ACTA IMEKO
ISSN: 2221-870X
December 2023, Volume 12, Number 4, 1 - 9

Machine learning models applied to estimate the water temperature of rivers and reservoirs

Jheklos G. Silva[1], Ricardo A. C. Souza[2], Obionor O. Nóbrega[2]

[1] *Graduate in Applied Informatics, Federal Rural University of Pernambuco, Rua Dom Manoel de Medeiros, 52.171-900, Recife, Brazil*
[2] *Departament of Computer Science, Federal Rural University of Pernambuco, Rua Dom Manoel de Medeiros, 52.171-900, Recife, Brazil*

ABSTRACT
Water temperature in rivers and reservoirs plays a crucial role in aquatic ecology, as inadequate conditions can promote the overgrowth of harmful algae and bacteria, resulting in the production of harmful toxins for human and animal health, and affecting water quality. To effectively manage water resources, continuous monitoring of these bodies is crucial. However, existing technological devices rarely offer continuous and real-time data collection, necessitating an alternative approach. The aim of this study was to compare the performance of four machine learning models (Linear Regression, Stochastic Model, Extra Tree, and Multilayer Perceptron Neural Network) in estimating water temperature in Pernambuco, Brazil's rivers and reservoirs. Statistical metrics showed that all models achieved a satisfactory capacity, with the Multilayer Perceptron Neural Network demonstrating slightly superior performance in reservoirs and rivers where it obtained the best result with a Mean Squared Error: 0.343, Root Mean Squared Error: 0.585, Mean Absolute Error: 0.445 and Coefficient of Determination: 0.595. Consequently, the MLPNN model was chosen for the development of virtual sensors. In addition to an interface that allows users to access a map and obtain estimated water temperature information for various locations, facilitating informed decision-making and resource management.

Section: RESEARCH PAPER

Keywords: measurement water temperature; machine learning; neural networks; statistical models

Citation: Jheklos G. Silva, Ricardo A. C. Souza, Obionor O. Nóbrega, Machine learning models applied to estimate the water temperature of rivers and reservoirs, Acta IMEKO, vol. 12, no. 4, article 23, December 2023, identifier: IMEKO-ACTA-12 (2023)-04-23

Section Editor: Laura Fabbiano, Politecnico di Bari, Italy

Received June 21, 2023; **In final form** November 8, 2023; **Published** December 2023

Funding: This research was financed by Coordination for the Improvement of Higher Level Personnel (CAPES). This work was also supported by Pernambuco Water and Climate Agency (APAC)

Corresponding author: Jheklos G. Silva, e-mail: jheklos.silva@ufrpe.br

1. INTRODUCTION

Water is a vital resource for human survival and socioeconomic development, rivers and reservoirs play a fundamental role in society, such as public supply, agricultural irrigation, hydroelectric power generation and industrial activities. Monitoring water temperature in reservoirs and rivers is an important practice to understand and assess environmental conditions and the effects of climate change and human activities on these environments. The continuous monitoring of water temperature is essential for understanding seasonal patterns, daily fluctuations, and long-term trends, providing valuable information for water resource management and the conservation of aquatic ecosystems [1]. These data are essential for mitigating possible negative effects on biodiversity and ecological balance.

The lack of monitoring of water temperature in water bodies, such as rivers and reservoirs, can lead to a series of problems and adverse impacts on aquatic ecosystems and human activities related to water, including water quality and its impact on aquatic life. Significant changes in temperature can affect several physicochemical parameters of the water, which in turn influences reproduction, growth and, in extreme cases, can lead to increased mortality of aquatic animals. Furthermore, Inadequate water temperatures caused by climate change can have substantial impacts on the spread of waterborne diseases, such as bacterial proliferation [2]. This increases the risk of waterborne diseases, posing a challenge to public health.

The methods employed for monitoring water quality cover a spectrum ranging from traditional manual methods to the most advanced, based on automated technology, which use wireless sensor networks [3]. Manual methods require immersing

thermometers in water to obtain temperature readings, while automatic methods involve the use of probes or temperature sensors that are installed at fixed points and transmit the data to recording stations. However, it is important to point out that, in practice, such devices are not widely present in most rivers or reservoirs. Measurements are often performed manually, but not continuously, resulting in time gaps in monitoring.

In view of this, it becomes necessary to seek alternative methods to estimate the water temperature in such environments so that it is possible to monitor it. A promising approach for monitoring water temperature is the application of statistical models based on machine learning, which are able to estimate the temperature based on historical data and relevant environmental variables. There are several applications and different machine learning models for predicting water temperature widely used in many fields of study [4]. These models are able to analyse large datasets and identify complex patterns, which can help to estimate water temperature at different times and locations. Statistical machine learning models such as artificial neural networks and regression models have shown good performance in predicting water temperature in different bodies of water. They can capture the complexity of the data and provide accurate estimates [5].

The two most used categories are deterministic models and statistical models. While process-based deterministic models provide a clear physical understanding of the energy balance and have their merits, they are often complex and require multiple inputs. Statistical models, on the other hand, allow for a more simplified approach, using the statistical relationships between the variables to estimate the water temperature efficiently [6].

Several researches were conducted using machine learning models to estimate water temperature in different contexts. Exploring different approaches to estimating river water temperature, considering the relationship with air and water temperature [7]. Although these researches have focused mainly on temperature prediction in rivers and reservoirs, as far as we know, there are no studies that explore the use of these trained models to develop virtual water temperature monitoring sensors in these environments, considering related meteorological variables.

In this sense, the following research question arises: "how to use statistical machine learning models to estimate water temperature in rivers and reservoirs, by combining historical data and air temperature, in order to develop an effective monitoring web application through virtual sensors?" This issue is motivated by the need to continuously monitor water temperature, aiming at proper management of water resources, preventing the growth and proliferation of pathogenic organisms, such as harmful bacteria and algae, and preserving water quality, with impacts positive effects on human health and the aquatic ecosystem.

Given this context, the objective of this research is to investigate a variety of statistical models using only air and water temperature variables, training them and evaluating their performance through validation tests. The final purpose is to compare the accuracy of the obtained results, in order to determine the most adequate model and approach to estimate the water temperature in rivers and reservoirs. From this, it is intended to develop a web application capable of providing reliable estimates of water temperature in real time. It is expected that this study will contribute significantly to the management of reservoirs and rivers, providing valuable information for informed decision-making. In addition, it is expected that the results of this research can inspire and drive theoretical advances, innovative methods and practical solutions in other areas of research, in addition to providing relevant empirical evidence for the scientific community.

The work is organized into several sections after this introduction. Section 2 discusses the related works that served as the basis for the development of the research. Then, in section 3, the materials and methods used are described. In section 4, we present the models that were adopted in the study. The metrics to evaluate the performance of the models are shown in section 5. Section 6 discusses the technologies used to develop the virtual sensors interface. The results of model validation are presented in section 7. Finally, section 8 brings the conclusion covering the main results and conclusions of this work.

2. RELATED WORK

Machine learning algorithms have stood out as fundamental components in digital solutions, attracting considerable attention in the digital area, in addition to being frequently used to assist in effective decision-making [8]-[9] They are widely used for measuring physical properties, exhibiting good predictive performance [10]. Machine learning has proven to be an efficient way of monitoring in several applications and sectors [11]-[12] This monitoring is performed through virtual sensors as an alternative to traditional physical sensors, which may have several limitations and costs [13].

Through virtual sensors, it becomes feasible to develop low-cost sensors with high precision, based on machine learning. This enables precise mapping of variables that, otherwise, could be costly when using traditional sensors [14]. In [15], a review of several existing virtual applications was carried out, focusing on virtual sensors and calibration of these devices, aiming at their application in environments that demand reliable sensing.

The use of virtual sensors as an alternative to physical sensors, aiming at reducing hardware costs, was explored in [16]. The authors used real data from sensor devices to compare it with virtual data using machine learning techniques. The result of this work was the implementation of a prototype designed to measure lighting attributes in industrial applications.

In [17], the authors propose a virtual sensor approach based on models such as computational fluid dynamics (CFD) for temperature monitoring in greenhouses. The main objective was to develop a real-time three-dimensional (3D) simulator using virtual sensors. A well-calibrated physical sensor was installed to collect and analyse the CFD. The quantitative result of the performance of the controller was that it reached the 25 °C set point in less than 45 seconds and maintained the desired temperature with an accuracy of ±0.3 °C. The result showed that virtual sensing can be applied in large greenhouses to monitor temperature, however, the 3D simulator requires hardware with good performance. The purpose in the present study is to offer an accessible interface for any type of device.

In [18], an approach was proposed to create virtual sensors in solar power plants using machine learning algorithms in order to replace defective sensors in an automated way. The authors employed IoT sensors and used the faulty sensor's historical data as predictors. The results showed that the linear regression obtained a MAE of 14299, the artificial neural network reached a MAE of 13922, and the Bayesian Ridge Regression presented a MAE of 14299. The main difference between this study and the current work is that, in solar energy, the algorithms are automatically trained using data from existing sensors so that when the physical sensors fail, the virtual sensors can take over

monitoring. On the other hand, in the present study, data collection was performed manually, depending exclusively on virtual sensors for monitoring.

In the context of estimating the volume of water in the soil, machine learning models are also employed. In [19], the authors used an IoUT sensing system with a low-cost soil moisture sensor. An augmented method was introduced that combines soil moisture sensor data with RSSI information to improve estimation accuracy. A comparison was also made with the virtual sensor, in which the increased sensor had a superiority (RMSE) of 1.84%. The main discrepancies with the present study are the sensor technologies using a set with a LoRaWAN transceiver for estimation and the application site with the objective of estimating the water volume.

The importance of monitoring systems to ensure sustainability and efficiency has been commonly explored. In [20], the authors analyse the state of the art of flexible electronics for IoT with a focus on sustainability and changes in design skills and tools, also recognizing the importance of monitoring as highlighted in this study. However, it does not mention specific evaluation metrics, as it does not focus on testing models but on analysing the state of the art and sustainability considerations.

Regarding water temperature monitoring, there is a lack of research that explores water temperature monitoring using virtual sensors. However, there are studies that apply machine learning techniques to estimate water temperature. In [21], the authors developed three distinct models of machine learning, namely the Artificial Neural Network (ANN), Gaussian Process Regression (GPR), and Aggregated Decision Trees with Bootstrap (BA-DT). In addition, standard models such as linear regression, non-linear regression, and stochastic models were developed and compared using different meteorological stations. The results obtained indicated that machine learning models are effective tools for predicting river water temperatures. In particular, for station No. 3, the best model was the GPR, which presented an RMSE of 1.4950, a correlation coefficient R of 0.9897, and a NSC coefficient of 0.9764.

It is worth noting that the study only evaluates the models and does not create a virtual sensor to estimate the river water temperature. It is therefore considered a research opportunity to propose the deployment of virtual sensors to compensate for the lack of physical sensors.

3. MATERIALS AND METHODS

3.1. Study site and data collection

Pernambuco, a state located in northeastern Brazil, have a significant network of rivers and important reservoirs. These bodies of water play essential roles, providing water resources to supply cities, rural communities and vital agricultural activities for irrigating crops, contributing to access to clean water, human and economic development in the region. In addition to being sources of hydroelectric power, they are also used for river transport, trade and tourism. In addition to economic and social benefits, these water bodies have significant ecological value, harbouring a variety of plant and animal species. They play an important role in preserving biodiversity and maintaining ecosystems.

In the present study, information was obtained from the rivers and reservoirs of Pernambuco, including their names and locations, provided by the Pernambuco Water and Climate Agency (APAC). Data referring to water temperature in rivers and air were provided by the same source. In total, 55 reservoirs and 45 rivers were considered for analysis. When examining the data, it was found that the measurements were performed manually in the period from 2011 to 2022, recording only water and air temperature at the time, date and place of collection, unfortunately other variables were not recorded. This data will be used as input and output for model training.

However, the datasets for each reservoir and river obtained were limited, with some of them containing less than 20 records. Given this limitation, it was decided to train the model using data from all rivers (for training rivers) and reservoirs (for training reservoirs), in order to obtain a more comprehensive and representative data set.

As a result, a total of 1919 records were used, 759 referring to rivers and 1160 to reservoirs. This approach allowed obtaining a more robust and suitable data set for model training, considering the limited availability of individual data for each body of water.

The next step consisted of calculating the correlation coefficient, with the aim of identifying the relationship between the two variables: air temperature and water temperature. This calculation makes it possible to determine the quantitative nature of this linear relationship, as well as its intensity and direction. The correlation coefficient can assume negative values, indicating an inverse or negative relationship, that is, when one variable increases, the other tends to decrease. Positive values, on the other hand, indicate a direct or positive relationship, that is, when one variable increases, the other tends to increase as well. A value of zero indicates no linear correlation between variables. For this purpose, Pearson's correlation coefficient formula was used, a statistical measure widely adopted to quantify the linear relationship between variables. Equation (1) represents the calculation performed to determine the value.

$$r = \frac{\sum_{i=1}^{n}[(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{n \cdot \sigma_x \cdot \sigma_y}, \tag{1}$$

where r is the correlation coefficient, Σ represents the sum, x and y are the values of the two variables, $\bar{x}$ and $\bar{y}$ are the means of the values of x and y, respectively, n is the number of pairs of observations. σ_x and σ_y are the standard deviations of variables x and y, respectively. The result of the calculation can be seen in Table 1.

From that point on, the data were divided into two distinct sets: one set comprising 80 % of the data, intended for the model adjustment phase, and another set comprising 20 % of the data, used for validating the model obtained after training. Before proceeding with the training and validation of the model, the normalization of the input and output data was performed using a technique that seeks to keep the variables within a specific range. This technique involves scaling the input and output data in order to obtain a normalized distribution, that is, with zero mean and unitary standard deviation. The equation (2) involves the operation of subtracting the mean of the data and dividing by the standard deviation.

$$x_{\text{norm}} = \frac{x - x_{\text{mean}}}{x_{\text{std}}}, \tag{2}$$

where x_{norm} is the normalized value, x represents the data value, x_{mean} represents the mean of the data set, and x_{std} represents the standard deviation of the data set.

Table 1. Air & water temperature correlation.

Rivers	Reservoirs
0.70	0.61

After completing the model training and validation process, it is essential to reverse the normalization applied to the data, in order to present them in the same structure as the original data set. This reversal step is extremely important to ensure that the results obtained by the model are correctly interpreted and can be compared with the original data. In this way, it is possible to obtain an accurate and contextualized analysis, considering the scale and characteristics of the original data.

4. WATER TEMPERATURE MODEL

4.1. Linear regression model

Linear regression is a statistical method used to model the relationship between a dependent variable (output) and one or more independent variables (or predictors). It is one of the simplest and most widely used methods in data analysis. One of the main advantages of linear regression is its ability to make predictions. Based on the patterns observed in the data, it is possible to estimate future values of the dependent variable with reasonable accuracy, providing valuable insights for planning and decision making [22]. Simple linear regression involves only one independent variable and one dependent variable. The general formula for simple linear regression is given by equation 3.

$$Y = \beta_0 + \beta_1 \cdot X_1 , \tag{3}$$

where Y is the dependent variable (output) that we are trying to estimate, β_0 is the constant intercept of the regression line, which represents the value of y when x is equal to zero, β_1 is the regression coefficient (or slope) which represents the mean change in the dependent variable associated with a unit change in the independent variable. and X_1 is the independent predictor variable. The objective of linear regression is to estimate the values of the β_0 and β_1 coefficients so that the regression line fits better to the observed data. This is done by minimizing the sum of squares of the residuals (or errors) between the observed values of y and the values predicted by the regression equation. The linear regression model can be expanded to incorporate several independent variables, however, in this study it is important to emphasize that we are limited to considering only the air temperature as a predictor variable. Therefore, based on this single variable, we will use the simple linear regression model to perform the analyses and obtain estimates.

4.2. Model Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a widely used approach in training deep learning models. This is the algorithm commonly used to solve optimization problems, adjusting the parameters iteratively through the learning rate [23]. The SGD formula, given by equation (4), is a variation of the classic gradient descent algorithm, which adjusts the parameters of a model to minimize a cost function.

$$W(t+1) = W(t) - \alpha \nabla w \left(J\big(W(t), x(i), y(i)\big) \right), \tag{4}$$

where $W(t+1)$ is the new value of the parameters at time $t+1$, $W(t)$ is the current value of the parameters at time t, α is the learning rate, which controls the size of the parameter update steps, $\nabla w \left(J\big(W(t), x(i), y(i)\big) \right)$ is the gradient of the loss function J with respect to the parameters $W(t)$, calculated based on a sample of data $\big(x(i), y(i)\big)$ specific. This formula represents the updating of model parameters using a stochastic approach, where the updating is done iteratively for each data sample individually. On each iteration, the loss function gradient is calculated based on this sample and used to adjust the parameters.

4.3. Model Extra Tree Regressor

Extra Tree Regressor is a machine learning algorithm based on decision trees that stands out for its efficiency and capacity to deal with regression problems. Is a relatively recent machine learning technique, which was proposed as an extension of the Random Forest algorithm and it was developed to further improve the performance and generalization of the model in relation to the Random Forest [24]. The main feature of Extra Tree Regressor is that it introduces additional randomness in the construction of decision trees, making them even more diverse. While traditional decision trees split the data at each node based on optimal attribute values, the Extra Tree Regressor randomly selects the split points and chooses the best one among them.

The formula used to build each individual tree in the set is similar to the conventional decision tree algorithm formula. Building each individual tree in the decision tree algorithm involves splitting the data based on criteria such as entropy given by equation (5) and information gain calculated by equation (6). Entropy is a measure of impurity or disorder in the data. The greater the entropy, the greater the disorder and uncertainty in the data.

$$Entropy = \sum_{i=1}^{c} -p_i \cdot \log_2 p_i , \tag{5}$$

where p_i is the probability of occurrence in the data set x belonging to the variable x, $\log_2$ is the logarithm in base 2. The information gain measures the reduction from the entropy calculation after the division of the data based on a predictor variable. It is used to determine the best variable to split the dataset and build the decision tree.

$$Gain\ (S, A) = Entropy(s) - H(S|A), \tag{6}$$

where $Entropy(s)$ represents the entropy measure of the random variable Y, and $H(S|A)$ is the conditional entropy of S given the variable A. These formulas are fundamental for the construction of decision trees and the determination of the best divisions in the data based on impurity and information gain criteria.

4.4. Multilayer perceptron neural network (MLPNN)

Multilayer Perceptron Neural Network represents a forward-feed type neural network architecture, based on the learning technique known as Backpropagation, structured and composed by an input layer of neurons that play the role of receptors, one or more hidden layers of neurons that perform iterative calculations with the data, and, finally, the output layer is responsible for predicting the final results of the network [25]. The MLPNN consists of several layers of neurons, including an input layer, one or more hidden layers, and an output layer. Each neuron in a MLP network is called a perceptron and operates similarly to a biological neuron. It receives weighted inputs, the network processes it and applies an activation function and generates an output. This process is called feed forward presented by the equation (7)

$$U = \sum_{i=1}^{N} x_i \cdot w_i + b , \tag{7}$$

where U is the output generated by the perceptron, Σ represents the weighted sum of the inputs multiplied by the corresponding synaptic weights, x_i is the vector of inputs, w_i is the vector of synaptic weights, b is the bias, an additional term that allows adjust the perceptron output.

After performing this operation, the output is added to a non-linear activation function, in this work the function ReLU (Rectified Linear Unit) determined by the equation (8) was used. ReLU returns the input value if it is positive, otherwise it returns zero.

$$ReLU(x) = \max(0, x)\,, \tag{8}$$

where $ReLU(x)$ is the output generated by the ReLU function, x is the input value to the function. During the training of an MLP network, the synaptic weights and biases are updated based on the backpropagation algorithm. The objective is to minimize a cost function, which measures the difference between the outputs predicted by the network and the desired outputs. The formula for updating the weights and biases using the backpropagation algorithm involves using gradient descent, which looks for the steepest descent direction in the cost function.

5. MODEL PERFORMANCE AND EVALUATION

In order to evaluate the performance of the statistical models, four different criteria were used: the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), the Absolute Mean Error (MAE) and the Coefficient of Determination (R^2).

5.1. MSE

The MSE, represented by equation (9), is a measure that quantifies the difference between the model's estimates and the actual observed values, which indicates the magnitude of this difference. The smaller the MSE value, the better the model's performance in terms of prediction accuracy.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2\,, \tag{9}$$

where n represents the number of observations, the term sum of squared errors indicates the sum of the squared differences between the predicted values of f_i and the observed values of y_i.

5.2. RMSE

The RMSE calculated by the equation (10) is a metric derived from the MSE in which the square root is applied at the end, resulting in an error measure expressed in the same unit as the target variable. The lower the RMSE value, the better the model's performance in terms of prediction accuracy.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - O_i)^2}\,, \tag{10}$$

where n represents the number of observations, the term sum of squared errors indicates the sum of the squared differences between the predicted values by f_i and the observed values O_i.

5.3. MAE

The MAE determined by the equation (11) represents the average of the absolute values of the errors between the model estimates and the actual values of the variable of interest. A smaller MAE value indicates better model performance in terms of prediction accuracy.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{11}$$

Where n represents the number of samples, the term y_i corresponds to the true value of the i-th sample and $\hat{y}_i$ represents the value predicted by the model for the i-th sample.

5.4. R^2

The coefficient of determination (R^2) presented by the equation (12) is a statistical metric that provides an indication of how well the regression model fits the observed data. In general terms, a higher value of R^2 indicates a superior performance of the model, that is, a better ability to explain the variation in the data.

$$R2 = \frac{SSE}{SST},$$
$$SSE = \sum_{i=1}^{n}(f_i - \hat{y})^2\,, SST = \sum_{i=1}^{n}(f_i - \bar{y})^2, \tag{12}$$

where SSE is the sum of squared errors and SST is the total sum of squares.

6. TECHNOLOGIES USED FOR INTERFACE DEVELOPMENT

In order to make the research results available through an interactive map that allows the user to obtain, with just one click, the estimated temperature for a reservoir or river at any time and place, it is necessary to develop a web prototype using HTML markup.

The focus of development is the creation of a map interface that presents points of location of water bodies. To facilitate this process, MapTiler [26], was used, a mapping and geocoding platform that provides access to map services, satellite imagery, and other related resources through an application programming interface (API). This API allows developers to easily integrate interactive maps and geolocation features into their apps and websites.

Inside the HTML code, the trained algorithm with the best accuracy to estimate the water temperature will be incorporated. The JavaScript programming language will be used for this purpose. The input data, which is the air temperature, will be acquired through a web API called Visual Crossing [27], as APAC still does not provide an API that allows obtaining the local air temperature at the time of access.

Thus, by combining these technologies and services, it will be possible to create an interactive interface that provides users with accurate and real-time information about water temperature in different places of interest.

7. RESULTS

In this section, the results of the validation tests of each model will be presented, taking into account the statistical indices used to evaluate the performance of the models. These indices provide objective metrics that allow a quantitative and comparative analysis of the results obtained and identify which has the best performance in terms of accuracy and predictability.

7.1. Result of rivers validation tests

During the training process of the data collected in the rivers, the neural network model MLPNN presented a superior performance compared to the other algorithms, followed by the linear regression model and the Stochastic. This finding is supported by the results obtained in the validation tests, which are represented in Table 2 and based on relevant statistical indices.

It is observed that in relation to the mean squared error (MSE), all models presented close results, ranging from 0.343 to 0.357. This indicates that the models have a good ability to estimate the correct values in relation to the observed values. The same pattern is observed for the root mean square error (RMSE), where differences between models are minimal, ranging from 0.585 to 0.598. Figure 1 illustrates plots stemming from the data presented in Table 2.

With regard to the absolute mean error (MAE), the Perceptron model (MLPNN) obtained the lowest value (0.445), indicating better precision in estimating the values compared to the other models. Regarding the coefficient of determination (R^2), which measures the proportion of data variability explained by the model, the Perceptron model (MLPNN) presented the highest value (0.595), indicating a better predictive capacity in relation to the other models. Based on the results of the validation test, we can conclude that the Perceptron model (MLPNN) showed the best overall performance, with the lowest mean absolute error and the highest coefficient of determination. However, it is important to highlight that all evaluated models presented similar results and showed a satisfactory capacity to estimate the observed values. Figure 2 shows the observed and estimated data from the validation test.

7.2. Result of reservoirs validation tests

During the training of the data collected in the reservoirs, once again the MLPNN neural network demonstrated a slight superiority in relation to the other models, followed by the linear regression model and the Stochastic. However, this time, even

Table 2. Performance of models with data collected in rivers.

Models:	*MSE*	*RMSE*	*MAE*	R^2
Perceptron:	0.343	0.585	0.445	0.595
Linear Regression:	0.351	0.592	0.460	0.586
Model Stochastic:	0.350	0.592	0.459	0.586
Model Extra Tree:	0.357	0.598	0.477	0.578

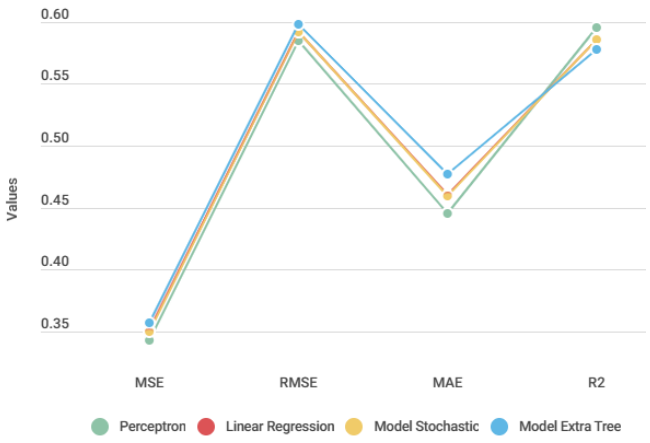

Figure 1. Plots of the statistical metrics of the validation test in the rivers.

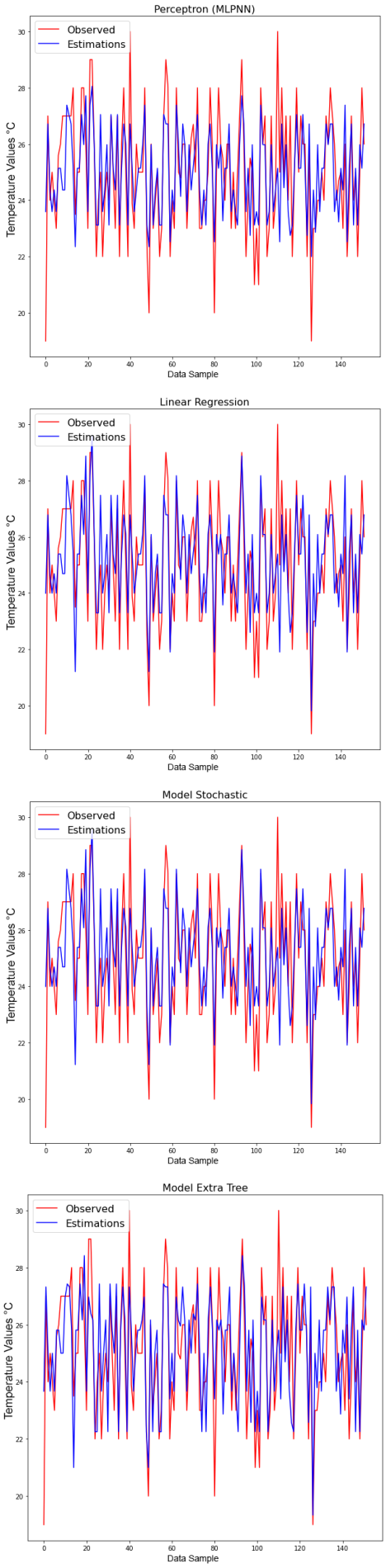

Figure 2. Estimates and observations for the river validation set.

with a larger data sample in the reservoirs, the results showed a lower accuracy compared to the data collected in the rivers.

A possible explanation for this difference in performance can be found in the previous section analysing the correlation calculations between air temperature and water temperature in the two types of water bodies. It was observed that the positive correlation between these variables in rivers is stronger than in reservoirs, which may have contributed to a better performance of the models on river data. This happens because, in cases with strong correlations, the patterns and relationships between the data become more evident and, consequently, are more easily learned by the model. Furthermore, in the case of reservoir data, the presence of noise, such as outliers, had a negative effect on model performance. On the other hand, in the river data, the dataset was smaller and contained less noise, which allowed the model to focus on the most important patterns. Table 3 presents the results of the statistical indices obtained in the reservoirs, providing an overview of the performance of the models in this specific context.

The results in Table 3 show that in terms of mean squared error (MSE) and root mean squared error (RMSE), the models showed results close to each other, ranging from 0.467 to 0.499 for the MSE and from 0.683 to 0.707 for the RMSE. This indicates that the models showed a moderate ability to estimate the correct values in relation to the values observed in the reservoirs.

With regard to the absolute mean error (MAE), the Perceptron model (MLPNN) obtained the lowest value (0.544), indicating better precision in estimating the values compared to the other models. Regarding the coefficient of determination (R^2), which measures the proportion of data variability explained by the model, all models showed low values, ranging from 0.310 to 0.356. This observation is evident in Figure 3, that illustrates plots stemming from the data presented in Table 3. Therefore, the models had difficulty in capturing the data variation in the reservoirs.

Table 3. Performance of models with data collected in reservoirs.

Models:	***MSE***	***RMSE***	***MAE***	***R²***
Perceptron:	0.467	0.683	0.544	0.356
Linear Regression:	0.493	0.702	0.559	0.320
Model Stochastic:	0.492	0.701	0.559	0.321
Model Extra Tree:	0.499	0.707	0.572	0.310

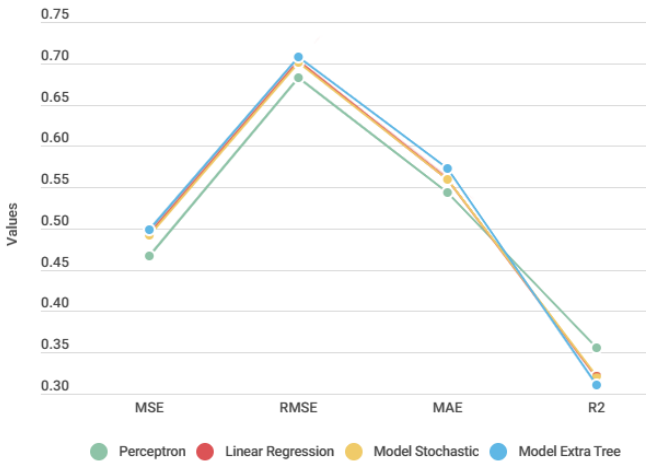

Figure 3. Plots of the statistical metrics of the validation test in the reservoirs.

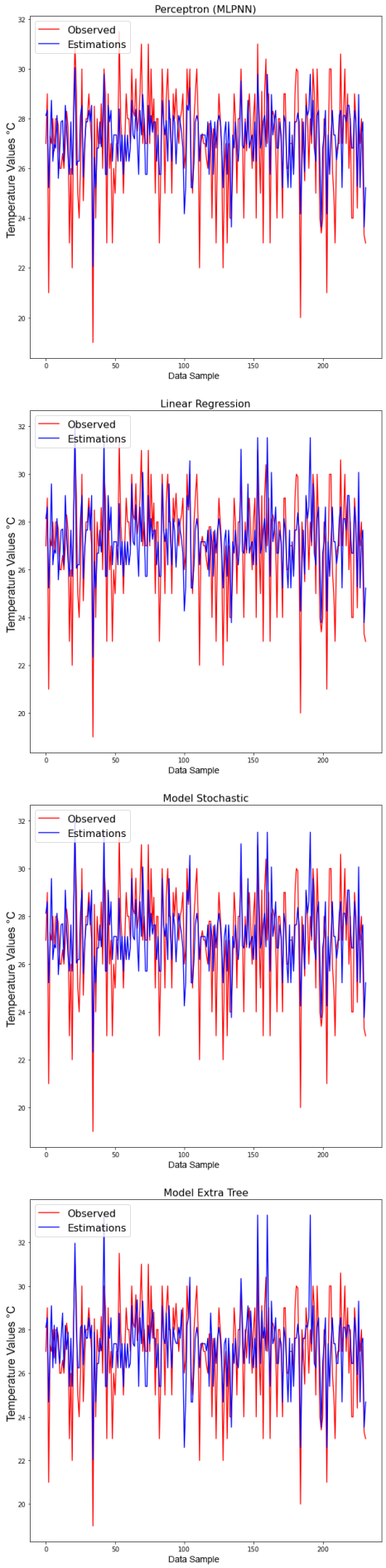

Figure 4. Estimates and observations for the reservoirs validation set.

Based on the results of the validation test, we can conclude that the models performed less well when dealing with reservoir data compared to river data. The Perceptron model (MLPNN) demonstrated a slight superiority in terms of lower mean absolute error. However, it is important to emphasize that all the evaluated models presented a moderate performance and showed a reasonable capacity to estimate the values observed in the reservoirs. Figure 4 presents the observed data and the estimates obtained during the validation test.

7.3. Web app user interface

After carrying out the model validation tests, it was possible to identify the MLPNN as the best performing model. Based on this information, neural network weights were obtained, and feedforward (predict) was implemented to calculate the estimated water temperature in reservoirs and rivers.

The developed system allows the user to browse the map and, by clicking on a given location (longitude and latitude), obtain real-time information on the estimated water temperature for the selected reservoir or river. This functionality provides the user with immediate access to the desired information, regardless of their location. In the prototype map interface visualized in Figure 5, the blue dots represent the rivers, while the green dots represent the reservoirs. This visual differentiation of the points allows a clear identification of the bodies of water under analysis, facilitating the interaction and understanding of the results by the users.

In this way, the developed prototype is an interactive and efficient web application that allows users to obtain information about the water temperature of specific water bodies quickly and accurately, using the best model (MLPNN) trained and the available data. This contributes to a better understanding and monitoring of water temperature in real time in different locations.

8. CONCLUSIONS

In this article, a study was carried out to estimate the water temperature in rivers and reservoirs in Pernambuco, using four machine learning models. The main objective was to evaluate and compare the performance of these models, in order to select the most accurate one and use its results in the creation of a web interface, in which users can obtain water temperature estimates for specific rivers and reservoirs.

Regarding the statistical indices used to evaluate the accuracy of the models, it was observed that they presented similar results for both rivers and reservoirs. However, the MLPNN model showed a slightly superior performance compared to the other models, obtaining a MSE index of 0.343 in the case of rivers. In the reservoirs and rivers, the Stochastic and Linear Regression models showed similar results in terms of RMSE, MAE and R^2. These results indicate that all trained models had a satisfactory ability to estimate the observed values, and the choice of the best model was based on a small difference in performance.

Furthermore, this study evidenced the importance of a high correlation between air temperature and water temperature as a fundamental requirement to obtain good accuracy in the developed models. However, a significant limitation of this work is related to the amount of data sampled for each river and reservoir. Due to the manual collection carried out on specific dates and times, some bodies of water had fewer than 20 samples available, which made individual training of models for each of them unfeasible. This scarcity of data may have affected the accuracy and generalizability of the developed models. In addition, another important limitation was the availability of only one predictor variable collected simultaneously, which was the air temperature. The absence of other meteorological variables, such as solar radiation, wind, and humidity, due to collection at random times restricted the ability of the models to capture more comprehensive and relevant information about environmental conditions and made it impossible to obtain data from meteorological stations, which usually provide hourly maximum and minimum temperature information. The absence of water flow data also impacted the results, since the amount of flowing water plays a crucial role in water temperature, influencing its heat transport capacity in the water body.

Therefore, it is recommended that future research seek to obtain more robust datasets, preferably for each river and reservoir, in order to enable individualized training for each dataset. This would allow achieving greater accuracy in the results of the algorithms. In addition, it would be beneficial to consider the inclusion of other predictive variables, such as solar radiation, wind speed and flow of water bodies, which also directly influence water temperature, as demonstrated in other studies.

It is hoped that this study will serve as a source of inspiration for the scientific community, stimulating the application of this method in other areas. In addition, the interface developed to estimate water temperature can be a viable alternative to monitor reservoirs and rivers in Pernambuco, contributing to proper management and preventing the proliferation of pathogenic organisms, which may be associated with inadequate water temperatures, aiming at protecting human health and the security of water resources.

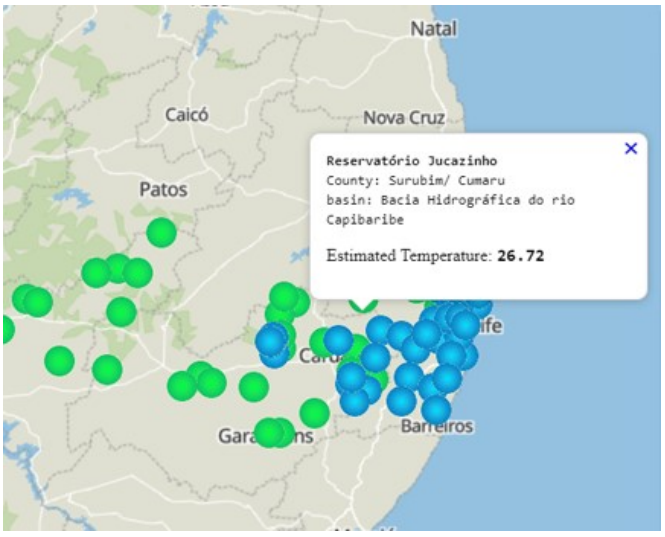

Figure 5. Interactive map user interface.

ACKNOWLEDGEMENT

This research is financially supported by Coordination for the Improvement of Higher Level Personnel (CAPES). we would like to extend our heartfelt appreciation to the Pernambuco Water and Climate Agency (APAC) for their invaluable assistance in providing us with data on water and air temperature of reservoirs and rivers.

REFERENCES

[1] F. J. Peñas, M. Álvarez-Cabria, M. Sáinz-Bariáin (+ another 17 authors), An evaluation of freshwater monitoring programs in ILTER nodes and mountain national parks: identifying key variables to monitor global change effects, Biodivers Conserv 32, 2023, pp. 65–94.
DOI: 10.1007/s10531-022-02466-x

[2] J. C. Semenza, Cascading risks of waterborne diseases from climate change, Nature immunology, 21(5) 2020, pp. 484-487.
DOI: 10.1038/s41590-020-0631-7

[3] K. S. Adu-Manu, C. Tapparello, W. Heinzelman, F. A. Katsriku, J. D. Abdulai, Water quality monitoring using wireless sensor networks: Current trends and future research directions, ACM Transactions on Sensor Networks (TOSN), 13(1) 2017, pp. 1-41.
DOI: 10.1145/3005719

[4] S. Zhu, E. K. Nyarko, M. Hadzima-Nyarko, S. Heddam, S. Wu, Assessing the performance of a suite of machine learning models for daily river water temperature prediction, PeerJ, 7, 2019, e7065.
DOI: 10.7717/peerj.7065

[5] S. Zhu, S. Heddam, S. Wu, J. Dai, B. Jia, Extreme learning machine-based prediction of daily water temperature for rivers, Environmental Earth Sciences, 78, 2019, pp. 1-17.
DOI: 10.1007/s12665-019-8202-7

[6] S. Zhu, S. Ptak, Z. M. Yaseen, J. Dai, B. Sivakumar, Forecasting surface water temperature in lakes: A comparison of approaches, Journal of Hydrology, 585, 2020, pp. 124809.
DOI: 10.1016/j.jhydrol.2020.124809

[7] G. T. Colombo, M. Mannich, Estimativa da temperatura da água em rios utilizando a média móvel da temperatura do ar, XXIII simpósio brasileiro de recursos hídrico, Foz do Iguaçu, Brazil, December, 2019. [In Portuguese]

[8] S. Ray, A quick review of machine learning algorithms, 2019 Int. conf. on machine learning, big data, cloud and parallel computing (COMITCon), Faridabad, India, 14-16 February 2019, pp. 35-39.
DOI: 10.1109/COMITCon.2019.8862451

[9] P. P. Shinde, S. Shah, A review of machine learning and deep learning applications, 2018 Fourth international conference on computing communication control and automation (ICCUBEA), IEEE, 2018, pp. 1-6.
DOI: 10.1109/ICCUBEA.2018.8697857

[10] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, Representation of compounds for machine-learning prediction of physical properties, Physical Review B, 95(14) 2017, pp. 144110.
DOI: 10.1103/PhysRevB.95.144110

[11] M. Flah, I. Nunez, W. Ben Chaabene, M. L. Nehdi, Machine learning algorithms in civil structural health monitoring: A systematic review, Archives of computational methods in engineering, 28, 2021, pp. 2621-2643.
DOI: 10.1007/s11831-020-09471-9

[12] M. Furdek, C. Natalino, F. Lipp, D. Hock, A. Di Giglio, M. Schiano, Machine learning for optical network security monitoring: A practical perspective, Journal of Lightwave Technology, 38(11) 2020, pp. 2860-2871.
DOI: 10.1109/JLT.2020.2987032

[13] M. Z. Zhang, L. M. Wang, S. M. Xiong, Using machine learning methods to provision virtual sensors in sensor-cloud, Sensors, 20(7) 2020, pp. 1836.
DOI: 10.3390/s20071836

[14] M. A. Zaidan, N. H. Motlagh, P. L. Fung (+ another 8 authors), Intelligent calibration and virtual sensing for integrated low-cost air quality sensors, IEEE Sensors Journal, 20(22) 2020, pp. 13638-13652.
DOI: 10.1109/JSEN.2020.3010316

[15] S. Yoon, Virtual sensing in intelligent buildings and digitalization, Automation in Construction, 143, 2022, 104578.
DOI: 10.1016/j.autcon.2022.104578

[16] M. Drakoulelis, G. Filios, V. G. Ninos, I. Katsidimas, S. Nikoletseas, Virtual sensors: an industrial application for illumination attributes based on machine learning techniques, Annals of Telecommunications, 76 (7-8) 2021, pp. 529-535.
DOI: 10.1007/s12243-021-00856-w

[17] V. Jaiswal, A. A. Brown, M. Yu, Virtual sensors for mooring line tension monitoring, Offshore Technology Conference OTC, Houston, Texas, USA, 2020, pp. D011S012R005.
DOI: 10.4043/30562-MS

[18] E. B. Ilyas, M. Fischer, T. Iggena, R. Tönjes, Virtual sensor creation to replace faulty sensors using automated machine learning techniques, 2020 Global Internet of Things Summit (GIoTS), IEEE, (2020), pp. 1-6.
DOI: 10.1109/GIOTS49054.2020.9119681

[19] M. Bertocco, S. Parrino, G. Peruzzi, A. Pozzebon, Estimating volumetric water content in soil for IoUT contexts by exploiting RSSI-based augmented sensors via machine learning, In: Sensors, 23(4), (2023), pp. 2033.
DOI: 10.3390/s23042033

[20] G. Scandurra, A. Arena, C. Ciofi, A Brief Review on Flexible Electronics for IoT: Solutions for Sustainability and New Perspectives for Designers, In: Sensors, 23(11), (2023), pp. 5264.
DOI: 10.3390/s23115264

[21] S. Zhu, E. K. Nyarko, M. Hadzima-Nyarko, Modelling daily water temperature from air temperature for the Missouri River, PeerJ, 6, (2018), pp. e4894.
DOI: 10.7717/peerj.4894

[22] D. C. Montgomery, E. A. Peck, G. G. Vining, Introduction to linear regression analysis, John Wiley & Sons, (2021), ISBN 9781119578741.

[23] N. Ketkar, Stochastic gradient descent. Deep learning with Python: A hands-on introduction, (2017), ISBN 978-1-4842-2766-4, pp. 113-132.
DOI: 10.1007/978-1-4842-2766-4_8

[24] M. W. Ahmad, J. Reynolds, Y. Rezgui, Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees, Journal of cleaner production, 203, (2018), pp. 810-821.
DOI: 10.1016/j.jclepro.2018.08.207

[25] M. Desai, M. Shah, An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN), Clinical eHealth, 4, (2021), pp. 1-11.
DOI: 10.1016/j.ceh.2020.11.002

[26] Maptiler, Maps for developers. Online [Accessed 31 July 2023] https://www.maptiler.com/

[27] Visualcrossing, weather data and API. Online [Accessed 31 July 2023] https://www.visualcrossing.com/