EDICIONES
COMPLUTENSE

Anales de Geografía de la Universidad Complutense
ISSN: 0211-9803

ARTÍCULOS

Using Xgboost models for daily rainfall prediction

Rafael Grecco Sanches
São Carlos School of Engineering, University of São Paulo (USP), São Paulo, Brazil ✉
Rodrigo Sanches Miani
School of Computer Science, Federal University of Uberlandia (UFU), Minas Gerais, Brazil
Bruno César dos Santos
Department of Environmental Sciences (DCAM), Federal University of São Carlos (UFSCar), São Carlo, Brazil
Rodrigo Martins Moreira
Environmental Engineering Department, Federal University of Rondônia (UNIR), Rondônia, Brazil
Gustavo Zen de Figueiredo Neves
São Carlos School of Engineering, University of São Paulo (USP), São Paulo, Brazil
Vandoir Bourscheidt
São Carlos School of Engineering, University of São Paulo (Brazil)
Pedro Augusto Toledo Rios
Faculty of Civil Engineering, Federal University of Uberlandia (UFU), Minas Gerais, Brazil

https://dx.doi.org/10.5209/aguc.98944

Received: November 7, 2024 / Submitted for review: November 11, 2025 / Accepted: June 11, 2025 / Published online: June X, 2025

Abstract. Machine learning models for predicting daily precipitation have gained traction in recent years. Understanding the benefits of using this technology in different regions is a relevant research topic. For this reason, this study aims to evaluate daily precipitation estimated forecasts from climate data between 1983 and 2019 in Itirapina, São Paulo, Brazil. We used a novel machine learning algorithm, XGBoost (eXtreme Gradient Boosting), to create several daily precipitation prediction models. Two tasks were modeled: the occurrence of daily precipitation (classification) and the amount of daily precipitation (regression). The results revealed that the occurrence of daily precipitation could be predicted with an accuracy of around 90%. Additionally, models were developed to predict the amount of daily precipitation with error rates of around 3mm. We observed that precipitation in the study area is directly associated with solar radiation, and estimated forecasts of precipitation and the corresponding months are characteristic of the tropical climate.
Keywords: precipitation; tropical climatology; machine learning; forecasting; XGBoost.

[ENG] Usando modelos XGBoost para la predicción diaria de lluvias

Resumen. Los modelos de aprendizaje automático para predecir las precipitaciones diarias han ganado fuerza en los últimos años. Comprender los beneficios del uso de esta tecnología en diferentes regiones es un tema de investigación relevante. Por esta razón, este estudio tiene como objetivo evaluar los pronósticos de lluvia diaria a partir de datos climáticos entre 1983 y 2019 en Itirapina, São Paulo, Brasil. Utilizamos un novedoso algoritmo de aprendizaje automático, XGBoost (eXtreme Gradient Boosting), para crear varios modelos de predicción de lluvia diaria. Se modelaron dos tareas: la aparición de precipitación diaria (clasificación) y la cantidad de precipitación diaria (regresión). Los resultados revelaron que la aparición de precipitaciones diarias se podía predecir con una precisión de alrededor del 90%. Además, se desarrollaron modelos para predecir la cantidad de lluvia diaria con tasas de error de alrededor de 3 mm. Observamos que la precipitación en el área de estudio está directamente asociada con la radiación solar, y los pronósticos de precipitaciones y los meses correspondientes son característicos del clima tropical.
Palabras clave: precipitaciones; climatología tropical; aprendizaje automático; previsión; XGBoost.

Cómo citar: Gedda-Shaheen, G., Caramés, E., Díaz-Pacheco, J., Hewitt, R. J. (2025). Large detail scale mapping from aerial orthophotos for understanding local-scale urban land use dynamics. The case of Madrid, Spain. *Anales de Geografía de la Universidad Complutense*, 45(1), 75-92.

1. Introduction

Estimated precipitation forecasting can be associated with various methods, including artificial intelligence techniques, such as Machine Learning (ML) and climate data automation (Facco et al., 2020; He et al., 2022; Pham et al., 2020; Kochhar et al., 2022). Considering specific characteristics in climate data, the application of fuzzy logic models has shown positive results in predicting short-term precipitation events (Liu et al., 2019). These models help analyze extreme events, climate patterns, and precipitation values, even without exploring recent prolonged droughts or heavy precipitation across continents (Facco et al., 2020; Pham et al., 2020; Ramirez & Lizarazo, 2017).

The eXtreme Gradient Boosting (XGBoost) is a state-of-the-art open-source package with high efficiency and scalable implementation of the gradient boosting framework (Chen & Guestrin, 2016; Chen et al., 2023; Das et al., 2024). XGBoost has been applied in various environmental studies related to precipitation, such as flood risk assessment (Mu et al., 2021), landslide displacement (Xu et al., 2022), groundwater levels (Ibrahem Ahmed Osman et al., 2021), reservoir inflow (Muhammad et al., 2020), water levels (Nguyen et al., 2021), and urban stormwater management (Zhou & Lau 2001; Zhou et al., 2022, 2021).

Compared to other ML methods, the XGBoost algorithm processes climate data more efficiently, especially for continental climates (Chen & Guestrin, 2016; Das et al., 2024). Additionally, XGBoost has demonstrated superior performance in predicting temperature regimes within forests (Ghafarian et al., 2022) assessing drought risks using ML methods (Prodhan et al., 2022), predicting landslides in high-risk areas (Stanley et al., 2020), estimating forecast river flow using hybrid ML models (Tan et al., 2022), and accurately predicting daily precipitation and precipitation patterns using AI-based estimated forecasts (Pham et al., 2020), among other applications.

Similar approaches have been employed in China, where machine learning techniques, specifically XGBoost, have been utilized to predict daily land surface temperatures and make seasonal predictions of precipitation across the Eurasian continent (Liu et al., 2022). Furthermore, in areas of Eurasia with non-monsoon winter precipitation, the dynamic mechanisms of precipitation variation have been studied using four machine learning models. These models provide average predictions for the winter precipitation ensemble, highlighting the highest temporal correlations. These findings contribute to seasonal estimated forecasts for winter precipitation across the Eurasian continent (Qian et al., 2021, 2020).

Similarly, various machine learning techniques were employed in Iran to simulate precipitation amounts (Nakhaei et al., 2023). The models showed satisfactory accuracy, particularly for the wet season compared to the dry season (Heydarizad et al., 2022). XGBoost proved to be robust and effective in predicting the occurrence of daily precipitation in Vietnam using daily meteorological databases (Pham et al., 2020).

This method effectively incorporates precipitation intensity, soil moisture, and snowmelt as indicators linked to landslide risks associated with precipitation in the region (Stanley et al., 2020). The accuracy of the machine learning model, combined with drought indices, significantly improved for estimated forecasting cereals in Morocco, achieving 93% accuracy in forecasting income variation on a national scale (Bouras et al., 2021).

Whether obtained from conventional weather stations, satellite precipitation estimates, or precipitation reanalysis, global precipitation data are crucial for machine learning algorithms. The ability to generate numerical and graphical results makes artificial intelligence approaches preferable, considering various time scales such as daily, monthly, semi-annual, and annual analyses (Oliveira et al., 2014; Parmar et al., 2017; Pham et al., 2020; Rasouli et al., 2012; Sachindra et al., 2018).

In Brazil, the methodology of estimated climate forecasting and the utilization of ML tools for transferring knowledge from historical data are current trends in atmospheric research. It is indeed feasible to enhance learning by incorporating ML with simulated data (Liu et al., 2022; Pham et al., 2020; Rolnick et al., 2022). The use of XGBoost algorithms to predict surface sediment concentrations in the Doce River basin was proven efficient, according to the results reported by Aires et al. (2023). The estimated forecast obtained satisfactory results in predicting the dry period based on historical data from a climatological station in the study area (Bouras et al., 2021; Monego et al., 2022).

The rest of this paper is organized as follows: Section 2 describes the materials and methods associated with our study. Section 3 details each step of our experimental evaluation. Section 4 presents the results of our prediction models. Section 5 discusses the impact of our results. Finally, Section 6 presents some concluding remarks and future research possibilities.

2. Materials and methods

2.1. XGBoost

Gradient boosting is a robust ML algorithm widely applied to climate studies to understand nonlinear relationships between dependent and independent variables (Fan et al., 2018; Ardabili et al., 2019; Tian et al., 2022). Usual ensemble techniques, such as random forests, are based on simple averaging of the models in the ensemble. A family of enhancement methods is based on different constructive integration strategies. The central concept behind Gradient Boosting is to add new models to the ensemble sequentially. One of the main challenges in using gradient boosting is adjusting parameter values.

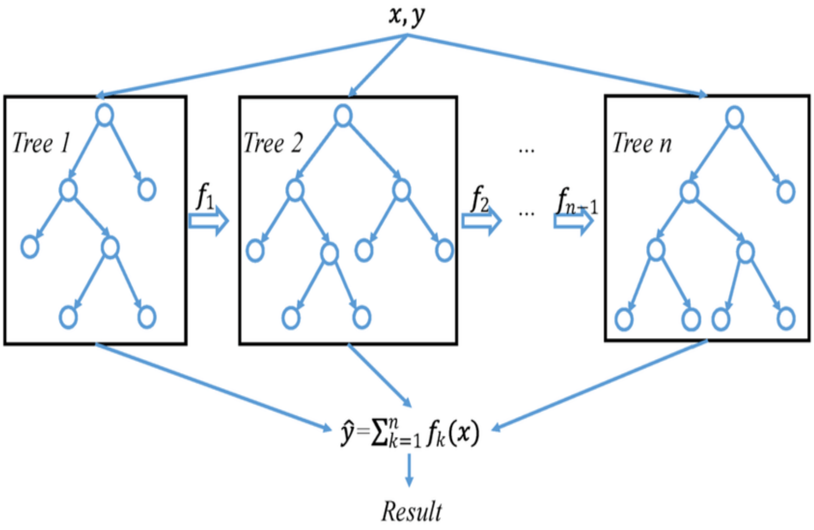
We used XGBoost (eXtreme Gradient Boosting), an efficient and scalable implementation of the gradient boosting framework (Chen et al., 2015). XGBoost uses a tree ensemble model consisting of a set of Classification and Regression Trees (CART) (Breiman, 2001). Since a single tree may not be sufficient to achieve good results, XGBoost uses multiple CARTs together, and the final prediction is the sum of each CART's score. We adopted the XGBoost algorithm due to its success in several machine learning competitions on a wide range of problems, such as store sales prediction, customer behavior prediction, and web text classification (Nielsen, 2019). Another advantage of XGBoost is its performance. According to Shilong (2019), the algorithm runs more than ten times faster than existing popular solutions.

ML concepts and algorithms on which XGBoost is based include supervised ML, decision trees, ensemble learning, and gradient boosting, among others (Chen & Guestrin, 2016; Wang et al., 2019; Tian et al., 2022). Likewise, the XGBoost tool is a scalable and distributed decision tree machine learning library (Bentéjac et al., 2021). It also uses algorithms to train a model that identifies patterns in labeled data and applies this knowledge to predict outcomes from new datasets (Figure 1).

XGBoost is an ensemble of decision trees, similar to random forests (an algorithm that combines several decision trees to produce a single result), and is used for classification and regression. It uses gradient values and sums them to evaluate the quality of splits in the training dataset (Figure 2). However, XGBoost focuses on reducing the complexity of finding the best split, which is the most time-consuming part of decision tree construction. Despite this, decision trees are easy to visualize and considered reasonably interpretable algorithms (Chen and Guestrin, 2016; Bentéjac et al., 2021; Tarwidi et al., 2023).considered

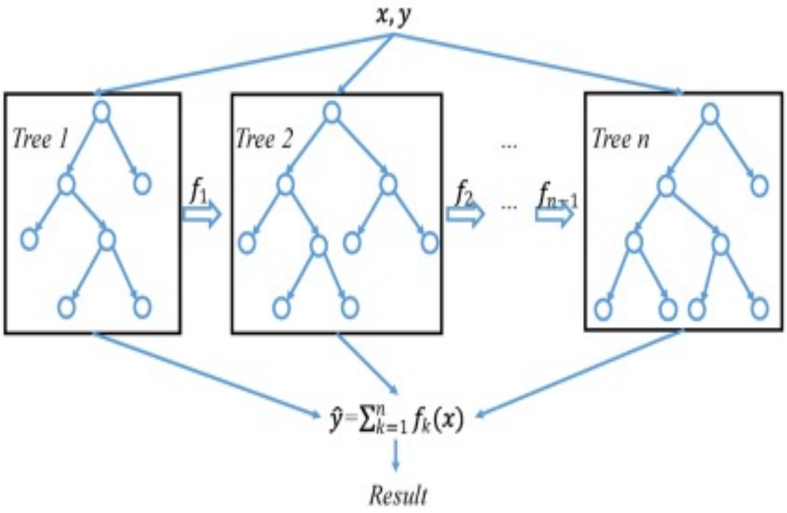
Feature importance analysis quantifies the degree to which each feature contributes to the raw daily weather data during the construction process, serving as a crucial metric for assessing feature significance. However, feature importance scores, which are based on the contribution of features to rainfall data, cannot elucidate individual weather feature predictions. Feature modeling, on the other hand, provides a more effective approach to estimating supervised regression feature importance scores, facilitating the interpretation of individual features contributions, as we conducted in the study area (Anwar et al., 2021; Althoff et al., 2022; Ma et al., 2024).

Figure 1. Organizational chart of the XGBoost model.



Source: Demajo & Lara (2020)

Figure 2. Architecture of the XGBoost algorithm with the dataset.



Source: Wang et al. (2019).

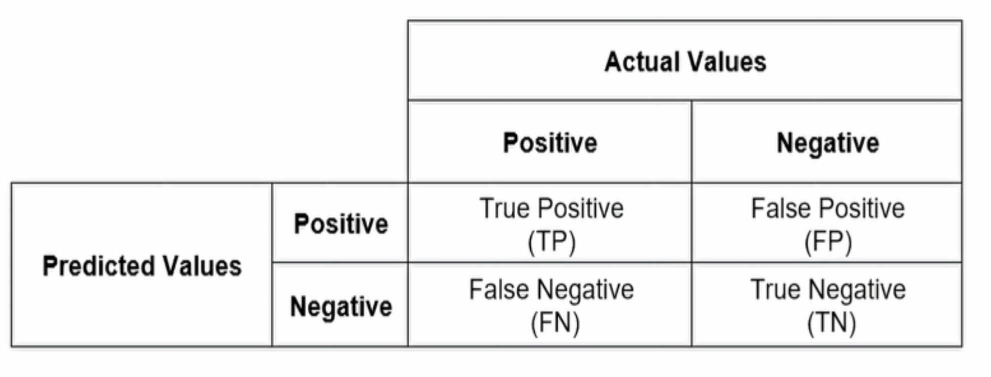
2.2. Validation Criteria

As discussed in the introduction, we created two types of models: 1) for predicting the occurrence of daily precipitation (a binary classification task) and 2) for predicting the amount of daily precipitation (a regression task).

In general, the confusion matrix is used to calculate performance metrics for binary classification tasks. Each row of the matrix represents the predicted class, while each column indicates the true class of the instance (Figure 3). The interpretation of each entry in the matrix is defined as follows:

- True Positive (TP): the model correctly classifies a rainy day.
- True Negative (TN): the model accurately classifies non-rain days.
- False Positive (FP): the model mistakenly classifies an instance as a rainy day, even though it is a non-rain day.
- False Negative (FN): the model misclassifies an instance as a non-rain day when, in reality, it is a rainy day.

Figure 3. Confusion matrix. Source: own elaboration.



		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Thus, when the model classifies an instance, one of the cells in the confusion matrix is updated. In binary and imbalanced classification problems, the minority class is considered the positive class. Therefore, in the daily precipitation forecasting problem, instances labeled as “Rain” are interpreted as positive, while instances labeled as “No rain” are interpreted as negative. We used the following metrics extracted from the confusion matrix to evaluate our binary classification models: Precision, Recall, and Accuracy.

- Precision indicates the ratio of days correctly classified as rainy among all those classified as rainy, including false positives:

$$Precision = \frac{TP}{TP + FP}$$

- Recall indicates the sensitivity of the model in classifying rainy days. It indicates the proportion of correctly classified rainy days among all days that actually experienced rain:

$$Recall = \frac{TP}{TP + FN}$$

- Accuracy represents the proportion of correctly classified days (both rainy and non-rainy) relative to the total number of instances:

$$Accuracy = \frac{TP + TN}{TN + FP + TP + FN}$$

To evaluate the models that predict the amount of daily precipitation, we use the following metrics, which measure the distance between real data and the predicted data:

- R-squared (R^2), also known as the coefficient of determination. It is a statistical measure used to assess how well a linear regression model fits the data points:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - y_p\right)^2}{\sum_{i=1}^{n} \left(y_i - \underline{y}_p\right)^2}$$

where i denotes the i-th day; yi denotes the actual value for the i-th day; yp denotes the average of predicted value for the i-th day; and yp denotes the predicted value for the i-th day.

- Mean Absolute Error (MAE), which represents the average of the absolute difference between the actual and predicted values in the dataset:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_p|}{n}$$

where i denotes the i-th day, y_i denotes the actual value for the i-th day and y_p denotes the predicted value for the i-th day.

- Mean Absolute Percentage Error (MAPE), which is a measure of the accuracy of a forecasting or prediction model. It calculates the average percentage difference between the predicted values and the actual values:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - y_p|}{max(\epsilon, |y_i|)}$$

where yi denotes the actual value for the i-th day, yp denotes the predicted value for the i-th day and is an arbitrary small yet strictly positive number to avoid undefined results when y is zero.

- Mean Squared Error (MSE), which represents the average of the squared differences between the predicted and actual values of a variable:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - y_p)\ 2}{n}$$

where i denotes the i-th day, y_i denotes the actual value for the i-th day and y_p denotes the predicted value for the i-th day.

- Root Mean Squared Error (RMSE), which represents the average of the squared difference between the original and predicted values in the data set. The idea here is to evaluate the variance of the residuals:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_p)\ 2}{n}}$$

where yi denotes the actual value for the i-th day and yip denotes the predicted value for the i-th day. A lower value of MAE and RMSE, implies a higher accuracy of the regression model.

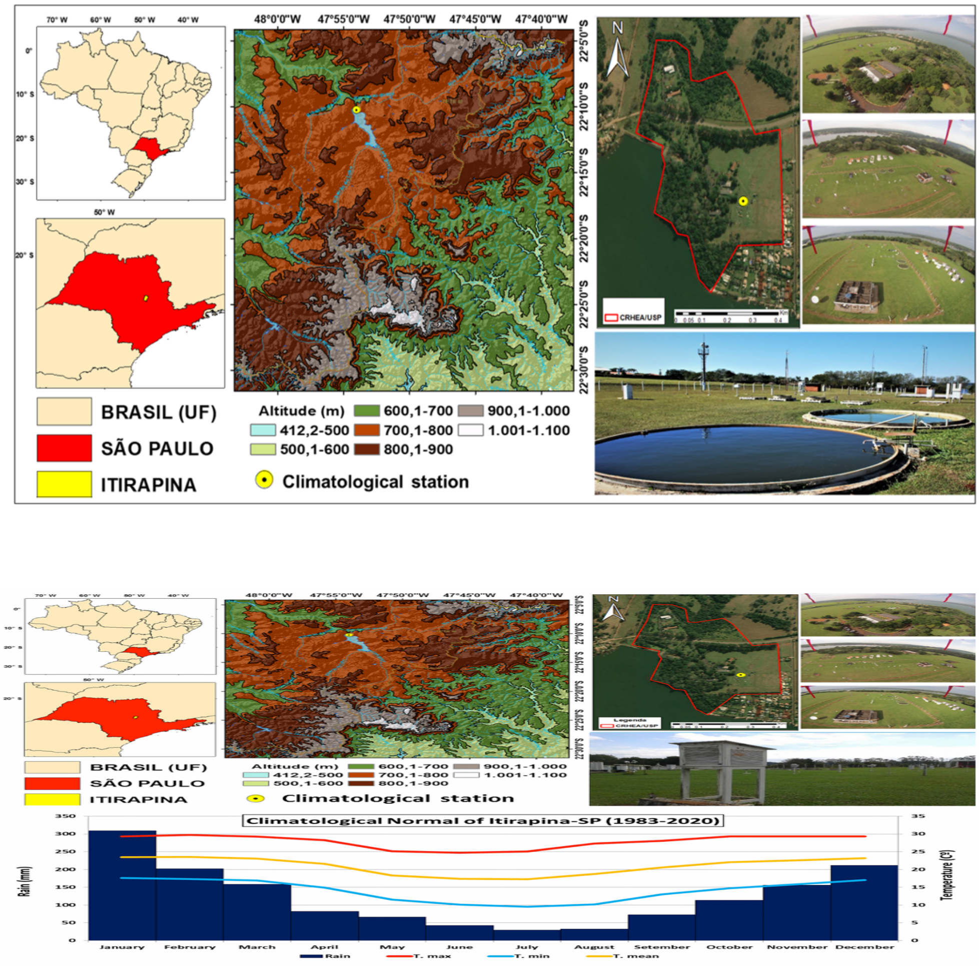
2.3. Study area

The city of Itirapina, located in the interior of the State of São Paulo, experiences an alternating rainy and dry seasons. Summers experience higher temperatures than winters, with highest precipitation totals occurring during summer. High temperatures alternate with milder temperatures in June, July, and August. Precipitation is concentrated in December, January, and February (Sanches, 2015; Sanches et al., 2020, 2018; Santos et al., 2017; dos Santos et al., 2022).

In the study area, air masses are influenced by the South Atlantic Convergence Zone (SACZ) and tropical and polar masses. These elements impact the Center-South region of Brazil, which is situated in a plateau region and on the edge of the Serra Geral sandstone cuestas. This situation reinforces orographic precipitation within the region's geographic dynamics. The rainiest quarter occurs between December and February, with annual precipitation ranging from 1500 to 1800 mm (Tolentino, 2007).

The climatic classification of the Itirapina region in São Paulo has been established using different classification systems. According to Tolentino (2007, p. 63-70), it is characterized using the classifications of Köppen (Cwa.i – Awi), Thornthwaite (BB'w), and Serebrenick (TUV°). Daily climatological data from 1983 to July 2020 were chosen from the climatological station located at CRHEA (Center for Water Resources and Environmental Studies) / EESC (São Carlos Engineering School) / USP (University of São Paulo) in Itirapina, São Paulo, as shown in Figure 4.

Figure 4. Location of the climatological station in Itirapina/SP.



Source: own elaboration.

2.4. Datasets

This study deliberately avoided using data-filling techniques, choosing not to fill in missing data. Notably, the proportion of missing data was below 5%, ensuring that at least 95% of the precipitation dataset was complete and reliable.

Table 1 showcases data collected between 1983 and 2019, without any processing. This dataset corresponds to the "Generic" model (GM).

Table 1. Summary of the complete dataset (01/01/1983 to 12/31/2019) for the GM model. Total Rainy Days: 3644. Total Dry Days: 7227.

	Max Temperature (°C)	Min Temperature (°C)	Average temperature (°C)	Wind Speed (km/h)	Solar Radiation (cal/cm²/h)	Pressure (mb)	Relative humidity (%)	Daily rain (mm)
mean	27.99	14.26	21.16	3.17	371.14	932.81	72.21	4.11
std	3.69	4.09	3.44	2.67	125.34	5.37	15.50	10.56
min	11.40	0.20	6.80	0.00	2.24	903.30	5.00	0.00
25%	26.00	11.40	19.00	1.52	286.10	928.50	63.80	0.00
50%	28.40	15.00	21.60	2.48	365.41	932.90	73.30	0.00
75%	30.60	17.40	23.80	4.03	459.22	937.10	82.00	1.70
max	38.80	28.60	34.20	99.20	1019.57	966.60	905.00	158.40

Table 2 presents data from 1983 to 2019, excluding extreme precipitation events and omitting days identified as outliers (Sanches et al., 2018, 2020). This refined dataset corresponds to the "Generic adjusted" model (GMA). Table 2. Summary of the adjusted dataset (01/01/1983 to 12/31/2019) for the GMA model.

Table 2. Summary of the adjusted dataset (01/01/1983 to 12/31/2019) for the GMA model. Total Rainy Days: 3495. Total Dry Days: 7227.

	Max Temperature (°C)	Min Temperature (°C)	Average temperature (°C)	Wind Speed (km/h)	Solar Radiation (cal/cm^2/h)	Pressure (mb)	Relative humidity (%)	Daily rain (mm)
mean	28.01	14.22	21.15	3.17	371.90	932.85	72.08	3.31
std	3.69	4.09	3.45	2.68	124.95	5.36	15.51	7.83
min	11.40	0.20	6.80	0.00	2.24	903.30	5.00	0.00
25%	26.00	11.40	19.00	1.52	287.22	928.50	63.80	0.00
50%	28.40	15.00	21.60	2.48	365.60	933.10	73.20	0.00
75%	30.60	17.40	23.80	4.03	459.22	937.20	81.80	1.40
max	38.80	28.60	34.20	99.20	1019.57	966.60	905.00	46.30

Table 3 provides an overview of the historical data for the period from October to March of the following year, corresponding to the "Rain season" model (RSM). Lastly, Table 4 presents data from the historical series between April and September of the same year, corresponding to the "Dry season" model (DSM).

In our models, a total of seven parameters (maximum temperature, minimum temperature, average temperature, wind speed, solar radiation, pressure, and relative humidity) were used as input variables, while daily precipitation data served as the output variable for generating training and testing datasets. A total of 13,729 samples were collected from January 1, 1983, to December 31, 2019.

Table 3. Summary of the adjusted dataset - Rain season (01/01/1983 to 12/31/2019) for the RSM model. Total Rainy Days: 2574. Total Dry Days: 2822.

	Max Temperature (°C)	Min Temperature (°C)	Average temperature (°C)	Wind Speed (km/h)	Solar Radiation (cal/cm^2/h)	Pressure (mb)	Relative humidity (%)	Daily rain (mm)
mean	29.30	16.32	22.86	3.37	421.86	931.26	72.92	5.10
std	3.12	2.78	2.49	2.63	130.00	4.88	17.59	9.34
min	12.20	1.60	10.30	0.00	6.60	903.30	6.00	0.00
25%	27.50	14.90	21.50	1.69	336.42	927.40	64.30	0.00
50%	29.60	16.70	23.20	2.67	433.40	930.90	73.80	0.00
75%	31.50	18.30	24.50	4.25	515.65	935.50	83.30	5.80
max	37.40	28.60	34.20	33.91	1019.57	952.40	905.00	46.30

Table 4. Summary of the adjusted dataset - Dry season (01/01/1983 to 12/31/2019) for the DSM model. Total Rainy Days: 921. Total Dry Days: 4405.

	Max Temperature (°C)	Min Temperature (°C)	Average temperature (°C)	Wind Speed (km/h)	Solar Radiation (cal/cm^2/h)	Pressure (mb)	Relative humidity (%)	Daily rain (mm)
mean	26.71	12.10	19.42	2.97	321.29	934.47	71.23	1.49
std	3.76	4.11	3.42	2.70	96.01	5.35	13.02	5.34
min	11.40	0.20	6.80	0.01	2.24	908.90	5.00	0.00
25%	24.60	9.40	17.30	1.38	266.50	930.20	63.20	0.00
50%	27.10	12.00	19.50	2.30	325.33	934.80	72.50	0.00
75%	29.20	15.00	21.80	3.77	382.59	938.80	80.30	0.00
max	38.80	22.80	29.50	99.20	816.74	966.60	245.30	45.50

According to the descriptive statistics, although the complete dataset recorded a maximum daily precipitation of 158.40 mm, this value was identified as an outlier and adjusted during the data preprocessing phase. Consequently, the adjusted dataset shows daily precipitation values ranging from 0.00 mm (no rain) to 46.30 mm, with a mean of 3.31 mm and a standard deviation of 7.83 mm.

Maximum temperatures range from 11.40 °C to 38.80 °C, with a mean of 28.01 °C and a standard deviation of 3.69 °C. Minimum temperatures vary from 0.20 °C to 28.60 °C, with a mean of 14.22 °C and a standard deviation of 4.09 °C. Average temperatures range from 6.80 °C to 34.20 °C, with a mean of 21.15 °C and a standard deviation of 3.45 °C. Wind speed ranges from 0.00 km/h to 99.20 km/h, with a mean of 3.17 km/h and a standard deviation of 2.68 km/h. Solar radiation values range from 2.24 cal/cm^2/h to 1019.57 cal/cm^2/h, with a mean of 371.90 cal/cm^2/h and a standard deviation of 124.95 cal/cm^2/h. Atmospheric pressure ranges from 903.30 mb to 966.60 mb, with a mean of 932.85 mb and a standard deviation of 5.36 mb.

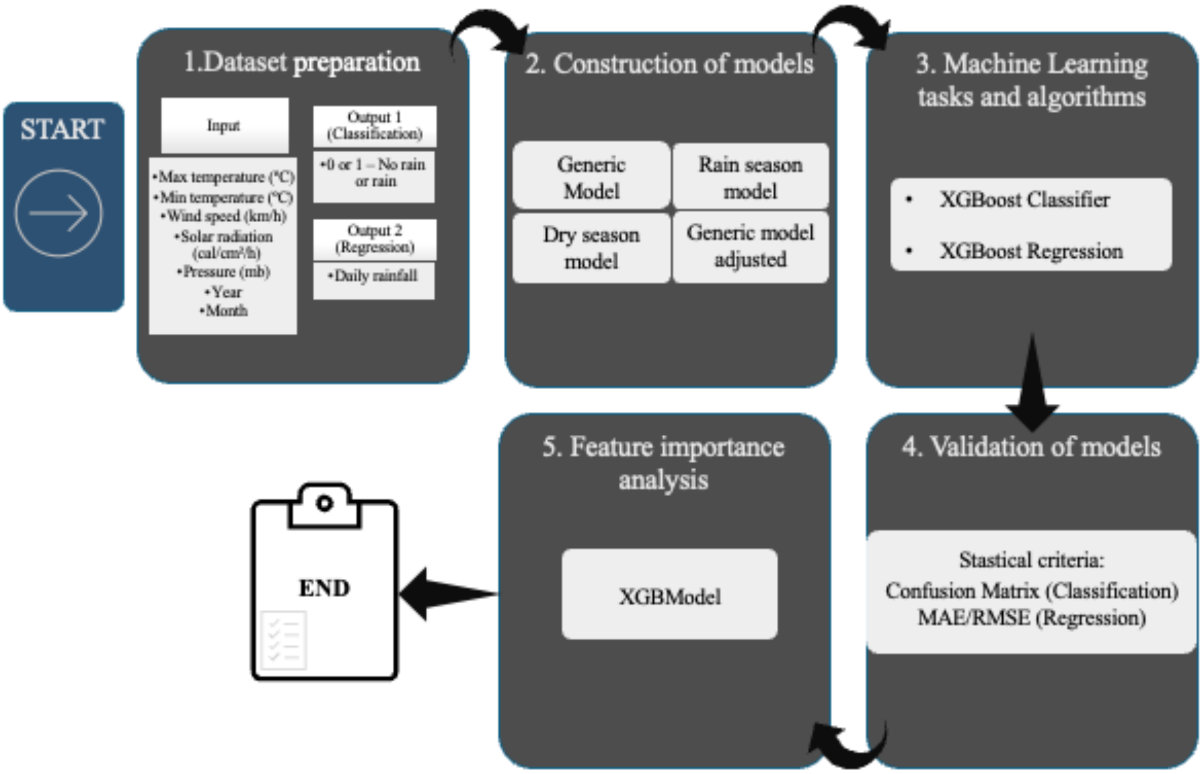
Finally, relative humidity ranges from 5.00% to 90.50%, with a mean of 72.08% and a standard deviation of 15.51%. During data inspection, an erroneous recorded value of 905.00% was detected and corrected.

These values characterize the weather patterns over a 37-year period, based on daily records of the climatological variables used in this study. This information supports the model's capability to predict next-day precipitation under both wet and dry conditions in the study area.

3. Proposal of a daily precipitation model

Figure 5 summarizes the daily precipitation modeling process, which includes the following steps: 1) dataset preparation; 2) construction of the models; 3) machine learning model training; 4) model validation; and 5) feature importance analysis. Next, we provide a detailed description of each step.

Figure 5. Daily precipitation prediction modeling.



Source: own elaboration.

During dataset preparation (step 1), the collected data were split into two parts: training and testing sets. The training set consists of data from 1980 to 2008, and the testing set consists of data from 2009 to 2019. The objective is to assess how effectively past data can be used to predict future outcomes.

It's important to note that we have four training and testing sets, each related to the four models: Generic (GM), Generic adjusted (GMA), Rain season (RSM), and Dry season (DSM). In this study, we conducted an experiment aimed at thoroughly verifying the model using daily data. The process involved training the model with data from 2009 to 2019 and testing it with data from 1980 to 2008. Due to limitations related to document length, editorial guidelines, and publication recommendations, only the generic model was evaluated.

For the model's construction (step 2), we used the training set to construct four precipitation prediction models (GM, GMA, RSM, and DSM) for each ML task. The aim here is to understand the impact of using different types of meteorological data on the models' performance. In this study, it is observed that for each model, "slices" of daily data from the same dataset were utilized. However, these "slices" significantly affect the models, as each contains periods of either excessive or low (drought) daily precipitation. For instance, in the "rainy season" scenario, notable differences are observed in the analyzed attributes. Moreover, the average rainfall value in the GM scenario, similar to that in the rainy season scenario, contrasts with the overall daily data.

For the ML training (step 3), as discussed in Section 2, we chose the XGBoost algorithm. We created two types of models: 1) predicting the occurrence of daily precipitation (binary classification task) and 2) predicting the amount of daily precipitation (regression task). For the binary classification task, we used the XGBoost implementation available in the scikit-learn package for Python, with its default parameters. For the regression task, we employed the XGBoost Regressor from scikit-learn, with the following parameters: colsample_bytree = 0.3, which defines the fraction of features randomly selected for each tree, reducing overfitting while potentially limiting the model's ability to capture complex patterns; learning_rate = 0.1, which controls the step size at each boosting iteration, where lower values improve generalization but require a greater number of iterations to reach convergence; max_depth = 100, which specifies the maximum depth of individual trees, allowing the model to capture intricate relationships but increasing the risk of overfitting; alpha = 1, which represents the L1 regularization term, promoting feature sparsity by driving some weights to zero, thereby reducing model complexity and enhancing generalization; and n_estimators = 100,000, which determines the number of boosting iterations, enabling the model to learn complex patterns at the expense of significantly higher computational cost. These parameter choices aim to balance model complexity and generalization, optimizing predictive performance for daily precipitation estimation.

Feature importance analysis (step 5) refers to procedures that assign a score to input features based on their usefulness in predicting the target variable. We utilized the built-in feature importance method from the XGBoost algorithm in scikit-learn (feature_importances). Since XGBoost is a tree-based algorithm, we used a technique called "weight," which is important based on the number of times a feature is used to split the data across all trees.

4. Results

4.1. Performance of classification models in predicting the occurrence of precipitation

First, the accuracy values for the investigated station were equal to or higher than 90% for all the proposed models. This is a strong indicator of the good performance of XGBoost in predicting the occurrence (or not) of precipitation. However, since the dataset is not well balanced (there are more days with no rain than the opposite), it is important to investigate the other metrics presented in Table 5.

Table 5. Evaluation of classification models in predicting the occurrence of precipitation across four datasets.

Model	Precision	Recall	Accuracy
Generic model (GM)	0,903	0,842	0,918
Generic model adjusted (GMA)	0,892	0,826	0,909
Rain season model (RSM)	0,877	0,891	0,890
Dry season model (DSM)	0,982	0,739	0,951

As defined in Section 2, high precision values indicate that a model could classify rainy days correctly. All four models exhibited good results, but the DSM model showed a precision value of 98%. We see here that applying a ML model to a specific dataset (months associated with the dry season) leads to an increase of almost 10% in terms of precision compared to the GM.

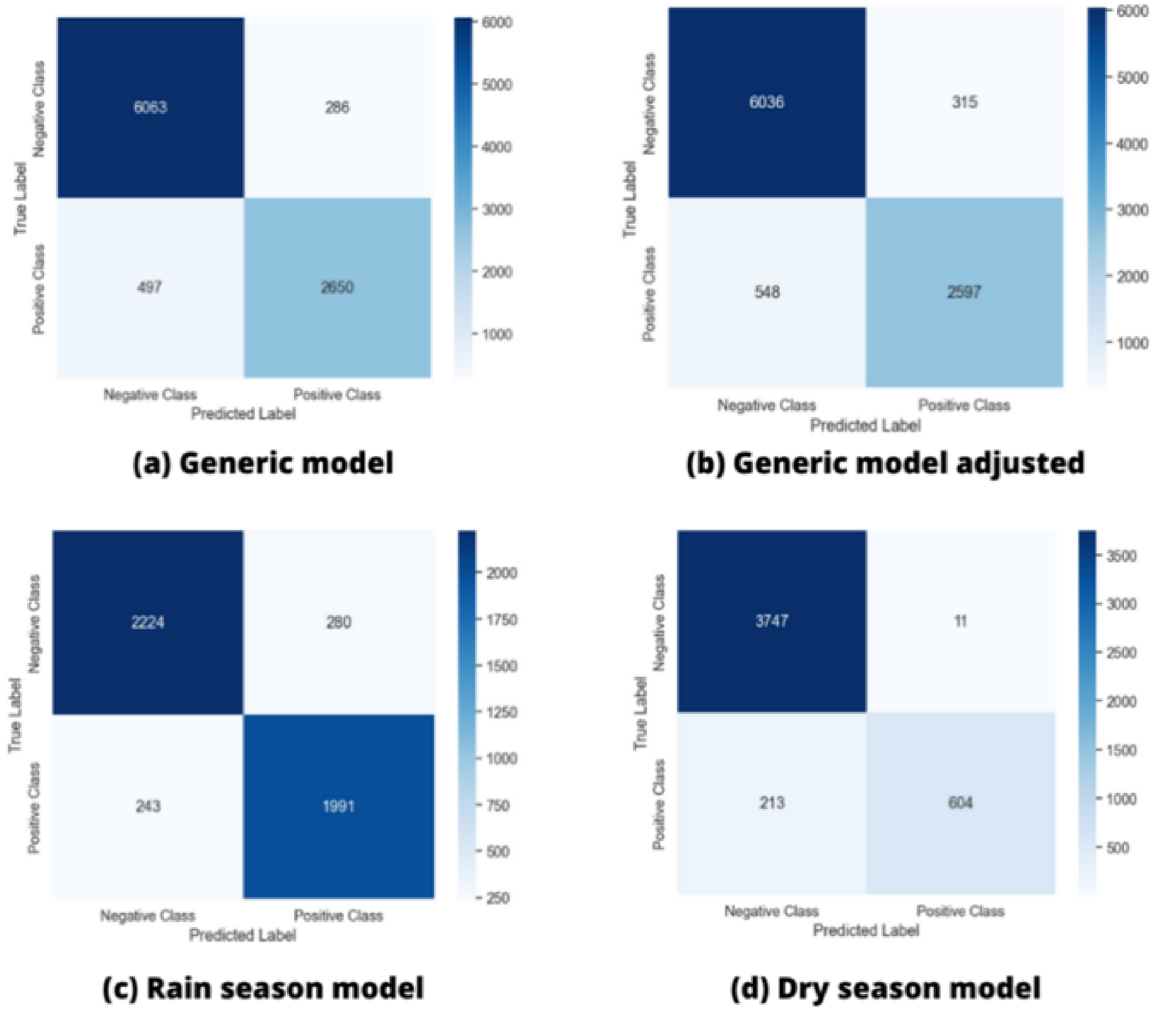
Another way to analyze the results is to look at the recall values. In this case, for the GM, the XGBoost algorithm was able to classify 84% of all rainy days. The maximum recall value achieved in our experiments was found in the RSM. In other words, these results indicate that XGBoost performed well in predicting the occurrence of daily precipitation.

It is interesting to see that removing extreme events (GMA) did not affect the performance of predicting the occurrence of precipitation compared to the GM, as observed in the RSM. However, differences were observed when analyzing the RSM and DSM models. We will discuss these differences when investigating the confusion matrix values.

Figure 6 depicts the confusion matrix for each classification model: generic (GM), generic adjusted (GMA), rain season (RSM), and dry season (DSM). Here, we can study the performance of each model regarding the two outcomes of our models: rain or no rain.

For example, for the GM, we have 6349 days with no rain (6,063 + 286). The model correctly predicted 6063 of them, which corresponds to 95%. On the other hand, we have 3147 days with rain (2,650 + 497), and the model correctly identified 2650 instances, corresponding to 84%. In other words, the generic model (GM) had an error rate of 5% for predicting days with no rain and 16% for predicting days with rain.

Figure 6. Confusion matrices for each classification model.



(a) Generic model

(b) Generic model adjusted

(c) Rain season model

(d) Dry season model

Source: own elaboration.

4.2. Performance of regression models in predicting the amount of daily precipitation

Table 6 summarizes the statistical methods (MAE, MAPE, RMSE, MSE, and R^2) discussed in Section 2.1. The validation process encompassed four subdatasets: the Generic model, Adjusted Generic model, Rainy Season model, and Dry Season model. Among them, the Adjusted Generic and Rainy Season models demonstrated reasonable performance, with R^2 values of 0.6, in contrast to the Generic model (0.03) and the Dry Season model (-0.1).

Upon closer examination, the MAE ranged from 3 (in the Adjusted Generic and Rainy Season models) to 25 (in the Dry Season model). A similar pattern emerged when considering the RMSE values. Smaller errors were evident in the Adjusted Generic and Rainy Season models, whereas the Generic model, particularly during the Dry Season, displayed larger errors in millimeters (approximately 50 mm).

Table 6. Evaluation of regression models in predicting the amount of daily precipitation across four datasets.

Model	MAE	MAPE	RMSE	MSE	R^2
Generic model (GM)	11.011	2.096	33.511	112.007	0.034
Generic model adjusted (GMA)	3.202	2.219	7.875	62.0116	0.693
Rain season model (RSM)	3.685	2.322	9.252	85.596	0.655
Dry season model (DSM)	25.786	6.683	50.427	252.882	-0.148

Overall, among the four tests, the GMA and RSM models demonstrated satisfactory performance and exhibited lower average error values that closely approximate the actual values utilized in this study. Conversely, the GM and DSM models exhibited low performance and higher average errors, with the GM model particularly affected. These differences can be attributed to two primary factors.

In the adjusted model, the monthly errors were the lowest, with error values close to the ideal (<4mm), and the dry season and rainy season models showed differences in the millimeter error values. However, during the rainy season (October to March), they were the best for daily precipitation estimated forecasts (<3.2mm) compared to the models where the errors remained relatively high (23.5mm) during the dry season (April to September).

Firstly, the presence of outliers within the historical series used for analysis led to differences between the GM and GMA models. The second model applied, the GMA model, was developed by excluding precipitation outliers from the historical series. Secondly, the alternating seasonality inherent in the tropical climate precipitation, characterized by rainy and dry periods, also influenced performance and errors in the Rainy and Dry Season models.

The RSM showcased superior estimated forecasting capabilities between these two models, owing to the atmospheric systems at play from October through March. Notably, the action of the South Atlantic Convergence Zone (SACZ) over the study area during the summer period results in a sequence of humid days and heavy precipitation. This propensity for wetter days fosters enhanced RSM model performance. Conversely, during the tropical climate period from April to September, the presence of humid days and substantial precipitation volumes diminishes due to the influence of the South Atlantic Subtropical High (ASAS) over the study area, particularly during winter. The transition months from the wet to the dry season of the tropical climate see the advent of precipitation events due to the advancement of frontal systems. These frontal advances can lead to atmospheric instabilities precipitating varying volumes of precipitation, which can affect the performance reduction seen in the DSM model.

Table 7 provides an overview of the four models' performance through a monthly analysis of precipitation errors. Considering these findings, the adjusted and wet period models demonstrated superior performance. This stems from the removal of outlier values for the adjusted model. The rainy period is associated with the volumes of precipitation that accumulate during the corresponding months of precipitation in the southeastern region of South America.

Table 7. Mean absolute error per month for each model.

Month	Generic model	Generic model adjusted	Dry season model	Rain Season Model
January	10,93	3,01	x	3,59
	11,01	3,07	x	3,69
	11,05	3,11	x	3,73
	11,09	3,15	21,22	x
	11,11	3,16	21,34	x
	11,14	3,17	23,42	x
	11,16	3,18	21,38	x
	11,16	3,19	21,32	x
	11,13	3,19	21,27	x
	10,73	3,02	x	3,58
	10,76	3,05	x	3,76
	10,79	3,09	x	3,66

Nonetheless, Figure 7 presents the predicted precipitation in comparison with the actual values from the base year in the test set. Data points clustering closely around the regression line indicate lower prediction errors. Notably, the generic model exhibits more pronounced prediction errors compared to the adjusted generic model, which excludes outliers. The adjusted model demonstrates improved performance in predicting daily precipitation during dry seasons.

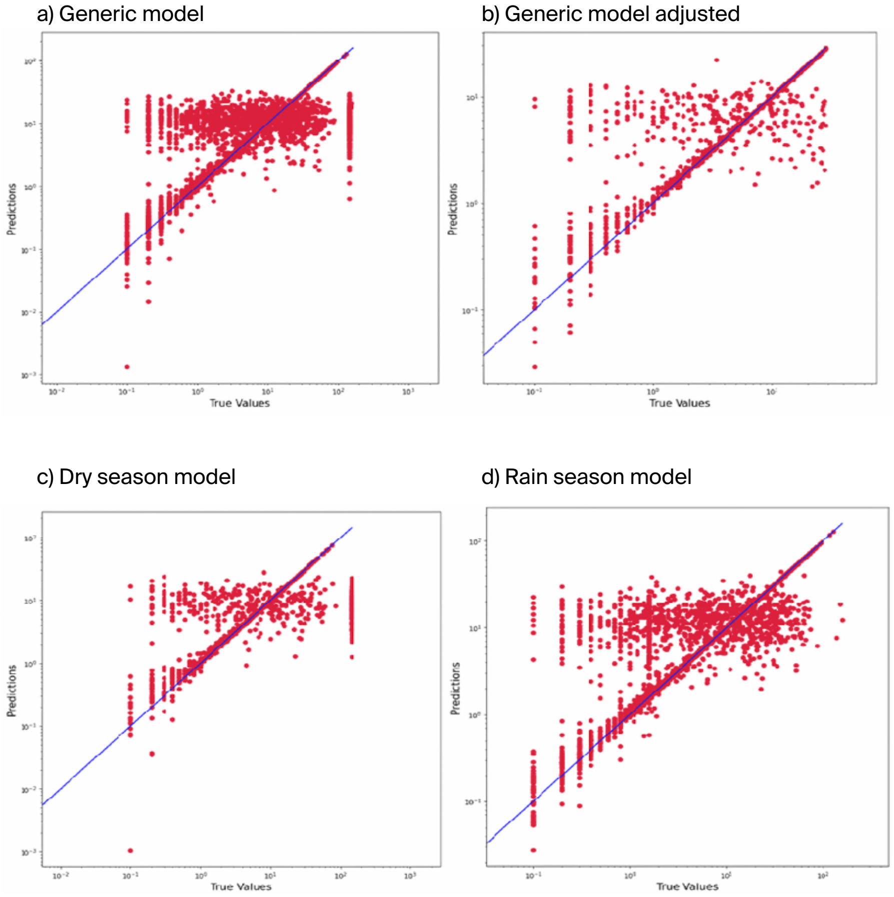
Given the variability of the data, constructing a model with very low error levels remains challenging. Figure 7 displays the behavior of all daily data points for each model, comparing the actual rainfall amounts with those predicted. The results show a higher concentration of values near the predicted range, while points that deviate significantly are considered outliers. This behavior is typical for the climate of the region studied.

Figure 7c illustrates the performance of the adjusted model, that is, with outliers removed, showing a smaller difference between predicted and actual values. This trend is reflected in the linear regression line shown in Figure 7.

4.3. Model Performance with Varying Training Sets

In the previous sections, the model evaluation process involved using historical data (1980–2008) to train the model and future data (2009–2019) to assess its performance. However, the past is only useful for predicting the future if models can operate effectively in both directions. With this in mind, a new set of experiments was conducted. In this alternative setup, we used data from 2009 to 2019 to train a new model and data from 1980 to 2008 to evaluate its performance. It is important to note that this analysis was carried out exclusively using data from the Generic Model (GM).

Figure 7. Predicted values versus actual values for each model. Source: own elaboration.



a) Generic model

b) Generic model adjusted

c) Dry season model

d) Rain season model

Source: own elaboration.

Table 8 summarizes the performance of the model trained on data from 2009 to 2019 and tested on data from 1980 to 2008. To assess the effectiveness of this new model, the results must be compared with those presented in the previous sections, where past data were used to predict future outcomes. By comparing the results in Table 5 and Table 8 for the classification task, we observe that the models achieved similar performance. A comparable outcome is observed when comparing the results in Table 6 and Table 8 for the regression task. The minor differences in metrics between the two experiments suggest that the models perform consistently well in both directions.

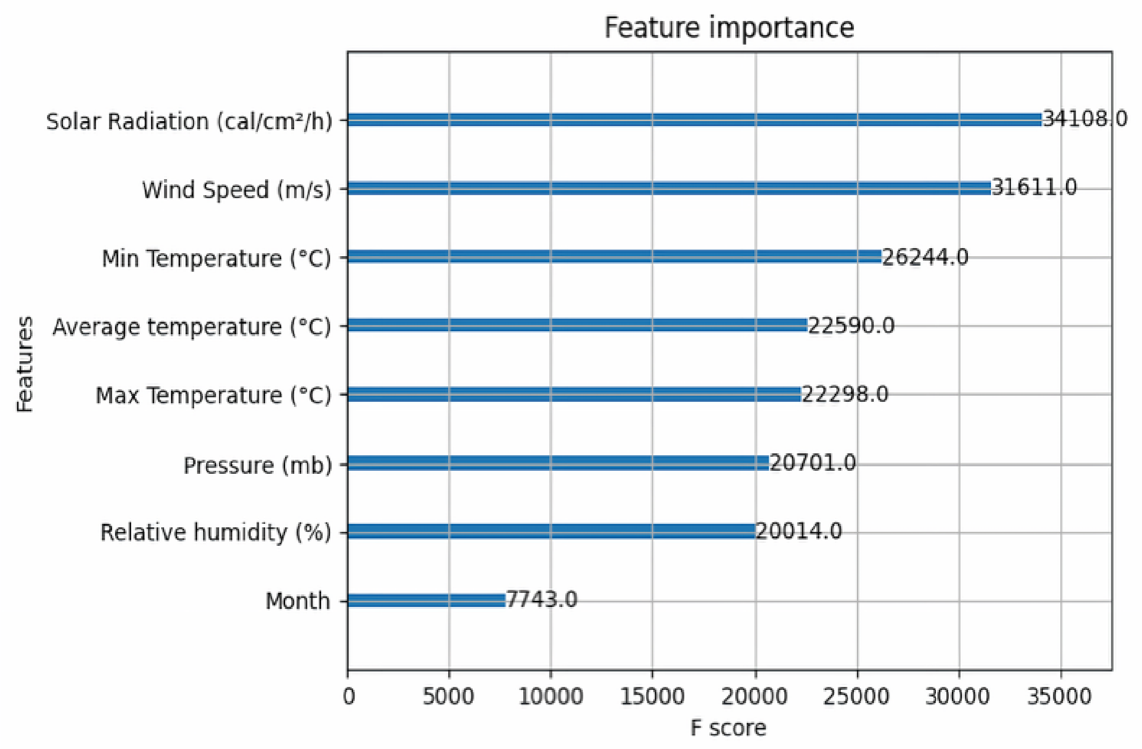
Table 8. Model's performance in both classification and regression tasks was evaluated using varying historical datasets.

Model (Classification)	**Precision**		**Recall**		**Accuracy**
Generic Model (GM)	0.828		0.776		0.872
Model (Regression)	MAE	MAPE	RMSE	MSE	R^2
Generic Model (GM)	11.217	8.689	15.202	231.087	-0.069

4.4. Feature importance analysis

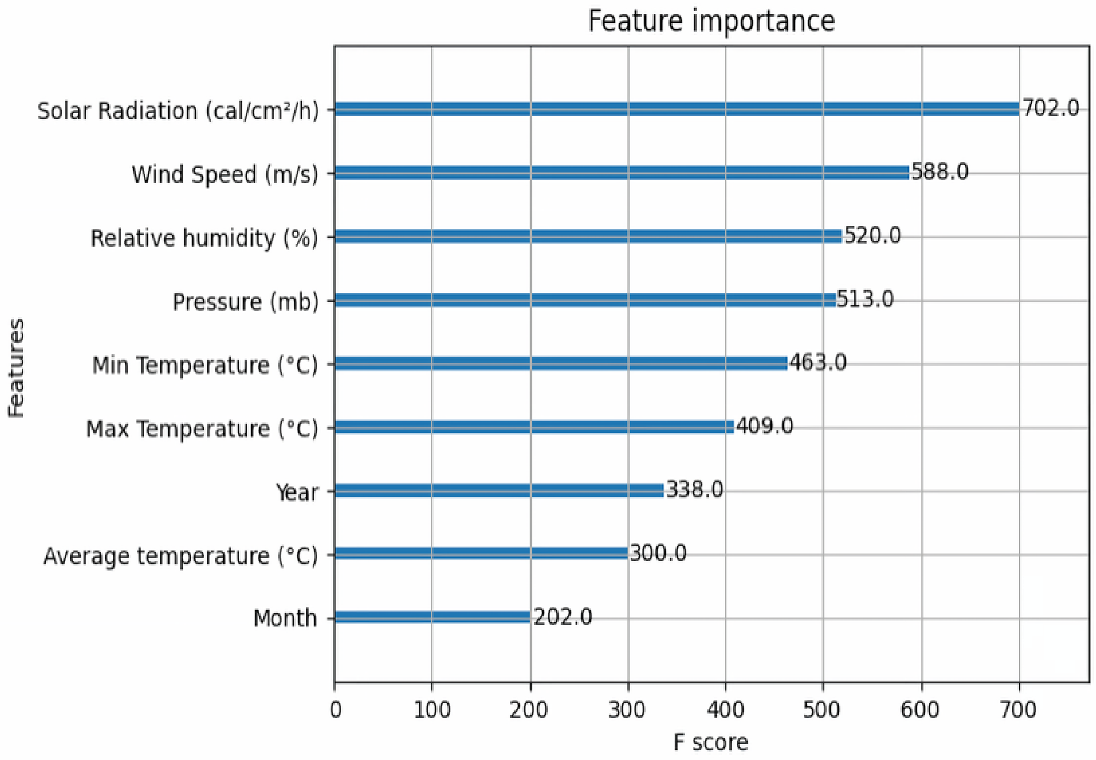
Figures 8 and 9 highlight the importance of the nine climatic elements specific to the study area. These elements encompass solar radiation (cal/cm^2/hour), wind speed (m/s), minimum and maximum temperatures (°C), atmospheric pressure (MB), relative humidity (%), average temperature (°C), Year and Month. This perspective underscores these characteristics' relevance in regression and classification tasks.

Figure 8. Feature importance analysis for the Generic model - regression task.



Source: own elaboration.

Figure 9. Feature importance analysis for the Generic model-classification task.



Source: own elaboration.

Although the models produce results that differ from the actual occurrences of daily precipitation, a strong connection emerges between solar radiation (cal/cm^2/hour), an atmospheric variable, and precipitation estimated forecasts. Similarly, wind speed values (km/h), especially those prevalent in the central region of the State of São Paulo, are intertwined with precipitation predictions alongside solar radiation.

The variations in minimum and maximum temperature (°C) and atmospheric pressure (mb) closely follow solar radiation and wind speed in terms of their predictive capability, aligning with the values of precipitation in the study area.

Moreover, daily surface data validates the accuracy of the ML predictions. The outcomes from the four ML models underscore the convergence of precipitation estimated forecasts with solar radiation and wind speed values. These variables rank highest in their impact on precipitation estimated forecasts, highlighting their pivotal role. Conversely, while influential, atmospheric pressure, average temperatures, and relative humidity do not take precedence in estimated forecasting daily data within the study area (Makarieva et al., 2014).

5. Discussion

The Machine Learning (ML) models produced by the XGBoost algorithm were standardized and validated by applying them to the seasonal data of the study area. This validation involved four tests: the generic model, adjusted generic model, rainy season model, and dry season model, following Hao & Bai (2023). This validation involved data from both the rainy and dry seasons. We evaluated the models using traditional classification and regression metrics, consistent with those used in related studies such as Anwar et al. (2021), Liyew & Melese (2021), and Manandhar et al. (2019).

In this article, the concepts of extreme precipitation and rainy and dry periods differ from those discussed in Liyew & Melese (2021) and Manandhar et al. (2019). Additionally, our choice of daily data aligns with other ML models in the literature. The primary difference lies in our dataset, the use of XGBoost, and the study of different models based on the rainy season (Anwar et al., 2021; Liyew & Melese, 2021; Manandhar et al., 2019; Dong et al., 2023).

The confusion matrix for each classification model (GM, GMA, RSM, and DSM) in predicting daily precipitation showed the lowest error rates, with the most notable improvements observed in the specific models for rainy and non-rainy days (RSM and DSM). Here, we can analyze each model's performance regarding the two outcomes: rain or no rain. According to Liyew & Melese (2021), the accuracy in predicting the amount of rain can increase if this data is considered. Therefore, the amount of rain and the presence or absence of precipitation leads to more effective predictions.

Additionally, the efficiency observed can be attributed to removing outlier values in the adjusted model for the rainy season, reflecting the significant precipitation volumes accumulated in southeastern South America and the influence of the South Atlantic Convergence Zone (SACZ) during the summer. Conversely, during the dry-winter period (April to September), precipitation volumes decrease due to the influence of the South Atlantic Subtropical High (SASH) in the study area.

Dry days (drought periods) serve as reliable indicators for accurately assessing and predicting drought risk based on daily precipitation data. These patterns also reaffirm the influence of air masses, orographic precipitation, and frontogenetic phenomena within the region's geographic and climatic dynamics.

Additionally, the climate pattern is marked by high temperatures alternating with milder temperatures during the months of June, July, and August, while precipitation is concentrated primarily in December, January, and February. This seasonal variability is clearly observed in two distinct periods: a hot and humid season, occurring between October and March (the hydrological year, with rainfall most concentrated in the first quarter), and a dry, milder season, extending from April to September.

Another interesting result is the high impact of two features on the model, solar radiation and wind speed, along with variations in minimum and maximum temperature and atmospheric pressure, in predicting precipitation values in southeastern South America.

The predicted precipitation values in our study are consistent with those reported for the Vietnam region by Pham et al. (2020). Similarly, they developed precipitation prediction models using artificial intelligence, and their results exhibited error margins comparable to those of our models (approximately 3 mm).

Although the present study was conducted in the southeast region of Brazil, which differs significantly from the context of Pham et al. (2020), the findings demonstrate the potential of the proposed method across diverse climatic conditions. Specifically, the model performed well in both rainy and drought-prone scenarios, indicating its robustness and applicability in various environmental settings.

Therefore, our results indicate that employing an ensemble algorithm (XGBoost) to predict the occurrence and amount of daily precipitation in our dataset has proven effective. Given the lack of robust estimated forecasting capabilities from satellite and radar data at the surface station, our proposed approach offers a viable solution for training models and predicting precipitation in real-time.

6. Conclusions and outlook

The estimated forecasts of daily precipitation in the central region of the State of Sao Paulo, based on the daily database in Itirapina/SP, generated by Machine Learning (ML) models for climatology, serve as a fundamental tool for drawing up future perspectives, prognoses, and climatological scenarios. In this work, we evaluated ML models to predict precipitation using data from a Brazilian meteorological station. We utilized maximum temperatures, minimum temperatures, average temperatures, wind speed, solar radiation, atmospheric pressure, and solar radiation of the estimate forecast day as input variables, with daily precipitation as the predictor (target class). We proposed four learning models using an ensemble algorithm (XGBoost) for two different machine learning tasks: binary classification (rain/no-rain) and regression. The learning models are associated with the following datasets: generic (GM), generic adjusted (GMA), rain season (RSM), and dry season (DSM).

The results of the four learning models for both tasks are promising. The GM, RSM, and DSM accuracy reached values equal to or higher than 90%. The DSM model, in particular, achieved accuracy values of around 96%, making it highly reliable in identifying days with no rain. The most efficient model in identifying days with rain was the RSM. Regarding predicting the daily amount of rain, the four learning models showed varying performances. The GMA and the RSM models exhibited satisfactory performance with low values. This result indicates the importance of working with models tailored to a specific scope. The four proposed precipitation estimated forecast models (generic, fitted generic, wet season, and dry season) have shown that the two elements that produce the most robust responses to real metrics are solar radiation and wind speed, along with minimum and maximum temperature and atmospheric pressure variations.

ML models for estimated forecasting, particularly the XGBoost algorithm, process climate data more efficiently, especially for continental climates. Additionally, XGBoost demonstrated superior performance (precision, recall, accuracy) in predicting precipitation regimes, particularly in generating the dry season model and the generic model adjusted for the study area. However, assessing drought risks, predicting landslides in risk areas, and accurately predicting daily precipitation and precipitation patterns using AI-based forecasts remain challenges in the four models.

The meteorological data used in this study include daily air temperatures, precipitation, wind speed, relative humidity, solar radiation, and more. These data were used to train and test models such as XGBoost and other machine learning (ML) models, which were developed to predict daily precipitation amounts with error rates of approximately 3 mm.

In Brazil, the methodology of climate estimated forecasting and the use of ML tools based on historical data are relatively recent in atmospheric research. In fact, it is feasible to enhance learning by incorporating ML, and this will improve the analysis of additional datasets in future work in this region. Furthermore, in terms of precipitation estimated forecasts for the study area, the corresponding months are characteristic of the tropical climate, as indicated by the historical climate data series.

Data availability

The code used for daily rainfall prediction, developed in Python (version 3.7) and based on scikit-learn, is available on GitHub: https://github.com/ditron6/USING-XGBOOST-MODELS-FOR-DAILY-RAINFALL-PREDICTION- . This repository was created by Rodrigo Sanches Miani (miani@ufu.br) in 2023 and contains notebooks and sample data. The author's experimental environment was as follows: OS: Linux Ubuntu; CPU: Intel Xeon E-2224G - 3.50GHz; RAM: 32GB. The daily rainfall data (raw_data.xlsx) was collected from the climatological station located at CRHEA (Center for Water Resources and Environmental Studies) / EESC (São Carlos Engineering School) / USP (University of São Paulo) in Itirapina, São Paulo.

CRediT authorship contribution statement

Rafael Grecco

Sanches: Conceptualization, Resources, Data Curation, Writing – Original Draft, and Writing. Miani, R. S. and Rios, P. A. T: Methodology, Writing – Original Draft, Software, Validation, and Formal Analysis, including data analysis, model coding, and programming. Santos, B. C.: Writing, Investigation, and Visualization. Moreira, R. M.: Writing, Investigation, and Visualization. Neves, G. Z. F.: Writing and Investigation. Bourscheidt, V.: Writing – Review and Editing, and Draft Preparation.

Acknowledgements and conflict of interest statement

We are grateful for the daily rainfall data (raw_data.xlsx), which was collected from the climatological station located at CRHEA (Center for Water Resources and Environmental Studies), EESC (São Carlos School of Engineering), USP (University of São Paulo), in Itirapina, São Paulo. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Althoff, D., Rodrigues, L. N., & Silva, D. D. (2022). Predicting runoff series in ungauged basins of the Brazilian Cerrado biome. *Environmental Modelling & Software*, 149, 105315. https://doi.org/10.1016/j.envsoft.2022.105315

Aires, U. R. V., Silva, D. D., Fernandes Filho, E. I., Rodrigues, L. N., Uliana, E. M., Amorim, R. S. S., Ribeiro, C. B. M., & Campos, J. A. (2023). Machine learning-based modeling of surface sediment concentration in Doce river basin. Journal of Hydrology, 619, 129320. https://doi.org/10.1016/j.jhydrol.2023.129320

Anwar, M. T., Winarno, E., Hadikurniawati, W., & Novita, M. (2021). Rainfall prediction using Extreme Gradient Boosting. Journal of Physics: Conference Series, 1869, 012078. https://doi.org/10.1088/1742-6596/1869/1/012078

Ardabili, S. F., Mosavi, A., Dehghani, M., & Várkonyi-Kóczy, A. R. (2019). Deep Learning and Machine Learning in Hydrological Processes, Climate Change and Earth Systems: A Systematic Review. https://doi.org/10.20944/preprints201908.0166.v1

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A Comparative Analysis of XGBoost. Artificial Intelligence Review, 54, 1937–1967. https://doi.org/10.1007/ s10462-020-09896-5

Bouras, E. H., Jarlan, L., Er-Raki, S., Balaghi, R., Amazirh, A., Richard, B., & Khabba, S. (2021). Cereal Yield Forecasting with Satellite Drought-Based Indices, Weather Data and Regional Climate Indices Using Machine Learning in Morocco. Remote Sensing, 13, 3101. https://doi.org/10.3390/rs13163101

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Bresciani, C., Boiaski, N. T., Ferraz, S. E. T., Rosso, F. V., Portalanza, D., de Souza, D. C., Kubota, P. Y., & Herdies, D. L. (2023). Brazilian Annual Precipitation Analysis Simulated by the Brazilian Atmospheric Global Model. Water, 15, 256. https://doi.org/10.3390/ w15020256

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chen, W., Huang, Y.-C., Lebar, K., & Bezak, N. (2023). A systematic review of the incorrect use of an empirical equation for the estimation of the rainfall erosivity around the globe. Earth-Science Reviews, 238, 104339. https://doi.org/10.1016/j.earscirev.2023.104339

Das, R., Chattoraj, S. L., Singh, M., & Bisht, A. (2024). Synergetic use of geospatial and machine learning techniques in modelling landslide susceptibility in parts of Shimla to Kinnaur National Highway, Himachal Pradesh. Modeling Earth Systems and Environment. https://doi.org/10.1007/s40808-024-01993-6

Dong, J., Peng, J., He, X., Cai, W., Gao, L., & Xiao, Z. (2023). Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China. Engineering Applications of Artificial Intelligence, 117, 105579. https://doi.org/10.1016/j.engappai.2022.105579

dos Santos, B. C., Sanches, R. G., de Melo Bolleli, T., Neves, G. Z. F., Pereira, D. N. B., & Tech, A. R. B. (2022). On the quality of satellite-based precipitation estimates for time series analysis at the central region of the state of São Paulo, Brazil. Theoretical and Applied Climatology. https://doi.org/10.1007/s00704-022-04287-y

Facco, M., Campos, M. A. A., Vargas, D. S., Silveira, R. B., & Bisognin, C. (2020). Algoritmos de Machine Learning Aplicados na Ocorrência de Chuvas na Cidade de Santa Maria. Ciência e Natura, 42, 28. https://doi.org/10.5902/2179460X40537

Ghafarian, F., Wieland, R., Lüttschwager, D., & Nendel, C. (2022). Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. Environmental Modelling & Software, 156, 105466. https://doi.org/10.1016/j.envsoft.2022.105466

Hao, R., & Bai, Z. (2023). Comparative Study for Daily Streamflow Simulation with Different Machine Learning Methods. Water, 15, 1179. https://doi.org/10.3390/ w15061179

He, R., Zhang, L., & Chew, A. W. Z. (2022). Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning. Knowledge-Based Systems, 251, 109125. https://doi.org/10.1016/j.knosys.2022.109125

Heydarizad, M., Pumijumnong, N., Sorí, R., Salari, P., & Gimeno, L. (2022). Fractional Importance of Various Moisture Sources Influencing Precipitation in Iran Using a Comparative Analysis of Analytical Hierarchy Processes and Machine Learning Techniques. Atmosphere, 13, 2019. https://doi.org/10.3390/atmos13122019

Ibrahem Ahmed Osman, A., Najah Ahmed, A., Chow, M. F., Feng Huang, Y., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 12, 1545–1556. https://doi.org/10.1016/j.asej.2020.11.011

Knighton, J., Pleiss, G., Carter, E., Lyon, S., Walter, M. T., & Steinschneider, S. (2019). Potential Predictability of Regional Precipitation and Discharge Extremes Using Synoptic-Scale Climate Information via Machine Learning: An Evaluation for the Eastern Continental United States. Journal of Hydrometeorology, 20, 883–900. https://doi.org/10.1175/JHM-D-18-0196.1

Liu, F., Wang, X., Sun, F., Wang, H., Wu, L., Zhang, X., Liu, W., & Che, H. (2022). Correction of Overestimation in Observed Land Surface Temperatures Based on Machine Learning Models. Journal of Climate, 35, 5359–5377. https://doi.org/10.1175/JCLI-D-21-0447.1

Liu, Y., Zhao, Q., Yao, W., Ma, X., Yao, Y., & Liu, L. (2019). Short-term rainfall forecast model based on the improved BP–NN algorithm. Scientific Reports, 9, 19751. https://doi.org/10.1038/s41598-019-56452-5

Liyew, C. M., & Melese, H. A. (2021). Machine learning techniques to predict daily rainfall amount. Journal of Big Data, 8, 153. https://doi.org/10.1186/s40537-021-00545-4

Ma, C., Yao, J., Mo, Y., Zhao, Y., Jiang, Y., & Jiang, Z. (2024). Prediction of summer precipitation via machine learning with key climate variables: A case study in Xinjiang, China. Journal of Hydrology: Regional Studies, 56, 101964. https://doi.org/10.1016/ j.ejrh.2024.101964

Makarieva, A. M., Gorshkov, V. G., Sheil, D., Nobre, A. D., Bunyard, P., & Li, B.-L. (2014). Why Does Air Passage over Forest Yield More Rain? Examining the Coupling between Rainfall, Pressure, and Atmospheric Moisture Content. Journal of Hydrometeorology, 15, 411–426. https://doi.org/10.1175/JHM-D-12-0190.1

Manandhar, S., Dev, S., Lee, Y. H., Meng, Y. S., & Winkler, S. (2019). A Data-Driven Approach for Accurate Rainfall Prediction. IEEE Transactions on Geoscience and Remote Sensing, 57, 9323–9331. https://doi.org/10.1109/TGRS.2019.2926110

Monego, V. S., Anochi, J. A., & de Campos Velho, H. F. (2022). South America Seasonal Precipitation Prediction by Gradient-Boosting Machine-Learning Approach. Atmosphere, 13, 243. https://doi.org/10.3390/atmos13020243

Mu, Y., Biggs, T., & Shen, S. S. P. (2021). Satellite-based precipitation estimates using a dense rain gauge network over the Southwestern Brazilian Amazon: Implication for identifying trends in dry season rainfall. Atmospheric Research, 261, 105741. https:// doi.org/10.1016/j.atmosres.2021.105741

Muhammad, A., Evenson, G. R., Unduche, F., & Stadnyk, T. A. (2020). Climate Change Impacts on Reservoir Inflow in the Prairie Pothole Region: A Watershed Model Analysis. Water, 12, 271. https://doi.org/10.3390/w12010271

Nakhaei, M., Mohebbi Tafreshi, A., & Saadi, T. (2023). An evaluation of satellite precipitation downscaling models using machine learning algorithms in Hashtgerd Plain, Iran. Modeling Earth Systems and Environment, 9, 2829–2843. https://doi.org/10.1007/s40808-022-01678-y

Nielsen, D. (2016). Tree boosting with XGBoost-why does XGBoost win "every" machine learning competition? [Dissertação de Mestrado]. NTNU.

Nguyen, D. H., Hien Le, X., Heo, J.-Y., & Bae, D.-H. (2021). Development of an Extreme Gradient Boosting Model Integrated With Evolutionary Algorithms for Hourly Water Level Prediction. IEEE Access, 9, 125853–125867. https://doi.org/10.1109/ACCESS.2021.3111287

Oliveira, G., Pedrollo, O., & Castro, N. (2014). O Desempenho das Redes Neurais Artificiais (RNAs) para Simulação Hidrológica Mensal. Revista Brasileira de Recursos Hídricos, 19, 251–265. https://doi.org/10.21168/rbrh.v19n2.p251-265

Parmar, A., Mistree, K., & Sompura, M. (2017). Machine Learning Techniques For Rainfall Prediction: A Review.

Pham, B. T., Le, L. M., Le, T.-T., Bui, K.-T. T., Le, V. M., Ly, H.-B., & Prakash, I. (2020). Development of advanced artificial intelligence models for daily rainfall prediction. Atmospheric Research, 237, 104845. https://doi.org/10.1016/j.atmosres.2020.104845

Prodhan, F. A., Zhang, J., Hasan, S. S., Pangali Sharma, T. P., & Mohana, H. P. (2022). A review of machine learning methods for drought hazard monitoring and forecasting: Current research trends, challenges, and future research directions. Environmental Modelling & Software, 149, 105327. https://doi.org/10.1016/j.envsoft.2022.105327

Qian, Q., Jia, X., Lin, H., & Zhang, R. (2021). Seasonal Forecast of Nonmonsoonal Winter Precipitation over the Eurasian Continent Using Machine-Learning Models. Journal of Climate, 34, 7113–7129. https://doi.org/10.1175/JCLI-D-21-0113.1

Qian, Q. F., Jia, X. J., & Lin, H. (2020). Machine Learning Models for the Seasonal Forecast of Winter Surface Air Temperature in North America. Earth and Space Science, 7, e2020EA001140. https://doi.org/10.1029/2020EA001140

Ramirez, S., & Lizarazo, I. (2017). Detecting and tracking mesoscale precipitating objects using machine learning algorithms. International Journal of Remote Sensing, 38, 5045– 5068. https://doi.org/10.1080/01431161.2017.1323280

Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. Journal of Hydrology, 414–415, 284–293. https://doi.org/10.1016/j.jhydrol.2011.10.039

Rodrigues, D. T., Gonçalves, W. A., Silva, C. M. S. E., Spyrides, M. H. C., & Lúcio, P. S. (2023). Imputation of precipitation data in northeast Brazil. Anais da Academia Brasileira de Ciências, 95, e20210737. https://doi.org/10.1590/0001-3765202320210737

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., & Bengio, Y. (2022). Tackling Climate Change with Machine Learning. ACM Computing Surveys, 55, 42:1-42:96. https://doi.org/10.1145/3485128

Sachindra, D. A., Ahmed, K., Rashid, M. M., Shahid, S., & Perera, B. J. C. (2018). Statistical downscaling of precipitation using machine learning techniques. Atmospheric Research, 212, 240–258. https://doi.org/10.1016/j.atmosres.2018.05.022

Sanches, R. G. (2015). As chuvas na região de São Carlos/SP: estudo do comportamento pluviométrico a partir de dados de estações climatológicas, 1993-2014 [Tese de Doutorado, Universidade de São Paulo]. https://doi.org/10.11606/D.18.2015.tde-16112015-100925

Sanches, R. G., Neves, G. Z. F., Santos, B. C., Silva, M. S. D., Pereira, D. N. B., & Tech, A. R.B. (2018). Intense Rainfall in São Carlos/SP: Determination of Threshold Values Using Climate Indices and Their Spatio-Temporal Repercussion. American Journal of Climate Change, 7, 388–401. https://doi.org/10.4236/ajcc.2018.73023

Sanches, R. G., Santos, B. C. D., Miani, R. S., Neves, G. Z. F., Silva, M. S. D., & Tech, A. R. B. (2020). Analysis of Daily Rainfall in São Carlos/SP, Brazil over 1979-2017 Using Laplace Trend Test. Journal of Geoscience and Environment Protection, 8, 104–125. https:// doi.org/10.4236/gep.2020.87006

Shilong, Z., Jianwei, L., Dawei, L., Pengfei, Z., Yuanfang, D., & Mengxuan, Z. (2021). Machine learning model for sales forecasting by using XGBoost. In IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 480-483). IEEE.

Stanley, T. A., Kirschbaum, D. B., Sobieszczyk, S., Jasinski, M. F., Borak, J. S., & Slaughter, S. L. (2020). Building a landslide hazard indicator with machine learning and land surface models. Environmental Modelling & Software, 129, 104692. https://doi.org/10.1016/ j.envsoft.2020.104692

Tan, W. Y., Lai, S. H., Teo, F. Y., & El-Shafie, A. (2022). State-of-the-Art Development of Two- Waves Artificial Intelligence Modeling Techniques for River Streamflow Forecasting. Archives of Computational Methods in Engineering, 29, 5185–5211. https://doi.org/10.1007/s11831-022-09763-2

Tarwidi, D., Pudjaprasetya, S. R., Adytia, D., & Apri, M. (2023). An optimized XGBoost- based machine learning method for predicting wave run-up on a sloping beach. MethodsX, 10, 102119. https://doi.org/10.1016/j.mex.2023.102119

Tian, D., He, X., Srivastava, P., & Kalin, L. (2022). A hybrid framework for forecasting monthly reservoir inflow based on machine learning techniques with dynamic climate forecasts, satellite-based data, and climate phenomenon information. Stochastic Environmental Research and Risk Assessment, 36, 2353–2375. https://doi.org/10.1007/ s00477-021-02023-y

Tolentino, M. (2007). Estudo crítico sobre o clima da região de São Carlos. EdUFSCar; Imprensa Oficial do Estado de São Paulo.

Wang, Y., Pan, Z., Zheng, J., Zhou, W., & Jia, J. (2019). A hybrid ensemble method for pulsar candidate classification. Astrophysics and Space Science, 364, 139. https:// doi.org/10.1007/s10509-019-3602-4

Xu, J., Jiang, Y., & Yang, C. (2022). Landslide Displacement Prediction during the Sliding Process Using XGBoost, SVR and RNNs. Applied Sciences, 12, 6056. https://doi.org/10.3390/app12126056

Zhou, J., & Lau, K.-M. (2001). Principal modes of interannual and decadal variability of summer rainfall over South America. International Journal of Climatology, 21, 1623– 1644. https://doi.org/10.1002/joc.700

Zhou, S., Liu, Z., Wang, M., Gan, W., Zhao, Z., & Wu, Z. (2022). Impacts of building configurations on urban stormwater management at a block scale using XGBoost. Sustainable Cities and Society, 87, 104235. https://doi.org/10.1016/j.scs.2022.104235

Zhou, Z., Zhao, L., Lin, A., Qin, W., Lu, Y., Li, J., Zhong, Y., & He, L. (2021). Exploring the potential of deep factorization machine and various gradient boosting models in modeling daily reference evapotranspiration in China. Arabian Journal of Geosciences, 13, 1287. https://doi.org/10.1007/s12517-020-06293-8