtained was 0.88 (0.01) by ANN M8C1N5, which received input set 8, with a hidden layer and five neurons in the hidden layer. The prediction obtained by the model is shown in Fig. 3. Discussions about the experiments will be presented below.

4.1 Architecture

Of the 21 proposed architectures, Fig. 4 shows the frequency of each configuration among the models with coefficient R^2 above 0.80. We can observe that simpler architectures (i.e. with fewer hidden layers and neurons) were able to perform predictions as effectively as the more complex ones. Therefore, there was no significant gain in model performance by adding hidden layers and neurons in these layers.

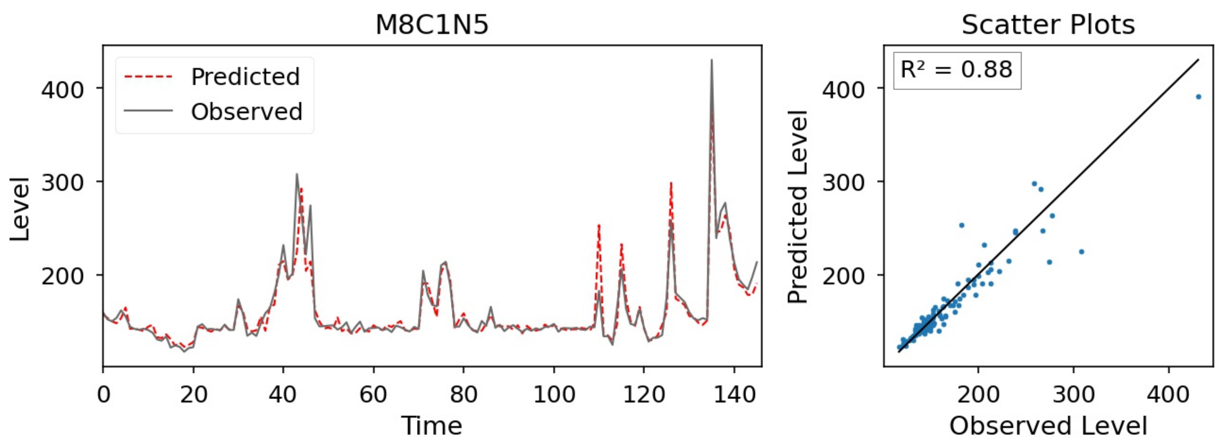
The architectures that generated the most good models were the C1N100, C2N10, C2N50 and C2N120. And ANNs with two hidden layers and 20 neurons in a layer, C2N20, were the least frequent among the best models with a coefficient of determination above 0.80.

4.2 Input variables

Table 4 presents the best ANN obtained for each input alternative presented in the Table 3. Next, the analysis of the contribution of the different features, the different lag days and the way the rainfall is represented are discussed.

4.2.1 Contribution of feature combinations

To analyze the contribution of the variables, we can observe the M7C2N120 network, which received only level data from the Fazenda Airis in $(t-1)$ and presented a good



(a) **(b)**

Figure 3: M8C1N5: (a) Predicted and observed values for the test dataset (b) Scatter plot.

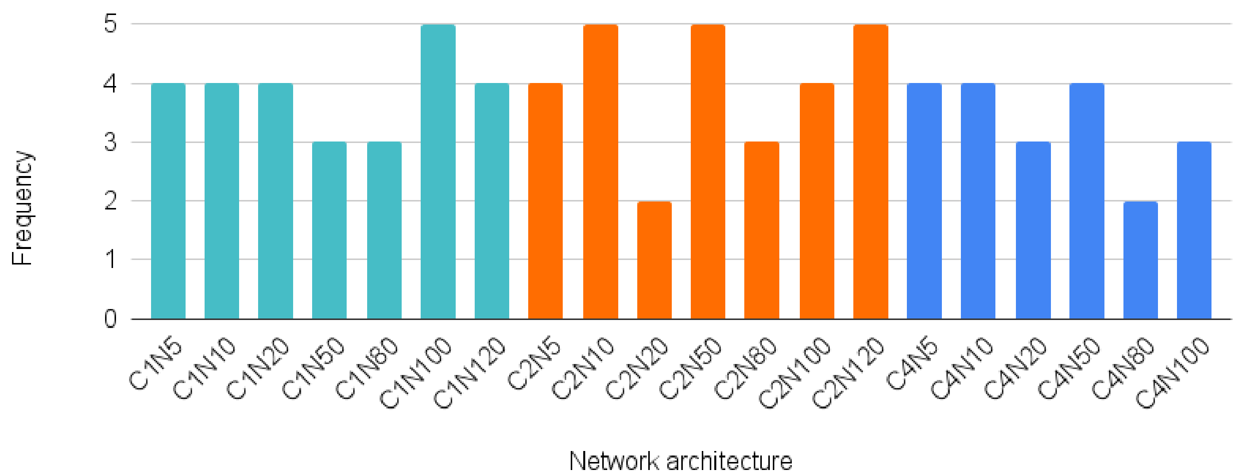


Figure 4: Frequency of architectures with coefficient of determination R^2 equal to or greater than 0.80.

performance with R^2 equal to 0.82 and with a standard deviation of 0.01. With the addition of average rainfall data, $\overline{P}$, we obtain the M5C4N20 model, which showed an improvement in the forecast with R^2 of 0.86 (0.07).

However, if the rainfall data were applied so that each rainfall station will be an input variable, we obtain the ANN M8C1N5. This ANN produced in the best prediction model, with R^2 equal to 0.88 with a standard deviation of 0.01.

Adding the data from the upstream and downstream fluviometric stations of the Fazenda Airis, we obtain M1C4N10. This change again produced a good prediction model, with a coefficient R^2 of 0.88 (0.04), but it did not represent a significant performance improvement.

Thus, Fazenda Airis station with a lag of one day is the most important attribute for predicting the model and that fluviometric data from other points did not significantly contribute to its efficiency.

Fig. 5 shows the frequency at which models of each input alternative obtained a coefficient of determination above 0.80. The input alternatives 1, 2, 5, and 8, which combine rainfall and fluviometric data, are the most frequent among ANNs with R^2 above 0.80. On the other hand, none of the models trained only with precipitation data had a coefficient of determination above 0.80 (alternatives 3 and 4). Therefore, we have from these results that the combination of precipitation and discharge data generates more accurate hydrological prediction models.

4.2.2 *Data lag*

Alternatives 1 and 2 receive the same variables with the difference in data lag time. Alternative 1 receives data on $(t-1)$, while alternative 2 input set receives data on $(t-1)$ and $(t-2)$.

In Table 4 we see that the addition of information

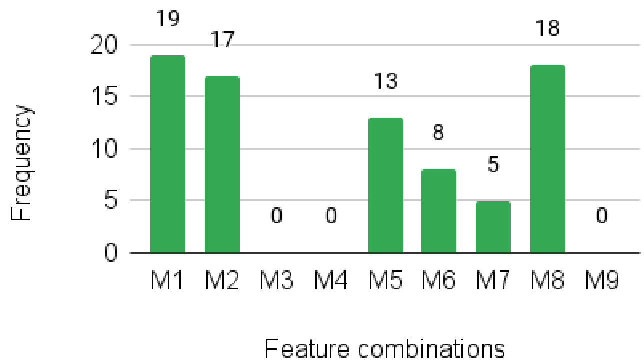


Figure 5: Frequency of input alternatives in models with coefficient R^2 above 0.80.

on rainfall and discharge with two days of lag did not contribute to the improvement of the forecast, since the M1C4N10 model had a higher coefficient in relation to the M2C1N5.

4.2.3 Processing of precipitation data

ANN models can receive rainfall data in different ways. Alternative 3 receives the rainfall data so that each rainfall station represents an input neuron. While in alternative 4, the ANN is fed by the average rainfall, calculated by the arithmetic mean. We see that the ANN M3C1N20 with a coefficient of R^2 of 0.49 had a better performance than the M4C4N20, which had a coefficient of 0.45.

The result obtained can be explained by the weights attributed by the ANNs to each variable during training. So that by the distribution of rainfall data from alternative 3, it is possible that the back-propagation algorithm assigns the weights according to the contribution of the station, generating better forecasts. It is noteworthy that there are other methods of obtaining the average rainfall and that these were not addressed in this work.

4.3 Comparison with Related Work

In this section, the results obtained are compared with related works that applied models based on Artificial Neural Networks for hydrological forecasting in Brazil in different temporal bases.

For each work, Table 5 presents the length of the historical series, the input and output variables, together with the metrics of the best model obtained. At the end, the data from this work were entered for comparison purposes.

Some performance metrics chosen by the authors were not addressed in this work, but they are comparable. The coefficient of determination R^2 is similar to the Nash-Sutcliffe (NS) performance coefficient when analyzing the model's fit, as discussed in Dawson and Wilby (2001). Meanwhile, Pearson's correlation coefficient (r_p) determines the degree of linear correlation between predicted and observed values, which may vary between -1 and 1. The other metrics MAE, MSE, MAPE and RMSE are error measures, where MAPE refers to the Absolute Percentage Mean Error and RMSE is the Root Mean Squared Error.

Some authors investigate the variation of input data (Machado, 2005, Sousa and de Sousa, 2010, Matos et al., 2014, Debastiani et al., 2016, Celeste et al., 2014, Barino and Bessa, 2020). But it is possible to observe that, for the most part, the best ANN data sets combine rainfall and fluviometric data, agreeing with the conclusions obtained in this work. Regarding the extent of the data, this work has the smallest series for forecasting on a daily basis among the studies presented in Table 5.

When comparing the results obtained in this study with the other works cited in Table 5, it is possible to notice that the proposed models had a satisfactory performance, being superior to other works that have a more complete historical series. However, we observe that it is possible to generate models with even more efficient forecasts, encouraging the continuity of this theme.

5 Conclusion

The goal of this work was to evaluate models based on Artificial Neural Networks for hydrological forecasting on a daily scale in the Macaé River drainage basin on the coast of the State of Rio de Janeiro. The database is composed of 6 rainfall stations and 6 fluviometric stations located in the Macaé river basin. The object of forecast was the daily level at the Fazenda Airis monitoring station and the study of the influence of the architecture and the training set were specific objectives of this work.

Different combinations of input variables were proposed. Furthermore, the number of hidden layers and number of neurons were varied. A total of 189 ANNs were trained, where 42.3% of the models obtained the correlation coefficient R^2 above 80%. These results are considered satisfactory or superior to other related works. The best model of ANN found, M8C1N5, received level data from Fazenda Airis and data from 6 rainfall stations with a time lag of one day and R^2 of 0.88 with a standard deviation of 0.01.

With regard to the feature combinations, it was observed that the performance of the ANNs was highly dependent on the input variables, so that the combinations of flow and precipitation data were more effective, with coefficient R^2 between 0.86 and 0.88 and MAE between 6.34 and 7.04 cm. It was also observed that models trained only with rainfall data did not obtain good fits, with R^2 between 0.45 and 0.49. The variable that was most relevant in the forecast was the daily level of the Fazenda Airis with a one-day lag. Having only this input variable, the ANN M7C1N180 produced a good fit to the test data and obtained R^2 equal to 0.84 (0.03).

The results obtained in the experiments encourage the use of the ANNs technique in hydrological modeling, even in basins where monitoring and data extent is limited, considering that the historical series of this work was less than 3 years. The results reinforce the relevance of the investment in the simultaneous monitoring of rainfall and the level of urban rivers and upstream stretches of urban areas, in order to develop more effective forecasts. So that such tools help in decision-making by public authorities

Table 5: Related works that uses ANN for hydrological forecasting with input set variables, length of the historical series, forecast model output and metrics.

Author	Features	Period	Output	Metrics
Machado (2005)	rainfall, discharge evapotranspiration	1976 a 1994	monthly discharge	r_p = 0.886 RMSE = 8.11
Batista (2009)	discharge	1990 a 2007	monthly discharge	MAPE = 0.59 MSE = 779.59
Sousa and de Sousa (2010)	rainfall, discharge	1964 a 2003	monthly discharge	R^2 = 0.92 RMSE = 8.29
Matos et al. (2014)	rainfall, discharge	1989 a 1994	daily discharge	NS = 0.97
Celeste et al. (2014)	rainfall, discharge	1969 a 1979	monthly discharge	NS = 0.82
da Silva et al. (2016)	rainfall, discharge	1986 a 1989	daily discharge	NS = 0.77 RMSE = 1.09
Debastiani et al. (2016)	rainfall, discharge	1997 a 1999	daily discharge	R^2 = 0.868 RMSE = 14.29
de Lima et al. (2016)	rainfall, stage	2013 a 2014	hourly stage	NS = 0.982 MAE = 0.69 RMSE = 1.50
da Silva et al. (2018)	rainfall sea surface temperature	1979 a 2016	monthly stage	R^2 = 0.999 MSE = 0.001
Cristaldo et al. (2018)	rainfall, stage	1995 a 2014	daily stage	NS = 0.93 MAE = 0.71
Gorodetskaya et al. (2018)	rainfall, discharge	2000 a 2016	daily discharge	R^2 = 0.917 MAPE = 0.108
Maraes et al. (2019)	stage, climate indices	1951 a 2017	monthly stage	r_p = 0.96
dos Santos Neto et al. (2020)	sea surface temperature, atmospheric pressure	2011 a 2016	monthly stage	R^2 = 0.845 RMSE = 0.233
Barino and Bessa (2020)	discharge, turbidity	2016 a 2019	monthly discharge	MAPE = 0.23
Mendonça et al. (2021)	rainfall, discharge	2009 a 2019	daily discharge	R^2 = 0.990 RMSE = 13.21 MAPE = 0.044
Alberton et al. (2021)	rainfall, stage	2009 a 2019	hourly stage	R^2 = 0.996 MAE = 4.82 MSE = 59.65
This work	rainfall, stage	2011 a 2013	daily stage	R^2 = 0.88 MAE = 7.03 MSE = 178.97

and in water resources planning.

As future works, it is recommended to implement the ANNs technique for hydrological modeling on a monthly and hourly scale, aiming at the use of forecasts in other sectors of water planning in the region. It is also possible to address the influence of new hydrological variables, such as the estimation of infiltration, soil moisture, effective rainfall and potential evapotranspiration in combination with the verification of attribute verification techniques. Another possible approach would be to explore the implementation of Recurring Neural Networks algorithms, widely used for time series data.

References

Alberton, G. B., Severo, D. L. and de Melo, M. N. V. (2021). Aplicação de redes neurais artificiais para previsão de enchentes no rio itajaí-açu em blumenau, sc, brasil, *Revista Ibero-Americana de Ciências Ambientais* **12**(4): 686–696. https://doi.org/10.6008/CBPC2179-6858.2021.004.0053.

Barino, F. O. and Bessa, A. d. S. (2020). Rede neural convolucional 1d aplicada à previsão da vazão no rio madeira, *XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, pp. 1–5. https://doi.org/10.14209/SBRT.2020.1570640893.

Batista, A. L. F. (2009). *Modelos de séries temporais e redes neurais artificiais na previsão de vazão*, PhD thesis, Engenharia de sistemas. Universidade Federal de Lavras. Available at http://repositorio.ufla.br/jspui/handle/1/1918.

Celeste, A. B., Menezes, T. and Fontes, V. (2014). Modelagem do processo chuva-vazão pelo tank model e por redes neurais artificiais na bacia do rio Japaratuba, Sergipe, *Scientia Plena* **10**(7). Available at https://www.scientiaplena.org.br/sp/article/view/1802/1005.

Cristaldo, M. F., Souza, C. C. d., Jesus, L. d., Oliveira, P. T. S. d., Padovani, C. R. and Viganó, H. H. d. G. (2018). Previsão probabilística de enchentes para uma pequena bacia hidrográfica do pantanal, *Revista Ambiente & Água* **13**(4). https://doi.org/10.4136/ambi-agua.1988.

da Silva, A. G., Castro, A. R. G. and Vieira, A. C. (2018). Modelo de previsão hidrológica utilizando redes neurais artificiais: um estudo de caso na bacia do rio xingu - Altamira-PA, *Revista Brasileira de Computação Aplicada* **10**(3): 55–62. https://doi.org/10.5335/rbca.v10i3.8779.

da Silva, M. R., Santos, L. B. L., Scofield, G. B. and Cortivo, F. D. (2016). Utilização de redes neurais artificiais em alertas hidrológicos: Estudo de caso na bacia do rio claro em Caraguatatuba, estado de São Paulo, *Anuário do Instituto de Geociências* **39**(1): 23–31. https://doi.org/10.11137/2016_1_23_31.

Dawson, C. and Wilby, R. (2001). Hydrological modelling using artificial neural networks, *Progress in physical Geography* **25**(1): 80–108. https://doi.org/10.1177/030913330102500104.

de Lima, G. R., Santos, L. B., de Carvalho, T. J., Carvalho, A. R., Cortivo, F. D., Scofield, G. B. and Negri, R. G. (2016). An operational dynamical neuro-forecasting model for hydrological disasters, *Modeling Earth Systems and Environment* **2**(2): 1–9. http://dx.doi.org/10.1007/s40808-016-0145-3.

de Lima, G. R. and Scofield, G. B. (2017). A machine learning-based model to improve shortterm forecasts of flooding in Nova Friburgo-RJ, *Revista Brasileira de Cartografia* **69**(1). Available at http://www.seer.ufu.br/index.php/revistabrasileiracartografia/article/view/44030.

de Lima, G. R. T. and Scofield, G. B. (2021). Estudo de viabilidade do uso operacional de redes neurais em um sistema de alertas antecipados de enxurradas, *Revista Brasileira de Recursos Hídricos* **26**(7): 1–11. https://doi.org/10.1590/2318-0331.262120200152.

de Mello, Y. R., Kohls, W. and de Oliveira, T. M. N. (2017). Uso de diferentes métodos para o preenchimento de falhas em estações pluviométricas, *Boletim de Geografia* **35**(1): 112–121. https://doi.org/10.4025/bolgeogr.v35i1.30893.

Debastiani, A. B., Silva, R. D. d. and Rafaeli Neto, S. L. (2016). Eficácia da arquitetura MLP em modo closed-loop para simulação de um sistema hidrológico, *Revista Brasileira de Recursos Hídricos* **21**(4): 821–831. https://doi.org/10.1590/2318-0331.011615124.

dos Santos Neto, L. A., Maniesi, V., Querino, C. A. S., da Silva, M. J. G. and Brown, V. R. (2020). Modelagem hidroclimatologica utilizando redes neurais multi layer perceptron em bacia hidrográfica no sudoeste da Amazônia, *Revista Brasileira de Climatologia* **26**. https://doi.org/10.5380/abclima.v26i0.73007.

dos Santos Sousa, R., Guedes, E. B. and de Oliveira, M. B. L. (2018). Previsão anual de precipitações em Manaus, Amazonas: Um comparativo de técnicas de aprendizado de máquina, *Anais do IX Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais*, SBC. https://doi.org/10.5753/wcama.2018.2929.

Freire, A. L., de Menezes Jr, J. M. P. and Barreto, G. A. (2009). Redes neurais recorrentes para predição recursiva de séries temporais caóticas: um estudo comparativo, *IX Congresso Brasileiro de Redes Neurais* pp. 25–28. http://dx.doi.org/10.21528/CBRN2009-009.

Gorodetskaya, Y., da Fonseca, L., Tavares, G. and Ribeiro, C. (2018). Daily streamflow forecasting for paraíba do sul river using machine learning methods with hydrologic inputs, *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, SBC, Porto Alegre, RS, Brasil, pp. 162–173. https://doi.org/10.5753/eniac.2018.4413.

Ikemoto, S. M., Oliveira, E., Cardoso, C. D. and Machado, T. (2015). Redução de risco de desastres associados a inundações, *Boletim águas & território*, Instituto Estadual do Ambiente (INEA). Available at http://www.inea.rj.gov.br/wp-content/uploads/2018/12/N%C2%BA-10-Redu%C3%A7%C3%A3o-de-Risco-de-Desastres-Associados-a-Inunda%C3%A7%C3%B5es.pdf.

Machado, F. W. (2005). *Modelagem chuva-vazão mensal utilizando redes neurais artificiais*, PhD thesis, Dissertação de Mestrado. Universidade Federal do Paraná. Available at https://hdl.handle.net/1884/2345.

Maraes, P., Costa, M. et al. (2019). Previsão de cheias sazonais para o rio negro usando redes neurais artificiais e árvores de decisão com aprendizado baseado em comitê, *Anais do X Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais*, SBC, pp. 47–56. https://doi.org/10.5753/wcama.2019.6419.

Matos, A., Pedrollo, O. C. and Castro, N. (2014). Efeito do controle de montante de sub-bacias embutidas na previsão hidrológica de curto prazo com redes neurais: Aplicação à bacia de ponte mística, *Revista Brasileira De Recursos Hídricos* **19**(1): 87–99. https://doi.org/10.21168/rbrh.v19n1.p87-99.

Mendonça, L. M., Gomide, I. d. S. and Blanco, C. J. C. (2021). Modelagem chuva-vazão via redes neurais artificiais para simulação de vazões de uma bacia hidrográfica da amazônia, *Revista de Gestão de Água da América Latina* **18**(2): 1–14. https://doi.org/10.21168/rega.v18e2.

Rodrigues, M., Costa, M. and Cícero Filho, C. (2015). Proposta de um método para previsão de cheias sazonais utilizando redes neurais artificiais: Uma aplicação ao rio amazonas, *Anais do VI Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais*, SBC, pp. 1–10. https://doi.org/10.5753/wcama.2015.10184.

Sarmento, L. (2021). Monitoramento hidrometeorológico no BRASIL: uma análise sob a ótica da coordenação de políticas públicas, *Revista de Gestão de Água da América Latina* **18**: 1–15. https://doi.org/10.21168/rega.v18e3.

Shamseldin, A. Y. (2009). Artificial neural network model for river flow forecasting in a developing country, *Journal of Hydroinformatics* **12**(1): 22–35. https://doi.org/10.2166/hydro.2010.027.

Sousa, W. d. S. and de Sousa, F. d. A. (2010). Rede neural artificial aplicada à previsão de vazão da bacia hidrográfica do rio piancó, *Revista Brasileira de Engenharia Agrícola e Ambiental* **14**(2): 173–180. https://doi.org/10.1590/S1415-43662010000200008.

Tucci, C. E. and Silveira, A. (2002). Gerenciamento da drenagem urbana, *Revista Brasileira de Recursos Hídricos* **7**(1): 5–27. https://doi.org/10.21168/rbrh.v7n1.p5-27.