A machine learning and statistical learning-based pipeline to perform multipoint rainfall forecasting

Eduardo Carvalho[1] **, Ewerton Oliveira**[1,2] **, Rafael Rocha**[1,2]**,**
Nikolas Carneiro[1]**, Renata Tedeschi**[1] **, Ronnie Alves**[1]

[1]Instituto Tecnológico Vale, Belém PA, BRAZIL

[2]Federal University of Pará, Belém PA. BRAZIL

{eduardo.costa.carvalho, renata.tedeschi, ronnie.alves,

nikolas.carneiro}@itv.org

{ewerton.oliveira, rafael.lima.rocha}@pq.itv.org

Abstract. *Analyzing and predicting precipitation is crucial for society, particularly when extreme rainfall and floods occur. Such events impact socioeconomic structures and can lead to fatalities. Understanding rain formation and associated variables helps develop predictive models for precipitation levels, aiding decision-making in production chains and urban mobility. Recent advancements in this field result from increased computer processing power and the availability of meteorological data worldwide. This study focuses on evaluating SARIMA, RNN, and XGBoost models for predicting monthly rainfall along a northern Brazilian railway. Using lagged precipitation values, SARIMA performed better for 11 out of 13 points (with R^2 ranging from 0.704 to 0.817), while RNN outperformed in the remaining points (15.39% of the evaluated points).*

Resumo. *Analisar e prever a precipitação é crucial para a sociedade, principalmente quando ocorrem chuvas extremas e inundações. Tais eventos impactam as estruturas socioeconômicas e podem levar a fatalidades. O entendimento da formação das chuvas e das variáveis associadas auxilia no desenvolvimento de modelos preditivos dos níveis de precipitação, auxiliando na tomada de decisões nas cadeias produtivas e na mobilidade urbana. Avanços recentes neste campo resultam do aumento do poder de processamento do computador e da disponibilidade de dados meteorológicos em todo o mundo. Este estudo se concentra na avaliação dos modelos SARIMA, RNN e XGBoost para prever a precipitação mensal ao longo de uma ferrovia no norte do Brasil. Usando valores de precipitação defasados, SARIMA teve melhor desempenho em 11 dos 13 pontos (com R^2 variando de 0,704 a 0,817), enquanto RNN superou nos pontos restantes (15,39% dos pontos avaliados).*

1. Introduction

Climate change has been occurring on planet Earth since its inception. However, in recent years, this issue has gained significant attention in various studies [Qerimi and Sergi 2022] [Berrang-Ford et al. 2021]. The study of rainfall patterns serves as an indicator of seasonal change in weather. Rainfall refers to the precipitation of rain

reaching the ground [Su et al. 2022] [Moazzam et al. 2022], which takes place in a specific location over a given period. Precipitation is typically measured in millimeters per specific time scale (mm/h, mm/day, mm/month, etc.), providing an estimate of the amount of rain that has fallen during that particular time frame.

One area of this article focuses on utilizing precipitation data to predict future rainfall amounts. For these predictions, various statistical models can be employed in weather forecasting. In recent years, the application of machine learning (ML) methods has garnered attention for estimating rainfall levels within a specific time horizon [Berrang-Ford et al. 2021].

The significance of predicting precipitation extends to various aspects of life. It ranges from simple decisions like planning to leave the house on a cloudy day to major companies organizing their activities based on this natural phenomenon [Yin et al. 2022]. Another critical application of rainfall prediction is related to population safety, as major natural disasters often stem from this meteorological occurrence.

Floods resulting from extreme precipitation have wide-ranging social and economic consequences. A study conducted by The Swiss Re Group estimated a loss of US$ 35 Billion in the first half of 2022, which is 22% higher than the average of the past 10 years [Group 2022]. In the Latin American region, research by Fang *et al.* [Fang et al. 2015] indicates that Brazil accounts for approximately 10% of estimated flood-related deaths. The study also highlights that Brazil and Argentina are the countries most affected by the economic risk of flood-related losses [Fang et al. 2015].

Globally, floods rank among the climatic events that have the greatest impact on society, both in terms of economic consequences and human fatalities [Cerveny et al. 2017]. In the United States alone, a study conducted between 2010 and 2022 indicates that floods have affected the lives of 1352 individuals [NWS 2023].

Companies have shown increasing interest in predicting precipitation to enhance their operations. Consequently, they have been actively seeking intelligent and impactful solutions for forecasting this meteorological variable. Company investments in intelligent methods for prediction span across multiple areas of knowledge, ranging from rainfall forecasting to financial market analysis [Santos and Qin 2019]. This growing focus on intelligent methods for rainfall forecasting represents a novel approach to identify patterns and take informed actions based on the results obtained.

Investment in research focused on machine learning has been widely recognized as an important aspect, leading to numerous applications aimed at asset maintenance. Among these assets, the railroad holds significant importance due to its role in facilitating massive transportation processes. To illustrate, consider a railway connecting two states in the north-northeast region of Brazil, stretching approximately 970km in length. The railway fleet consists of 300 locomotives and around 20,000 wagons, organized into compositions of 330 wagons each. Several studies have proposed the use of machine learning in the maintenance of these assets, as seen in works such as [Fernandes et al. 2018] and [Rocha et al. 2019].

In order to address the problem of rainfall forecast, we intend to employ statistical and machine learning techniques, including Seasonal Autoregressive Integrated Moving Average (SARIMA), Recurrent Neural Network (RNN), and eXtreme Gradient Boost-

ing (XGBoost). These methods will be utilized to forecast rainfall using only a single meteorological variable. The objective is to determine the best-performing method for each approach, considering the evaluation metrics of mean squared error (RMSE) and the Coefficient of Determination R^2.

The present work is structured into the following sections: Section 2: This section provides an overview of application of machine learning methods in climate studies. Section 3: Problem Statement: The problem of predicting seasonal rainfall is presented and discussed in this section. The specific challenges and objectives of the research are outlined. Section 4: Methodology: The methodology employed in this study is described in detail. It encompasses the extraction of the precipitation database, as well as the process of conducting statistical analyses and implementing machine learning algorithms. Section 5: Rainfall Prediction using a Single Variable: This section presents the results of rainfall prediction using only a single meteorological variable. The performance of different machine learning algorithms, such as SARIMA, RNN, and XGBoost, is evaluated using the defined metrics. Section 6: Conclusions and Future Directions.

2. Machine Learning methods and applications in climate

The research conducted by McGuffie and Henderson-Sellers [McGuffie and Henderson-Sellers 2001] provides insights into the evolution of numerical climate models over four decades of research. The study highlights the advancements in information processing capabilities, with models being capable of reaching more than 1 Teraflop (trillions of floating-point operations per second) by the early 2000s.

Overall, the research by McGuffie and Henderson-Sellers [McGuffie and Henderson-Sellers 2001] emphasizes the continuous development of numerical climate models, the importance of parameterization in precipitation predictions, and the challenges associated with predicting precipitation accurately due to the complex and chaotic nature of climate systems.

When it comes to predicting rainfall on a seasonal scale, the complexity of the problem decreases compared to shorter-term predictions. In this context, identifying patterns and establishing correlations between variables in meteorology become crucial.

While high-resolution models and extensive computational power are often required for detailed short-term predictions, seasonal-scale predictions can benefit from a more focused analysis of key variables and their correlations. These variables may include sea surface temperatures, atmospheric pressure patterns, moisture content, and other large-scale climate indicators.

By examining the relationships and correlations between these variables, researchers can identify patterns and trends that may influence rainfall patterns over a particular season. Statistical and machine learning techniques can be employed to analyze historical data and identify meaningful connections between variables.

The study conducted by Zhou *et al.* [Zhou et al. 2021] focuses on the comparative analysis of different machine learning models, including the Autoregressive Integrated Moving Average (ARIMA) model, for forecasting purposes. The ARIMA model is a widely used method that combines autoregressive and moving average components to

predict future values based on past data.

In their research, Zhou *et al.* [Zhou et al. 2021] compare the performance of the ARIMA model with other machine learning models. The results indicate that the machine learning models, as a whole, outperformed the ARIMA model in terms of forecasting accuracy. Additionally, among the machine learning models considered, the Random Forest (RF) model exhibited relatively better performance compared to the other models.

The study conducted by Chhetri *et al.* [Chhetri et al. 2020] explores the application of Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units for rainfall prediction. Additionally, the authors investigate the effectiveness of combining LSTM with another type of RNN called Gated Recurrent Unit (GRU) in improving rainfall predictability.

The proposed model, which combines LSTM and GRU, demonstrates significant improvement in rain predictability compared to using LSTM alone. Specifically, the model achieved a 41% reduction in Mean Square Error (MSE) for rainfall prediction [Chhetri et al. 2020].

The study conducted by Monego *et al.* [Monego et al. 2022] introduces the use of XGBoost, an efficient gradient boosting algorithm, as a machine learning model for seasonal rainfall forecasting in South America. The researchers aimed to improve the precipitation identification performance by leveraging XGBoost and optimizing its parameter configuration using the Optuna framework through Bayesian optimization.

By applying this XGBoost model to a database spanning from 1980 to 2020, the researchers observed improved predictability in three seasons compared to a Tensor Flow deep neural network. However, it is important to note that the proposed model did not achieve superior results for all seasons of the year [Monego et al. 2022].

To address this limitation and further enhance prediction accuracy, the authors recommend the use of a hybrid model. Combining multiple models or approaches can potentially leverage the strengths of different methods, leading to more robust predictions across various seasons and meteorological conditions.

3. Statment of the problem on predicting seasonal rainfall for specific points in a railroad

The use of machine learning methods in precipitation identification and prediction has gained significant attention in recent years. The works discussed in the previous section demonstrate the application of various machine learning models, such as multiple linear regression, ARIMA, LSTM, GRU, and XGBoost, for rainfall forecasting and identification.

These models utilize historical data and relevant meteorological variables to identify patterns, classify rainfall types, and make predictions at specific points or locations. By leveraging the power of machine learning algorithms, these studies aim to improve the accuracy and reliability of precipitation forecasts, which can have significant implications for various sectors and decision-making processes.

The models consider factors such as climate variables, atmospheric conditions, historical patterns, and even geographical features to capture the complex relationships

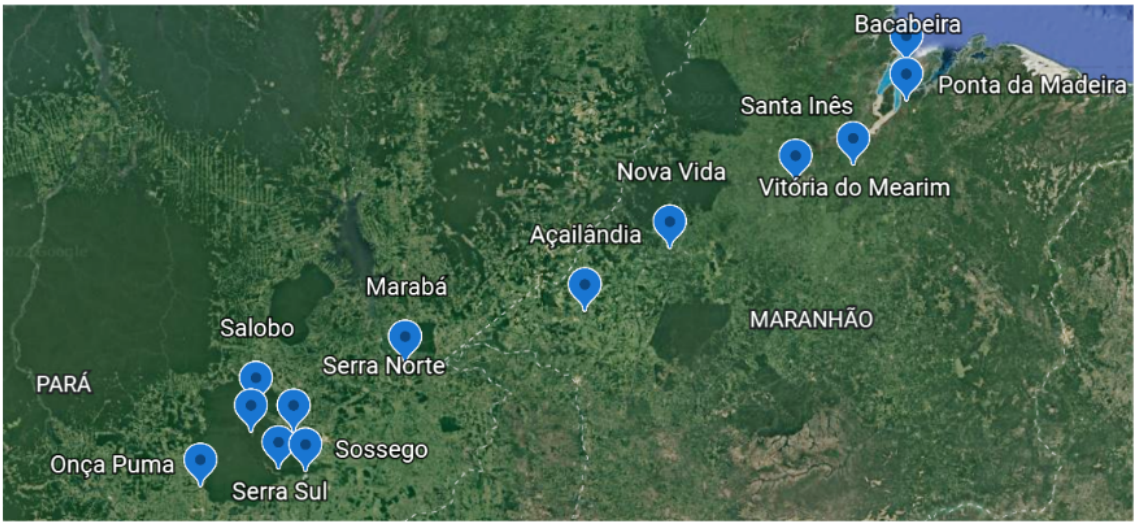

Figure 1. Map identification of the 13 points.

and dynamics associated with rainfall. This allows for the identification of precipitation patterns and the prediction of future rainfall events, enabling better preparedness and planning in various fields, including agriculture, disaster management, water resource management, and urban planning.

The main problem addressed in this work is to examine whether the precipitation variable alone, measured at specific points along a train track, can be used to project or predict the amount of precipitation using statistical probability methods. The objective is to determine if the historical data of precipitation at these points can provide sufficient information to estimate or forecast future rainfall levels.

By analyzing the relationship between past precipitation data and future rainfall, the study aims to assess the predictive capability of statistical models such as SARIMA, RNN, and XGBoost. These models will be evaluated based on their performance in predicting monthly precipitation at selected points of interest along the train track in northern Brazil.

The study will investigate whether the lagged values of precipitation alone, without considering additional meteorological variables, are sufficient to achieve predictions. The comparison between the different models will provide insights into the effectiveness of each approach in capturing the underlying patterns and variability in precipitation.

The research outcome will contribute to understanding the feasibility and reliability of using statistical probability methods with precipitation data as the sole input for precipitation prediction.The main problem addressed in this work is to examine whether the precipitation variable alone, measured at specific points along a train track, can be used to project or predict the amount of precipitation using statistical probability methods. The objective is to determine if the historical data of precipitation at these points can provide sufficient information to estimate or forecast future rainfall levels.

By analyzing the relationship between past precipitation data and future rainfall, the study aims to assess the predictive capability of statistical models such as SARIMA, RNN, and XGBoost. These models will be evaluated based on their performance in pre-

dicting monthly precipitation at selected points of interest along the train track in northern Brazil.

The study will investigate whether the lagged values of precipitation alone, without considering additional meteorological variables, are sufficient to achieve predictions. The comparison between the different models will provide insights into the effectiveness of each approach in capturing the underlying patterns and variability in precipitation.

The research outcome will contribute to understanding the feasibility and reliability of using statistical probability methods with precipitation data as the sole input for precipitation prediction.

A railroad chosen between two Brazilian states was divided into 13 points indicated by the Fig. 1.

Fig. 1 visually represents the geographical locations of the 13 identified points along the railway track where the precipitation data will be monitored and analyzed. These points are strategically selected to cover a specific region of interest and ensure adequate coverage for the study.

By monitoring the precipitation at these points, the study aims to gain insights into the local precipitation patterns and understand how they may impact the production flow along the railway. By examining the historical precipitation data at these locations, researchers can identify any potential risks or disruptions caused by extreme or prolonged rainfall events.

This monitoring process is crucial for industries or organizations relying on the efficient transportation of goods and services along the railway. By being aware of the precipitation levels in the region, proactive measures can be taken to mitigate any adverse effects caused by heavy rainfall, such as implementing drainage systems, adjusting schedules, or implementing contingency plans.

4. The methodology for rainfall forecasting

This section outlines the methodologies employed to predict rainfall at the points of interest along the railway track. It encompasses various steps, including database extraction, pre-processing of data, and model generation. Fig. 2 provides a visual summary of the part of the process.

4.1. Pre-processing the data

For the precipitation forecasting at the 13 specific points along the railroad track, the ERA-5 reanalysis precipitation database from ECMWF was utilized. This database offers a comprehensive collection of oceanic, atmospheric, and surface variables, which are derived from various models and data assimilation techniques, incorporating historical observations. The data is provided in NetCDF format.

In this work, the monthly total precipitation (TP) data from the ERA-5 reanalysis database was extracted for the period spanning from 1979 to 2020. This data served as the basis for training and testing the rainfall prediction models. To process and analyze the precipitation data, the Python programming language was employed.

Fig. 3 provides a visual representation of the time series of total precipitation (TP) over the years in the Serra Norte (SEN) region, showcasing the seasonal patterns

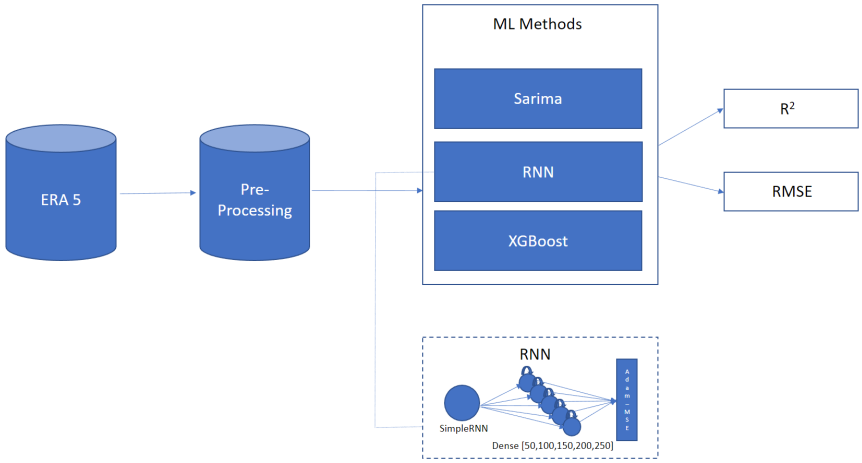

Figure 2. Work methodology. From the online data repository part of ERA5, its transformations for inputs to different machine learning methods and its precipitation assessments.

of rainfall along the railroad. It demonstrates how the precipitation varies throughout the months and years, highlighting any notable trends or fluctuations.

The TP data in the ERA-5 reanalysis database is originally presented on a scale of meters per day (m/day). However, for the purpose of monthly forecasting, it is more suitable to convert the data to millimeters per month (mm/month). This conversion allows for a consistent and standardized scale for monthly precipitation measurements.

The pre-processing stage, as illustrated in Fig. 2, involves two steps to prepare the total precipitation (TP) data for further analysis and forecasting. The first step entails multiplying the TP values by a scale factor specific to each month. This step is performed for the TP time series of all 13 points of interest along the railroad.

The scale factors are determined based on the number of days in each month. For months with 31 days, the scale factor is set to 31.000. For months with 30 days, the scale factor is 30.000. February, being a variable-length month depending on leap years, has a scale factor of 28.000 applied consistently across all years.

By applying these scale factors, the TP values are adjusted to a standardized unit of millimeters per month (mm/month). The multiplication of TP with the appropriate scale factors allows for the normalization and uniformity of the data.

An advantage of the methodology of listing points is the decrease in the difficulty of the model to learn precipitation patterns, however a disadvantage is the loss of characterization of the region, because as rain clouds are very dynamic it may be that at one point it has a precipitation characteristic , and in a few meters or kilometers it has another pattern, being in just one city/region.

In the second step of the pre-processing stage, the total precipitation (TP) time series for each point of interest along the railroad is split into two datasets: the training dataset and the test dataset.

The training dataset comprises the TP data from 1979 to 2010, spanning a 32-

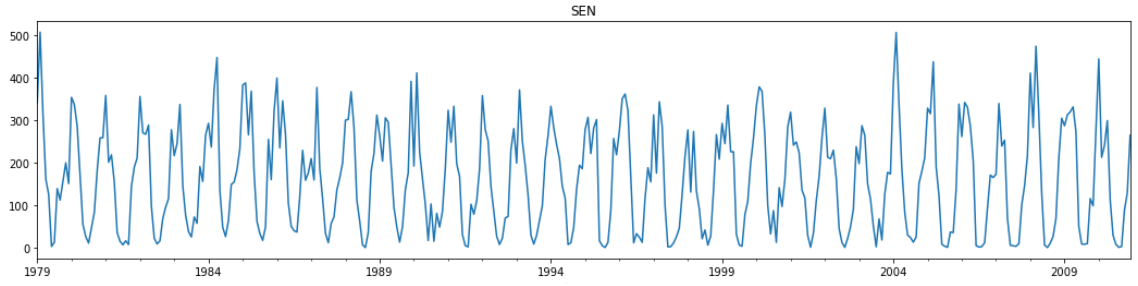

Figure 3. Time series of monthly precipitation data at latitude -6.05 and Longitude 309.83 Serra Norte (SEN).

year period. This data is used to train the machine learning models, allowing them to learn the patterns and relationships between the input variables and the target variable (precipitation) over this historical period. The models utilize this training data to establish their internal parameters and optimize their performance.

On the other hand, the test dataset encompasses the TP data from 2011 to 2020, covering a 10-year period. This data is held out from the training process and is used to evaluate the performance and generalization ability of the trained models.

4.2. Evaluated algorithms: RNN, XGBoost and SARIMA

In the precipitation prediction task, three models were employed: SARIMA, RNN (Recurrent Neural Network), and XGBoost. Hyperparameter tuning for each model was performed using grid search during the training stage to optimize their performance.

4.2.1. RNN parameters

For the RNN algorithm, the grid search was conducted to find the optimal number of neurons in the hidden layer. The number of neurons was varied across different scales: 50, 100, 150, 200, and 250. The activation functions used in the dense layer of the SimpleRNN model were Relu-Sigmoid.

The RNN model was trained using the TensorFlow package for Python. It underwent 25 epochs, which represents the number of times the entire training dataset was passed through the model during training. This parameter affects the convergence and learning capability of the model.

These neurons were made available in just one layer with the SimpleRNN method of Tensorflow[1], as already mentioned. Its second layer of the network was the activation functions (ReLU and Sigmoid), as previously identified, in addition its parameterization was optimized with the ADAM algorithm, and the calculation of its error with the MSE.

[1]https://www.nablasquared.com/creating-a-simple-rnn-from-scratch-with-tensorflow/

4.2.2. XGBoost parameters

In the case of the XGBoost algorithm, a grid search was conducted to determine the optimal hyperparameters for the model. The grid search evaluated different values for two key hyperparameters: the number of estimators and the maximum depth.

For the number of estimators, the grid search considered the following values: 1500, 2000, 3000, 4000, and 6000. The number of estimators refers to the number of boosting rounds or decision trees that are sequentially added to the model.

The maximum depth parameter, on the other hand, was evaluated for values of 10, 15, 20, and 30. The maximum depth determines the maximum depth of each decision tree in the ensemble.

During the grid search, the learning rate was fixed at 0.01, which controls the contribution of each tree in the ensemble. The learning rate affects the speed and accuracy of the model's convergence.

4.2.3. SARIMA parameters

The SARIMA model, as described by Eq. (1), combines the autoregressive integrated moving average (ARIMA) model with seasonality. It is a popular time series forecasting model that takes into account both non-seasonal and seasonal components.

The non-seasonal portion of the model is represented by the parameters (p, d, q). On the other hand, the seasonal portion of the model is represented by the parameters (P, D, Q).

By specifying the appropriate values for these parameters, the SARIMA model captures both the non-seasonal patterns and the seasonal patterns in the time series, allowing for accurate and robust forecasting of rainfall at the specific points along the railroad track.

$$SARIMA = f[(p, d, q)(P, D, Q)_s] \tag{1}$$

For the SARIMA model tunning, the grid search was applied for the following intervals: p (1, 2, 3), d (1, 2), q (1, 2), P (1, 2, 3), D (1, 2), and Q (1, 2). The parameter s was fixed in 12, which corresponds to monthly seasonality.

4.2.4. Forecast Evaluation

The Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2) are commonly used metrics to evaluate the performance of rainfall prediction models.

The RMSE measures the average deviation between the observed values of precipitation and the predicted values. It is calculated using the following equation:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

where n is the number of samples, y_i represents the actual precipitation value, and $\hat{y}_i$ represents the predicted precipitation value.

The Coefficient of Determination, often denoted as R^2, measures the proportion of the variance in the observed data that is captured by the predicted values. It ranges from 0 to 1, where 1 indicates a perfect fit between the predicted and observed values. It is calculated using the following equation:

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

where y_i represents the actual precipitation value, $\hat{y}_i$ represents the predicted precipitation value, and $\bar{y}$ represents the mean of the observed precipitation values.

These metrics provide quantitative measures of the accuracy and goodness-of-fit of the rainfall prediction models, allowing for the comparison and selection of the best performing model among SARIMA, RNN, and XGBoost.

To evaluate the performance of the machine learning algorithms (SARIMA, RNN, and XGBoost) for precipitation estimation at each of the 13 points, the equations for RMSE and R^2 will be applied individually for each model and each point.

4.2.5. Summary methodology

This paper focuses on the utilization of three machine learning algorithms to generate forecasts based solely on precipitation as the input variable. Each algorithm was selected based on its specific characteristics and the existing literature on these methods. The XGBoost algorithm holds significant appeal within the field of meteorology, as it incorporates memory to retain past information in its neurons. The RNN algorithm, which is chosen for its ability to capture sequential data patterns, aligns with the premise of storing previous information. Lastly, SARIMA, a machine learning algorithm that considers seasonality, serves as a valuable point of comparison. By employing these diverse algorithms, the study aims to explore and evaluate their effectiveness in predicting precipitation.

It is important to note that each point was evaluated individually, meaning that each method was generated per point. Furthermore, for each of these points, the complete time series was utilized, spanning from 1979 to 2010 for training and from 2011 to 2020 for testing purposes. This approach ensures comprehensive analysis and accurate evaluation of the performance of the machine learning algorithms for each specific location.

5. Analysis of the rainfall forecasting

Based on the evaluation using RMSE and R^2 values, the efficacy of the SARIMA, RNN, and XGBoost models for monthly precipitation forecasting at the 13 locations of interest during the 2010-2020 period can be assessed.

The RMSE values provide an indication of the average error between the predicted and actual precipitation values. Lower RMSE values indicate better accuracy, as they represent smaller deviations from the observed data.

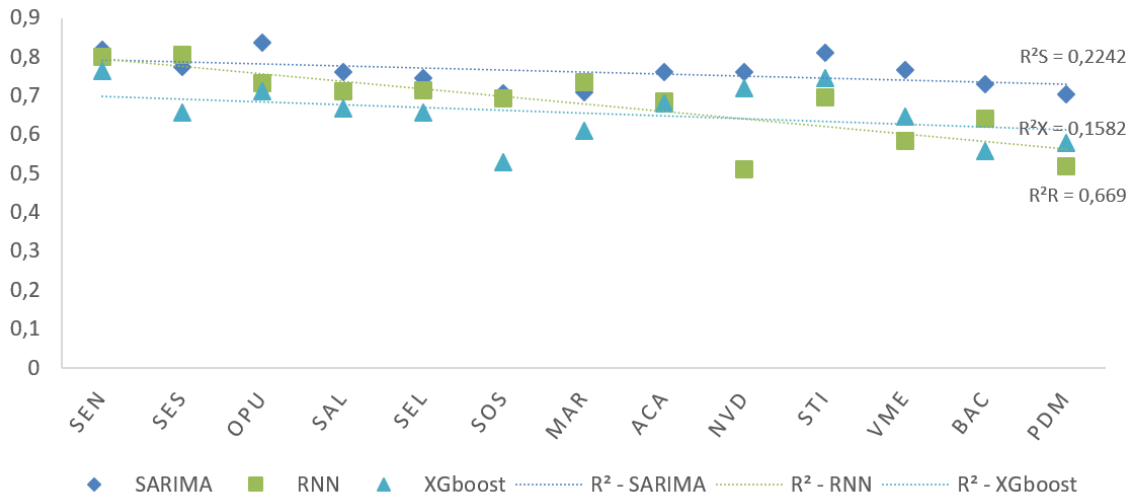

Figure 4. Values of R^2 metric reached by SARIMA, RNN, and XGBoost for all 13 points.

The R^2 values, on the other hand, assess the goodness of fit of the predicted precipitation values compared to the actual values. Higher R^2 values indicate a better fit, suggesting that the model captures a larger proportion of the variance in the data.

According to the results presented in Fig. 4, it is observed that the SARIMA model generally outperformed the RNN and XGBoost models in terms of R^2 values for most of the points of interest. SARIMA achieved the highest R^2 values for 11 out of the 13 points, ranging from 0.704 to 0.817. This indicates that SARIMA provides a better fit to the observed data and captures a larger proportion of the variance in the precipitation patterns at those locations.

On the other hand, for the Serra Sul (SES) and Marabá (MAR) regions, the RNN model achieved the best performance with R^2 values of 0.805 and 0.735, respectively. This suggests that the RNN model may be more suitable for capturing the temporal dynamics and dependencies in precipitation patterns in these specific regions.

Although the XGBoost algorithm did not achieve the best performance for any of the points, it still exhibited relatively close R^2 values to the best-performing models. However, the relatively lower R^2 value of 0.2242 suggests that the XGBoost model may not capture the underlying patterns of precipitation as effectively as SARIMA and RNN.

Overall, the SARIMA model showed the best performance across most of the points, indicating its efficacy for monthly precipitation forecasting in the study area. However, it is worth noting that the performance of the models may vary depending on the specific characteristics of each location and the temporal dynamics of precipitation in those regions.

The linear coefficients shown in Fig. 4 provide additional insights into the variability of the R^2 values for the different models. A linear coefficient of 0 indicates that the R^2 values are evenly distributed around the mean, while coefficients closer to 1 indicate

less variability.

According to the linear coefficients provided, SARIMA achieved a coefficient of 0.224, indicating a relatively higher variability in the R^2 values across the 13 points. This suggests that the performance of SARIMA varied more significantly among different locations, with some points achieving higher R^2 values than others.

On the other hand, the RNN model exhibited a higher linear coefficient of 0.669, indicating a lower variability in the R^2 values. This suggests that the RNN model consistently performed well across the majority of the points, with less variation in its performance.

Interestingly, the XGBoost model achieved the lowest linear coefficient of 0.158, indicating the lowest variability in the R^2 values among the 13 points. This suggests that the performance of XGBoost was relatively consistent across different locations, with less variation compared to the SARIMA and RNN models.

Overall, these results indicate that while XGBoost may not have achieved the highest R^2 values, it exhibited a more consistent performance across the 13 points, with less variability in its predictions. This suggests that XGBoost may provide stable and reliable rainfall predictions, albeit with a lower overall fit to the observed data compared to SARIMA and RNN.

Table. 1 provides an overview of the RMSE values achieved by each model for the monthly precipitation prediction at each of the 13 points of interest. In this case, a lower RMSE value indicates better model performance, as it represents a smaller difference between the predicted values and the observed values.

By examining the table, it can be observed that the SARIMA model achieved the lowest RMSE values for 11 out of the 13 points. This indicates that SARIMA performed well in terms of accurately predicting the precipitation at these locations, with relatively small errors.

The RNN model achieved the lowest RMSE values for 2 points, indicating its better performance at these specific locations. It should be noted that although RNN did not achieve the lowest RMSE values for the majority of the points, its performance was still competitive.

On the other hand, XGBoost did not achieve the lowest RMSE values for any of the points. However, as mentioned earlier, it exhibited a more consistent performance across the 13 points based on the linear coefficient of R^2.

Overall, the results in Table. 1 complement the findings based on R^2 values, indicating that the SARIMA model generally performed well in terms of minimizing the RMSE and accurately predicting monthly precipitation at most of the points. However, it is important to consider both the RMSE and R^2 values together to obtain a comprehensive understanding of the model performance.

The results from both the RMSE values and R^2 values provide a comprehensive assessment of model performance, highlighting the strengths and weaknesses of each method across different points of interest.

The results in Table 1 indicate that the XGBoost model did not outperform the

Table 1. RMSE results for each machine learning algorithm tested

Point	SARIMA	RNN	XGBoost
Serra Norte (SEN)	**49,466**	51,930	56,440
Serra Sul (SES)	54,643	**50,656**	67,382
Onça Puma (OPU)	**43,336**	55,330	57,515
Salobo (SAL)	**54,069**	59,476	63,922
Serra Leste (SEL)	**58,910**	62,415	68,311
Sossego (SOS)	**52,926**	54,130	67,205
Marabá (MAR)	70,547	**67,278**	81,628
Açailandia (ACA)	**52,480**	60,258	60,782
Nova Vida (NVD)	**51,295**	73,234	55,473
Santa Inês (STI)	**57,350**	72,751	66,502
Vitória do Mearim (VME)	**71,369**	95,281	87,822
Bacabeira (BAC)	**91,234**	105,119	116,776
Ponta da Madeira (PDM)	**88,663**	113,204	105,868

other models, including SARIMA, at any of the 13 evaluated points. While XGBoost may have shown better performance compared to RNN at certain points, it did not surpass the performance of SARIMA in terms of RMSE.

This suggests that for the specific task of monthly precipitation prediction at the given points of interest, the SARIMA model demonstrated better overall performance in terms of minimizing the RMSE error. It is important to note that the performance of different models can vary depending on the specific dataset, variables, and characteristics of the problem being addressed. In this particular study, SARIMA proved to be the most effective model for the given task.

6. Conclusions and future works

The study aimed to explore the efficacy of SARIMA, RNN, and XGBoost models in predicting monthly rainfall at the 13 selected points of interest along the railroad track. By focusing on lagged values of precipitation, the models were evaluated based on their performance using the RMSE and R^2 metrics for the test dataset.

The study addressed the fundamental concepts related to rainfall formation and the application of regression models in meteorology. It also provided insights into the preprocessing steps, model selection, hyperparameter tuning, and performance evaluation, as depicted in the proposed methodology.

By comparing the results obtained from the models, it was found that SARIMA consistently achieved the best performance in terms of both RMSE and R^2 values for the majority of the points. RNN showed promising results for two specific points, while XGBoost did not surpass the performance of the other models at any of the evaluated points.

The study was able to assess the effectiveness of the different machine learning methods in predicting precipitation at specific locations. The focus on data configuration, along with the evaluation of the models using standardized metrics, contributed to a comprehensive analysis of the results obtained.

The results obtained in the study, as presented in Fig. 4 and Table. 1, highlight the performance of the SARIMA, RNN, and XGBoost models in monthly rainfall prediction at the 13 points of interest along the railway. The SARIMA model demonstrated superior performance by achieving the best results in 11 out of the 13 points (84.61% of cases) during the period from 2011 to 2020. The RNN model exhibited the best performance in 2 points (15.39% of cases) during the same period.

The findings of the study have practical implications, as the SARIMA and RNN models can provide valuable insights for weather prediction along the railway. This can contribute to decision-making processes, especially in cases where extreme precipitation events are anticipated. By utilizing the lagged values of precipitation, these models can assist in mitigating potential risks and optimizing operational strategies in response to precipitation patterns.

Overall, the study demonstrates the potential of the proposed method and the utility of machine learning models in enhancing precipitation prediction along the railway, thereby aiding in important decision-making processes related to weather conditions.

The planned future work presents promising avenues for further improving the prediction of precipitation at the identified points of interest. The exploration of 1 Dimensional Convolutional Neural Network (CNN 1D) for precipitation prediction can bring new insights and potentially enhance the accuracy of the forecasts. CNNs are known for their ability to capture spatial and temporal patterns in data, and applying them to precipitation prediction can leverage their strengths in handling sequential data.

Segmenting the database into rainy and dry periods for each point of interest is another interesting direction for future research. By dividing the data into distinct periods based on precipitation patterns, it allows for the development of more targeted models that can capture the specific dynamics of rainfall formation during different conditions. Incorporating lagged values of precipitation along with other relevant meteorological variables associated with rainfall formation can further enhance the predictive capability of the models.

By considering these future research avenues, the study can advance the understanding of rainfall prediction along the railway and potentially improve the accuracy and reliability of the forecasts. These efforts will contribute to the development of more robust and effective models for precipitation forecasting, enabling better decision-making and planning in response to changing weather conditions.

References

Berrang-Ford, L., Sietsma, A. J., Callaghan, M., Minx, J. C., Scheelbeek, P. F., Haddaway, N. R., Haines, A., and Dangour, A. D. (2021). Systematic mapping of global research on climate and health: a machine learning review. *The Lancet Planetary Health*, 5(8):e514–e525.

Cerveny, R. S., Bessemoulin, P., Burt, C. C., Cooper, M. A., Cunjie, Z., Dewan, A., Finch, J., Holle, R. L., Kalkstein, L., Kruger, A., et al. (2017). Wmo assessment of weather and climate mortality extremes: lightning, tropical cyclones, tornadoes, and hail. *Weather, climate, and society*, 9(3):487–497.

Chhetri, M., Kumar, S., Pratim Roy, P., and Kim, B.-G. (2020). Deep blstm-gru model for monthly rainfall prediction: A case study of simtokha, bhutan. *Remote sensing*, 12(19):3174.

Fang, J., Li, M., and Shi, P. (2015). Mapping flood risk of the world. *World Atlas of Natural Disaster Risk*, pages 69–102.

Fernandes, E., Rocha, R. L., Ferreira, B., Carvalho, E., Siravenha, A. C., Gomes, A. C. S., Carvalho, S., and de Souza, C. R. (2018). An ensemble of convolutional neural networks for unbalanced datasets: A case study with wagon component inspection. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Group, R. S. (2022). Re swiss group. `http://bit.ly/SwissReFlood`. Accessed: 2023-04-18.

McGuffie, K. and Henderson-Sellers, A. (2001). Forty years of numerical climate modelling. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21(9):1067–1109.

Moazzam, M. F. U., Rahman, G., Munawar, S., Tariq, A., Safdar, Q., and Lee, B.-G. (2022). Trends of rainfall variability and drought monitoring using standardized precipitation index in a scarcely gauged basin of northern pakistan. *Water*, 14(7):1132.

Monego, V. S., Anochi, J. A., and de Campos Velho, H. F. (2022). South america seasonal precipitation prediction by gradient-boosting machine-learning approach. *Atmosphere*, 13(2):243.

NWS (2023). Nws preliminary us flood fatality statistics. `https://www.weather.gov/arx/usflood`. Accessed: 2023-04-18.

Qerimi, Q. and Sergi, B. S. (2022). The case for global regulation of carbon capture and storage and artificial intelligence for climate change. *International Journal of Greenhouse Gas Control*, 120:103757.

Rocha, R. L., Silva, C. D., Gomes, A. C. S., Ferreira, B. V., Carvalho, E. C., Siravenha, A. C. Q., and Carvalho, S. R. (2019). Image inspection of railcar structural components: An approach through deep learning and discrete fourier transform. In *Anais do VII Symposium on Knowledge Discovery, Mining and Learning*, pages 33–40. SBC.

Santos, R. S. and Qin, L. (2019). Risk capital and emerging technologies: innovation and investment patterns based on artificial intelligence patent data analysis. *Journal of Risk and Financial Management*, 12(4):189.

Su, B., Xiao, C., Zhao, H., Huang, Y., Dou, T., Wang, X., and Chen, D. (2022). Estimated changes in different forms of precipitation (snow, sleet, and rain) across china: 1961–2016. *Atmospheric Research*, 270:106078.

Yin, K., Cai, F., and Huang, C. (2022). How does artificial intelligence development affect green technology innovation in china? evidence from dynamic panel data analysis. *Environmental Science and Pollution Research*, pages 1–25.

Zhou, Z., Ren, J., He, X., and Liu, S. (2021). A comparative study of extensive machine learning models for predicting long-term monthly rainfall with an ensemble of climatic and meteorological predictors. *Hydrological Processes*, 35(11):e14424.