Chaos
An Interdisciplinary Journal of Nonlinear Science

RESEARCH ARTICLE | JULY 07 2025

Rainfall forecast in Brazil using machine learning

Special Collection: Advances in Mathematics and Physics: from Complexity to Machine Learning

Sidney T. da Silva ; Letícia C. Milani; Enrique C. Gabrick ; Kelly C. Iarosz ; Ricardo L. Viana ; Iberê L. Caldas ; Antonio M. Batista

Chaos 35, 073116 (2025)
https://doi.org/10.1063/5.0259222

Rainfall forecast in Brazil using machine learning

Cite as: Chaos 35, 073116 (2025); doi: 10.1063/5.0259222
Submitted: 18 January 2025 · Accepted: 20 June 2025 ·
Published Online: 7 July 2025

Sidney T. da Silva,[1] Letícia C. Milani,[1] Enrique C. Gabrick,[2] Kelly C. Iarosz,[2,3,a)] Ricardo L. Viana,[4] Iberê L. Caldas,[5] and Antonio M. Batista[2,5,6]

AFFILIATIONS

[1]Department of Chemical, Federal University of Paraná, Curitiba 81531-980, PR, Brazil
[2]Graduate Program in Science, State University of Ponta Grossa, Ponta Grossa 84030-900, PR, Brazil
[3]University Center UNIFATEB, Telêmaco Borba 84266-010, PR, Brazil
[4]Department of Physics, Federal University of Paraná, Curitiba 81531-980, PR, Brazil
[5]Institute of Physics, University of São Paulo, São Paulo 05508-090, SP, Brazil
[6]Department of Mathematics and Statistics, State University of Ponta Grossa, Ponta Grossa 84030-900, PR, Brazil

Note: This paper is part of the Special Topic on Advances in Mathematics and Physics: from Complexity to Machine Learning.
[a)]**Author to whom correspondence should be addressed:** kiarosz@gmail.com

ABSTRACT

Rainfall forecasting through machine learning can play a crucial role in several areas, such as agriculture, energy, infrastructure, and public safety. The machine learning models have the ability to anticipate climate patterns and extreme events, allowing plantation planning, water resource management, and forecasting energy demands, as well as adopting preventive measures against natural disasters. In this work, we explore three machine learning models (random forest, long short-term memory, and bidirectional long short-term memory) to predict the amount of precipitation in five Brazilian regions (South, Southeast, Central-West, Northeast, and North). We use three-variable reanalysis climate data: local temperature, Atlantic Ocean temperature, and total precipitation. The models are trained by means of the local and Atlantic Ocean temperatures as input features and the total precipitation as a label. Our results indicate that all models perform satisfactorily in their predictions. We verify that the random forest exhibits average absolute errors less than the errors related to the recurrent neural network models. Our results show the effectiveness of machine learning models in predicting rainfall patterns.

Rainfall is an atmospheric phenomenon that impacts not only the local weather but also the global atmospheric circulation. The analysis of precipitation plays a crucial role in the hydrologic cycle, ecosystem processes, and climate patterns. Various approaches and algorithms of machine learning have been used by researchers for rainfall forecasting. In this work, we analyze the precipitation in Brazilian regions by means of machine learning models. For the rainfall forecasting, we use random forest, long short-term memory, and bidirectional long short-term memory. To do that, we consider the Atlantic Ocean and local temperatures, as well as previous precipitation data. Our findings demonstrate a good accuracy for rainfall prediction in Brazil using some techniques related to machine learning.

I. INTRODUCTION

In Brazil, the rainfalls exert a significant influence on the country's economy, especially due to its close relationship with key sectors, such as agriculture and energy production. Brazil is one of the largest agricultural producers in the world, hence a reliable and accurate rainfall forecast is extremely important.[1] These forecasts play a crucial role in the production of crops, such as soybeans, corn, coffee, and sugar cane. Adequate rainfall during the planting and growing seasons is essential to ensuring healthy and abundant harvests. However, periods of prolonged drought[2] or excessive rainfall can result in crop failures, negatively impacting farmers and the entire supply chain. Therefore, a highly accurate forecast is essential for farmers to plan their activities appropriately.

The rainfall is one of the most complex meteorological phenomena[3] due to its non-linear nature and the variety of factors involved in its occurrence. Even in seemingly similar weather conditions, the chance of rainfall can vary considerably from one moment to the next. In the past, weather forecasting was often associated with the image of meteorologists interpreting meteorological maps based on their experience and theoretical knowledge accumulated over the years.[4,5] In recent years, machine learning[6] has become a powerful tool when it comes to rain forecasting. Through sophisticated algorithms and analysis of historical data, it is possible to develop models capable of anticipating weather patterns and providing more accurate predictions about when and where rainfall can occur. The historical rainfall data is essential for training machine learning models, allowing them to learn from past patterns and make future predictions. The algorithms used in rainfall forecasting can range from simple models, such as linear regression,[7] to more complex models, such as artificial neural networks[8] and decision tree algorithms.[9]

One of the challenges in forecasting rainfall is dealing with the complex and dynamic nature of weather,[10] which can be influenced by a variety of factors,[11] as well as long-term climate change.[12] Various studies have demonstrated the efficiency of algorithms in predicting rainfall. Markuna *et al.*[13] studied the application of four machine learning techniques for long-term rainfall prediction. Machine learning models were used to predict Indian summer monsoon rainfall.[14] Sahai *et al.*[15] employed the error backpropagation algorithm to predict summer monsoon rainfall in India using monthly and seasonal time series. They based their predictions on data from the previous 5 years of average monthly and seasonal precipitation values. Philip and Josheph[16] adopted the neural network (adaptive basis function) to predict the annual rainfall in the Kerala region. Poornima and Pushpalatha,[17] using a long short-term memory (LSTM) based intensified recurrent neural network with weighted linear units, demonstrated the importance of accurate rainfall forecasting in meteorology and showed a new approach using deep learning techniques. Somvanshi *et al.*[18] carried out rainfall forecasts in the Hyderabad region, India, utilizing an artificial neural network (ANN) model. They conducted a comparison between ANN and the autoregressive integrated moving average (ARIMA) technique. To feed the neural network model, they used precipitation data from the last 4 months. Wu *et al.*[19] carried out rainfall forecasts in India and China using the modular artificial neural network (MANN). They compared the performance of MANN with logistic regression (L), k-nearest neighbor (KNN), and ANN methods. Aswin *et al.*[20] showed that the LSTM and ConvNet architectures contribute to rainfall forecasting, effectively capturing temporal and spatial patterns in the data, respectively. Caseri *et al.*[21] addressed heavy rainfall forecasting by means of weather radar data and convolutional recurrent neural networks (CNN-LSTM). Considering LSTM, de Araújo *et al.*[22] proposed an approach to predict extreme precipitation events in the southeast region of Brazil. Using ANNs, Esteves *et al.*[23] introduced a softcomputing technique to predict the occurrence of rainfall over short periods of time.

In this work, we focus on the rainfall forecasting in the five Brazilian regions (South, Southeast, Centalr-West, Northeast, and North). We consider the Atlantic Ocean temperature (divided into Atlantic and South Atlantic), local temperature, and previous precipitation data. Previous works reported the strong relationship between Atlantic meteorological conditions and the region's precipitation regime.[24–27] For the predictions, we use three models: random forest, LSTM, and bidirectional LSTM. We demonstrate that these machine learning models are able to predict precipitation in Brazilian regions. They perform satisfactorily in their predictions when trained with data on climate variables. Moreover, in our simulations, we do not find significant improvement in forecasting across the three techniques related to machine learning. We show that the random forest algorithm exhibits the smallest average absolute error value compared with LSTM and bidirectional LSTM.

In our study, the main novelty is the rainfall forecast in five Brazilian regions considering not only the local temperatures but also the Atlantic Ocean temperatures. In addition, we do not observe a significant improvement in the forecasting when previous precipitation data are added in the training of the machine learning algorithms.

This paper is structured as follows: in Sec. II, we describe the machine learning methods as data acquisition and processing. Section III is devoted to present and discuss our results. Finally, our conclusions are drawn in Sec. IV.

II. METHODS

We collect the data from the Copernicus program (www.copernicus.eu/en). Copernicus is the Earth observation component of the European Union's space program, which records and analyzes data about our planet and its environment from 1940 to 2023. The data are the monthly averages for each region, measured by the average of their values between the latitudes and longitudes for each region, as shown in Table I. The average is taken over the range of the coordinates with a precision of 0.25. We utilize the temperature measured 2 m from the surface, ocean temperature measured 2 m from the surface, and total precipitation in meters.

When there are many outliers in the data set, a standardization technique can reduce the error in the results. In this work, we use the Robust Scaler method, which works by subtracting the median [med(X)] from the data (X) and scaling in the interval between the first (Q_1) and the third (Q_3) quartiles. The Robust Scaler equation is given by

$$RS(x_i) = \frac{x_i - \text{med}(x)}{Q_3 - Q_1}. \quad (1)$$

TABLE I. Latitude and longitude of each region where the reanalysis data were obtained.

Regions	Longitude (West, East)	Latitude (North, South)
South	(−57.6, −48.35)	(−22.72, −33.72)
Southeast	(−52.97, −39.97)	(−14.54, −25.29)
Midwest	(−61.47, −46.22)	(−7.91, −23.66)
Northeast	(−48.51, −35.01)	(−1.3, −18.3)
North	(−73.73, −46.23)	(5.21, −13.56)
Atlantic	(−49.2, −17.2)	(12.25, −34, 12)
South Atlantic	(−65.73, −56.19)	(−34.2, −56.19)

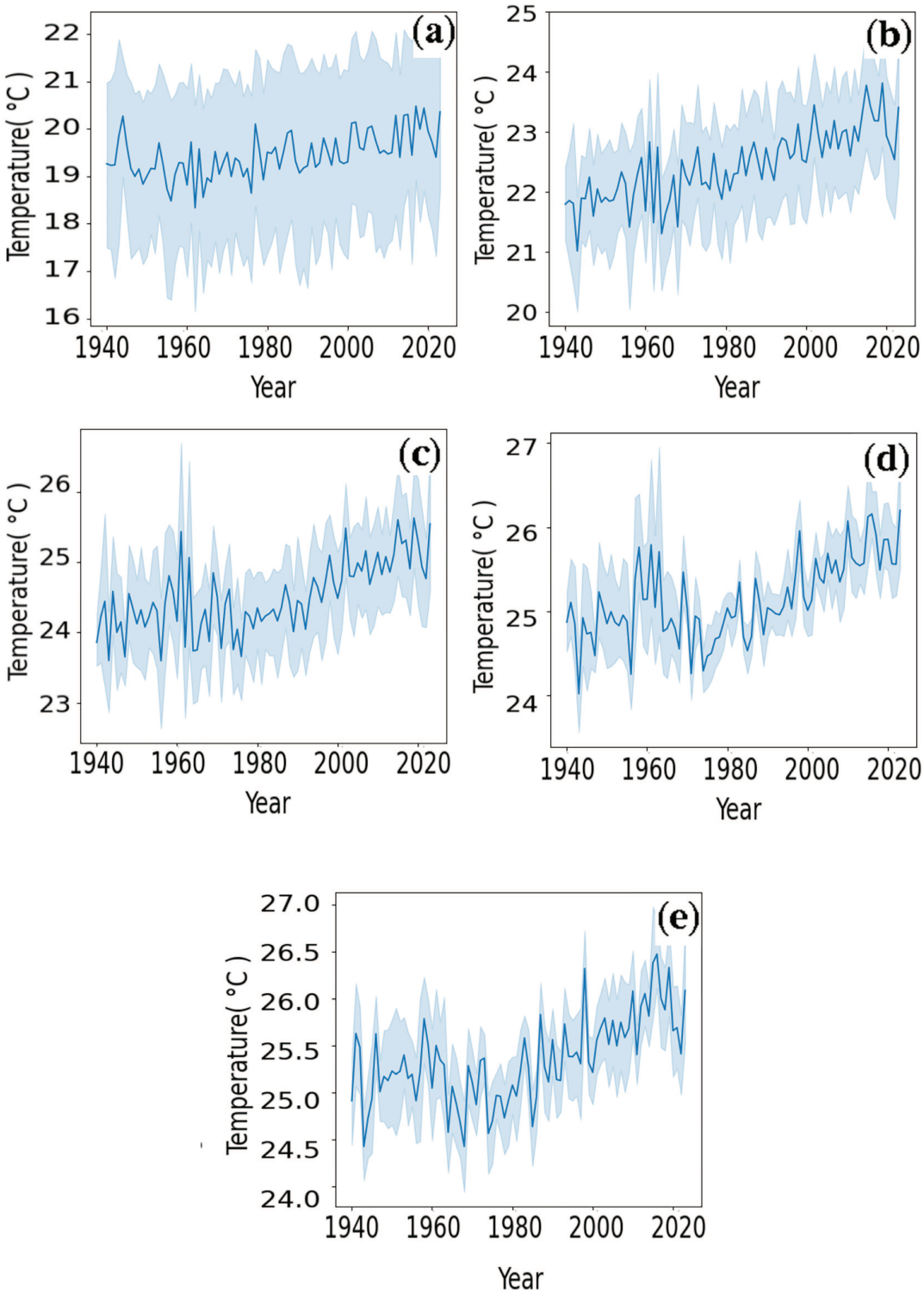

FIG. 1. Temperature as a function of years for (a) South, (b) Southeast, (c) Midwest, (d) North, and (e) Northeast. The curve represents the annual averages and the bars correspond to the monthly temperatures for each year. The temperature scale is measured in degrees Celsius (°C).

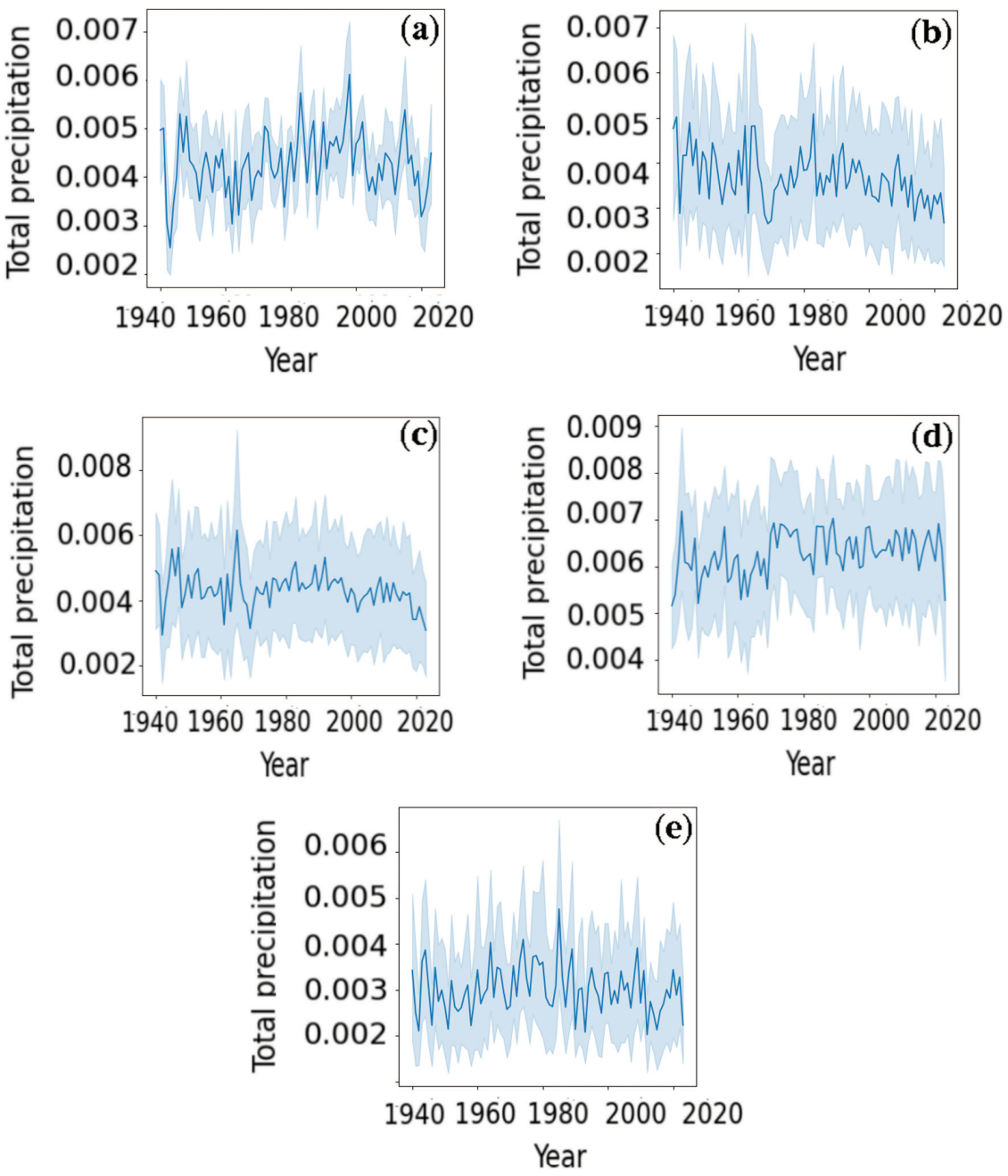

FIG. 2. Total precipitation as a function of years for (a) South, (b) Southeast, (c) Midwest, (d) North, and (e) Northeast. The curve represents the annual averages and the bars correspond to the monthly total precipitation for each year. The total precipitation scale is measured in meters (m).

The median and interquartile range ($Q_3 - Q_1$) are stored and used in future data as the transformation applied in the forecast.

In our simulations, we chose to use the Random Forest (RF)[9] algorithm. RF is a supervised learning algorithm based on the concept of Ensemble Learning, which combines multiple models to improve prediction accuracy. It comprises multiple decision trees and is known for its effectiveness in handling complex data sets and reducing overfitting. RF uses an initial data set (D) to

generate predictions or classifications through multiple decision trees.

Long short-term memory (LSTM) models capture long-term dependencies in sequential data. They stand out for their ability to retain and select information within the model's memory cell, using specialized control mechanisms such as forget, input, and output gates. Due to their complex architecture, which includes gate operations controlled by sigmoid functions and hyperbolic tangent functions,[28] LSTMs are capable of handling and processing sequential information in extended sequences.

Successful applications of LSTM networks cover areas such as human trajectory prediction, traffic prediction, speech recognition, and weather forecasting. These RNN cells have the ability to capture dependencies from at least two previous states as well as the current state. The evanescent gradient problem is reduced by incorporating three gates along with the hidden state. These gates, commonly known as entry, exit, and forget gates, regulate how much information from the new state is relevant. The input port defines how much information from the new state is used. The output port determines the amount of information used from previous states. The forget gate controls the amount of internal state information that is transmitted to the next layer. The LSTM model equations are defined as

$$\begin{aligned} f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \\ i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \\ o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \\ \tilde{c}_t &= \sigma_c(W_c x_t + U_c h_{t-1} + b_c), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ h_t &= o_t \odot \sigma_h(c_t), \end{aligned} \tag{2}$$

where x_t is the input vector to the LSTM unit, h_t is the hidden state vector (output vector of the LSTM unit), f_t is the forget gate's activation vector, σ_g is the sigmoid function, σ_c is the hyperbolic tangent function, σ_h is the hyperbolic tangent function or $\sigma_h(x) = x$, i_t is the input/update gate's activation vector, o_t is the output gate's activation vector, $\tilde{c}$ corresponds to the cell input activation vector, and h_t is the hidden state vector.

During the training, the weight matrices (W and U) and bias vectors (b) are the learnable parameters. The operator $\odot$ denotes the element-wise multiplication (Hadamard product). Stacked LSTM networks are composed of two or more LSTM networks successively connected as hidden layers. This stacked architecture can provide, in some applications, a higher level of representation of time series data than individual LSTM networks.

We also compute bidirectional long short-term memory (LSTM), that is, a variation of RNNs capable of learning dependencies on previous and future states. This type of architecture has shown good results in domains such as natural language processing for speech and handwriting recognition. Bidirectional LSTM involves LSTM cells that capture left-to-right time series data (standard LSTM cells) and LSTM cells that capture data in reverse.[28,29] The architecture of a bidirectional network-LSTM has two hidden layers. Replacing RNN cells with LSTM cells in bidirectional RNN results in bidirectional-LSTM networks.

III. RAINFALL FORECAST IN BRAZIL

In this work, our objective is to demonstrate that, despite rainfall being a complex time series, it is possible to make predictions with high accuracy using machine learning models, such as Random Forest and LSTM. In order to simplify and make the model more robust, we chose to use as few variables as possible as input characteristics in our models, focusing mainly on the climate variables of local temperature (Fig. 1), Atlantic Ocean temperature (Fig. 2) and delayed data on total precipitation (Fig. 3). In our simulations, the length of the training set is equal to 806 and the length of the test set is equal to 202, where the attributes are related to the climate features.

In Fig. 1, the temperature reanalysis data are exhibited over the years, covering the interval from 1940 to 2023. The graphs show the temperatures measured in the five regions of Brazil (South, Southeast, Midwest, North, and Northeast). Figure 2 displays the total precipitation in each region, also using the reanalysis data from 1940 to 2023. Figure 3 shows the temperature data for the Atlantic Ocean and South Atlantic in blue and orange, respectively.

Figure 4 displays the correlation matrices between the climate variables in each region. We analyze the correlations between the climate variables (Atlantic and South Atlantic) related to the temperature and total precipitation. In Fig. 4(a), which corresponds to the South region, it is observed that the correlations are less than 0.5, indicating a low correlation between the temperature and precipitation. Analyzing the Southeast region [Fig. 4(b)], we verify a more significant correlation, with emphasis on the temperature which exhibits a stronger correlation, above 0.5. This pattern is repeated in the Central-West region [Fig. 4(c)]; however, it is the temperature of the Atlantic Ocean that shows the greatest correlation. It is important to highlight that, despite these stronger correlations, we see a relatively low correlation pattern in these three regions.

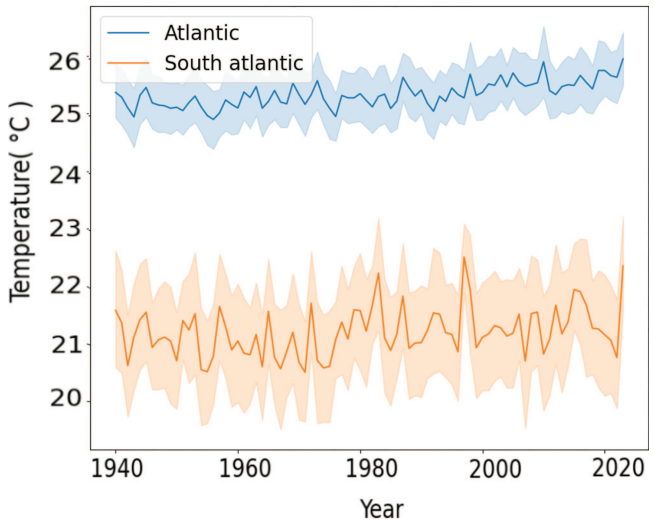

FIG. 3. Temperature as a function of years, the curve represents the annual average, and the bars are the monthly averages. The temperatures are measured in the Atlantic (blue) and South Atlantic (orange) Oceans.

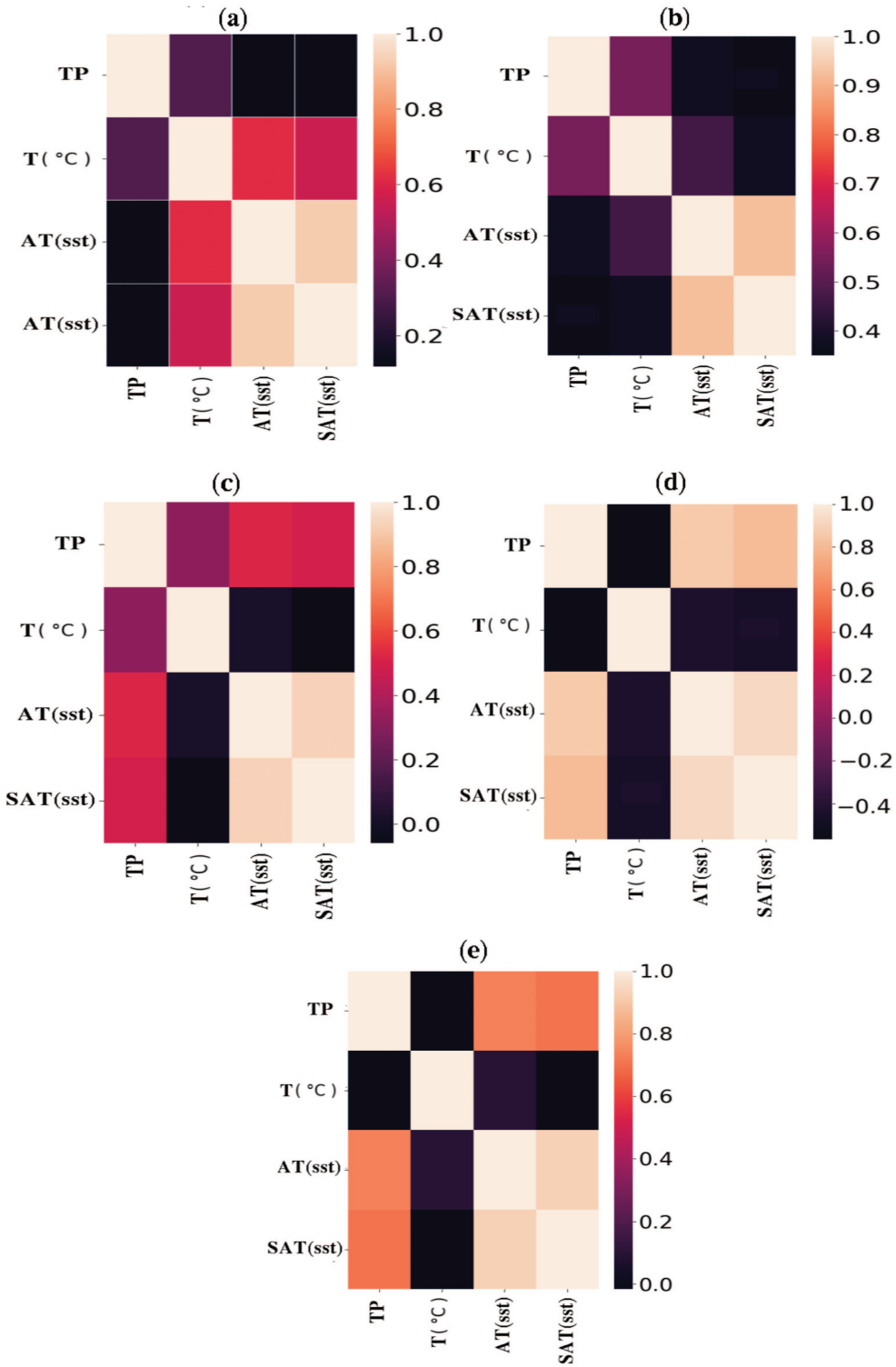

FIG. 4. Correlation matrix for (a) South, (b) Southeast, (c) Central-West, (d) North, and (e) Northeast.

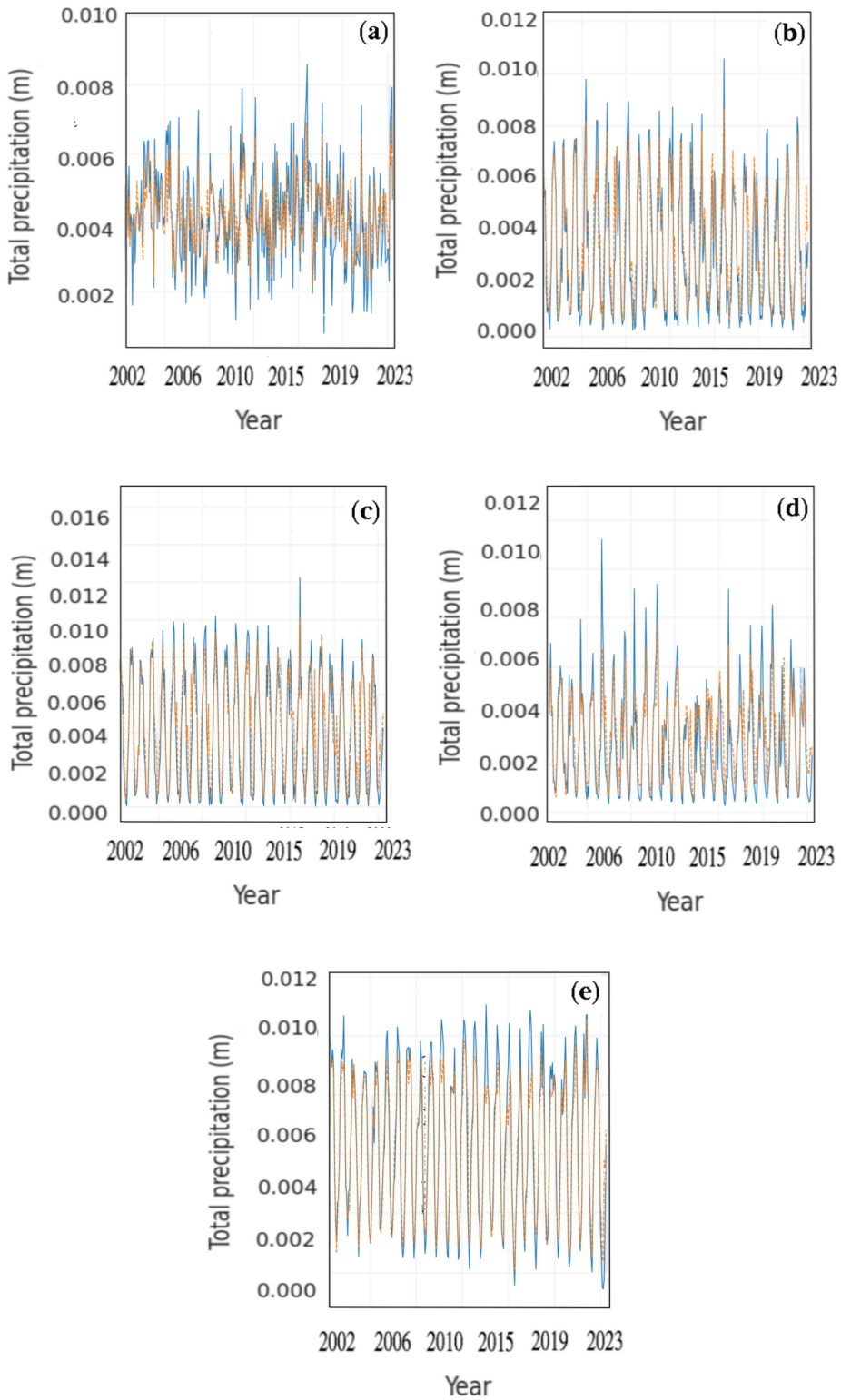

FIG. 5. Rainfall prediction for the five regions of Brazil using the local and Atlantic Ocean (North and South) temperatures. The blue and orange lines are the actual and predicted values for (a) South, (b) Southeast, (c) Midwest, (d) Northeast, and (e) North.

TABLE II. The mean absolute error (MAE) for each region considering without (second column) and with (third column) rainfall. It includes local temperature and ocean temperature, both lagged by 2 months.

Regions	MAE (no rainfall)	MAE (with rainfall)
South	6.7×10^{-4}	6.5×10^{-4}
Southeast	7.5×10^{-4}	7.5×10^{-4}
Midwest	9.2×10^{-4}	6.0×10^{-4}
Northeast	8.4×10^{-4}	5.5×10^{-4}
North	7.2×10^{-4}	6.0×10^{-4}

Analyzing the northernmost regions of Brazil, a change in the correlation matrix is notable. In this area, practically, all variables exhibit correlations above 0.5, with some reaching higher values, such as 0.7 and 0.8. This significant increase in the correlations suggests a more robust relationship between climate variables and total precipitation in the northern regions of Brazil.

In the northern region [Fig. 4(d)], where most of the Amazon forest is located, we see that the ocean temperature has a high correlation in relation to the rainfall profile, with emphasis on the northernmost Atlantic region, which has a correlation equal to 0.86. An interesting behavior is the negative correlation in relation to the temperature. In the northeast region [Fig. 4(e)], we also observe a large correlation in relation to precipitation and ocean temperature profiles, with emphasis on the northernmost Atlantic Ocean region and a correlation equal to 0.73. The South Atlantic Ocean region has a correlation with a value equal to 0.70. A temperature with a value of −0.02 shows a zero correlation with the precipitation.

Random forest is a supervised learning method that uses Ensemble Learning, combining multiple models to improve prediction accuracy. It comprises multiple decision trees and is recognized for its effectiveness in handling complex data sets and reducing overfitting. For the training process, we first separate the samples into 70% and 30% for the training and test, respectively. To predict the amount of precipitation $(y_i + \alpha)$, we consider as input $SSTN_i$, $SSTS_i$, and T_i,

$$(y_{i+\alpha}) = f(SSTN_i, SSTS_i, T_i, y_i). \tag{3}$$

The training objective is to find the best function f that correlates the inputs $(SSTN_i, SSTS_i, T_i, y_i)$ with the output $y_i + \alpha$.

Two different learning processes are carried out. In the first process, we use the local temperature and the Atlantic Ocean temperature (North and South) as input. For the training phase, we utilize the data with a size of 750 times, that is, all input data (local and Atlantic Ocean temperatures) from 1940 to 2002, as shown in Fig. 5. For the test, we consider data 258 in length (2002–2023). In Table II, we see that the model is able to predict precipitation with an average absolute error with an order of $\sim 10^{-4}$. The best delay with the best result is equal to 2 months ($\alpha = 2$).

In Fig. 6, we consider the temperatures and previous precipitation data (2 month delay). It is not observed a significant improvement in the results, as shown in Table II by means of the mean absolute error. In the time series, we see that the southern region is the noisiest; however, it is the one with the lowest prediction error.

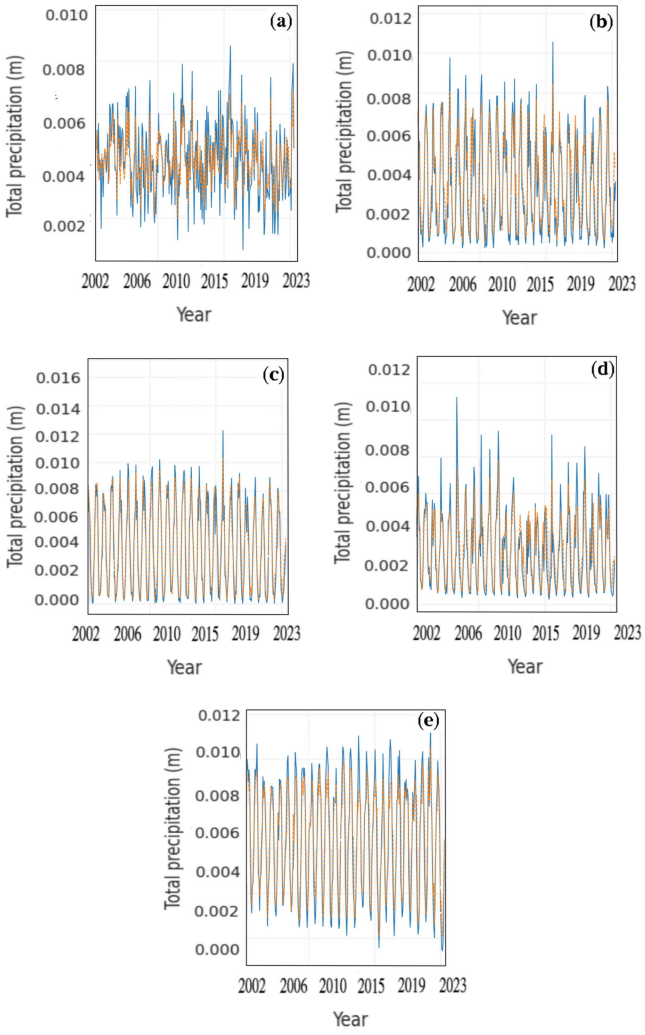

FIG. 6. Rainfall prediction for the five regions of Brazil considering the local temperature, the temperature of the Atlantic Ocean (North and South), and previous precipitation data. The blue lines are the actual values and the orange dashed lines are the predicted values for (a) South, (b) Southeast, (c) Midwest, (d) Northeast, and (e) North.

The long short-term memory (LSTM) neural network, introduced by Hochreiter and Schmidhuber,[30] was designed to address the evanescent gradient problem, that is, a common challenge in recurrent neural networks. During the training phase, the network weights are updated iteratively using an error gradient to adjust the network output. However, the calculation of this gradient can become insignificant due to the long temporal range of temporal sequences.[31]

The LSTM neural network is trained with time series of climate variables related to total precipitation. The sequences of climate features are separated into 80% and 20% of training and test sets, respectively. We used other partitions to split between training

and testing data sets, for instance, 70% and 30%. The best results are obtained using an 80/20 split. The input data are normalized through RobustScaler before entering the LSTM network. The activation function is a specific non-linear sigmoid (S-shaped) function, the logistic function, which allows nodes to learn complex structures.[32] The loss function is given by the Mean Square Error (MSE) metric with the Adam optimizer.[33]

In this work, each sample/instance is composed of a sequence of five consecutive values, each with the values of the climatic characteristics (local temperature, Northernmost Atlantic Ocean temperature, and Southernmost Atlantic Ocean temperature). To find

TABLE III. The mean absolute error (MAE) for each region considering without (second column) and with (third column) rainfall. It includes local temperature and ocean temperature, both lagged by 2 months (LSTM).

Regions	MAE (no rainfall)	MAE (with rainfall)
South	2.0×10^{-3}	2.0×10^{-3}
Southeast	3.0×10^{-3}	2.6×10^{-3}
Midwest	3.2×10^{-3}	3.8×10^{-3}
Northeast	2.3×10^{-3}	3.1×10^{-3}
North	2.1×10^{-3}	2.6×10^{-3}

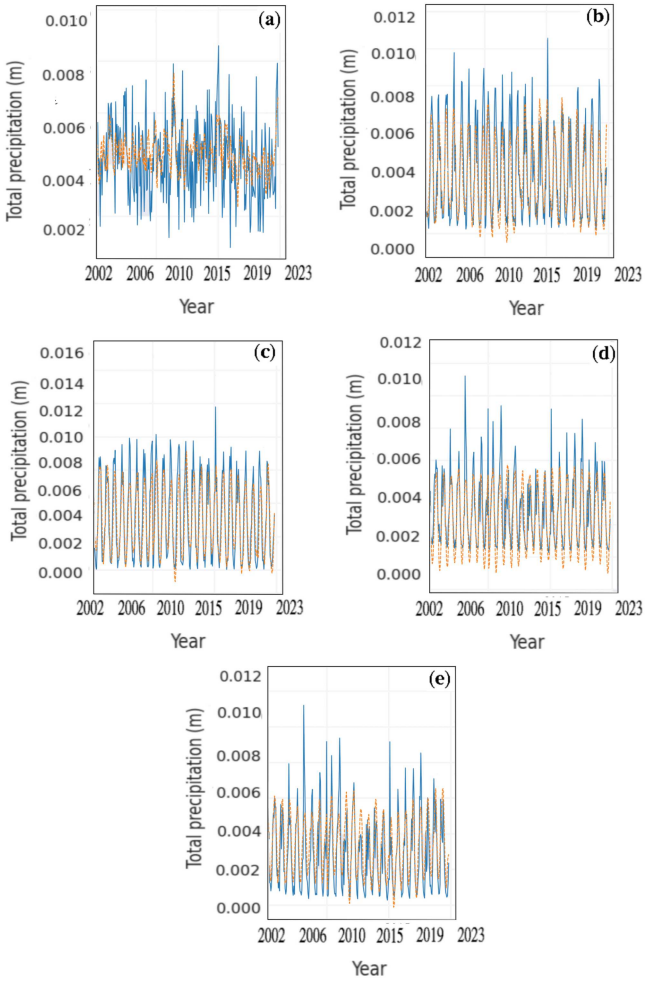

FIG. 7. Rainfall prediction for the five regions of Brazil considering the local temperature and the temperature of the Atlantic Ocean (North and South). The black vertical line separates the training and testing region. The blue lines are the actual values and the orange dashed lines are the predicted values for (a) South, (b) Southeast, (c) Midwest, (d) Northeast, and (e) North.

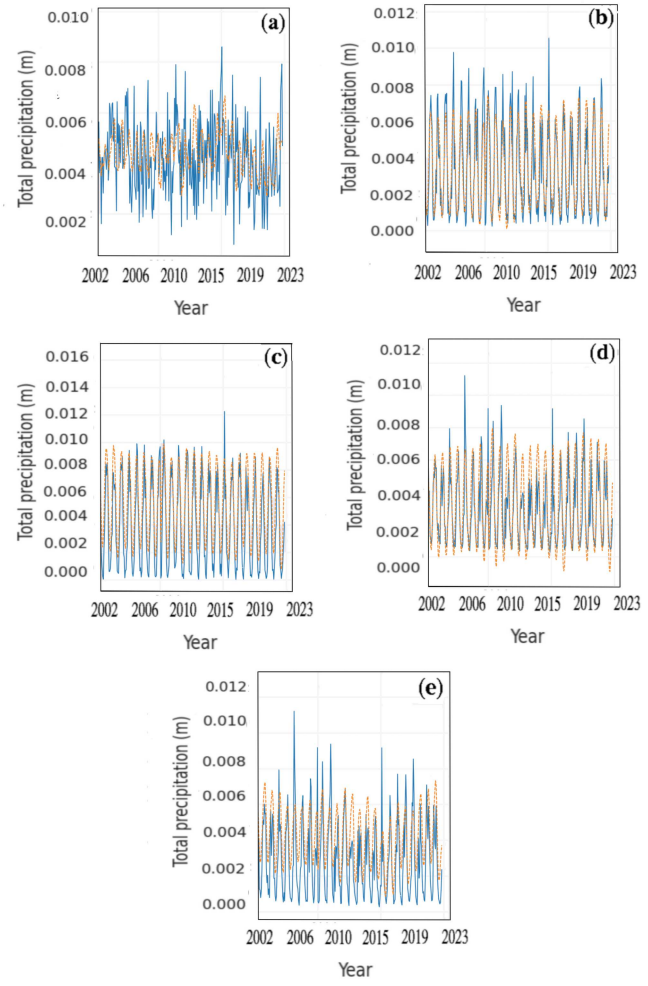

FIG. 8. Rainfall prediction for the five regions of Brazil considering the local temperature and the Atlantic Ocean (North and South) temperatures. The blue and orange lines are the actual and predicted values for (a) South, (b) Southeast, (c) Midwest, (d) Northeast, and (e) North.

TABLE IV. The mean absolute error (MAE) for each region considering without (second column) and with (third column) rainfall. It includes local temperature and ocean temperature, both lagged by 2 months (bidirectional LSTM).

Regions	MAE (no rainfall)	MAE (with rainfall)
South	1.4×10^{-3}	1.4×10^{-3}
Southeast	2.8×10^{-3}	2.8×10^{-3}
Midwest	3.6×10^{-3}	3.6×10^{-3}
Northeast	2.7×10^{-3}	3.1×10^{-3}
North	2.6×10^{-3}	2.6×10^{-3}

an optimal number of epochs, i.e., the number of times in which the training is performed, some tests are carried out for 50, 80, and 100. The prediction error decreased from 10 to 80 epochs. There is no significant change above 80 epochs and hence the number of epochs adopted is 80. At each training epoch, a validation is performed using 10% of the data to evaluate the convergence of the training process and analyze whether overfitting has occurred. A convergence is observed in the training and validation data, indicating that the neural network does not suffer from overfitting.

After the model training phase, we apply the training network to the test data (Fig. 7). In Fig. 7, the blue and orange lines represent the true data and values predicted, respectively. Compared with RF, we verify that these results are not better than RF. Table III exhibits the values of the mean absolute error with and without precipitation. Comparing the error with or without precipitation, we observe that there is no significant change.

In Fig. 8, we compute the bidirectional long short-term memory (LSTM). We use two bidirectional layers, one with 64 units and the other with 28 units, dropout of 0.5 and 0.25, respectively. We also utilize a third dense layer with one unit. The activation functions are ReLU, ReLU, and linear, respectively. In the training process, we consider batches of 64 trained in 100 epochs. We use 10% of the training data for validation and find that the neural network is not overfitted. In Table IV, we observe that there is no significant change in the mean absolute error with or without the use of precipitation data. Figure 8 shows that the southern region has worsened compared to the other previous models. By means of the mean squared error, this worsening is not very clear. The other regions exhibit results similar to the ones obtained by the LSTM model.

Comparing the values of the mean absolute error, we see that RF stands out as the best model. The RF error has values about 10^{-4}, while the neural networks have their errors about 10^{-3}. Figures 6, 7, and 8 show that RF stands out in predicting rainfall. We compute the mean absolute percentage error (MAPE) for the regions using Random Forest (RF), LSTM, and bidirectional LSTM, considering no rainfall, as exhibited in Table V. Our results show that the MAPE values using the RF algorithm are less than 10%, indicating a highly accurate prediction.

Considering RF for Malaysian rainfall prediction, Zainudin *et al.*[34] achieved an accuracy of under 90% and our current model achieves an accuracy of over 90% in most regions. Nguyen-Duc *et al.*[35] used LSTM for seasonal prediction of monthly rainfall across Vietnam and found absolute errors about 10^{-1}. In our findings, we observe absolute errors about 10^{-3}, highlighting the effectiveness of combining oceanic and local temperatures in Brazilian contexts.

TABLE V. Mean absolute percentage error (%) for the regions using Random Forest (RF), LSTM, and bidirectional LSTM, considering no rainfall.

Regions	RF	LSTM	Bidirectional LSTM
South	9.24	30.58	30.83
Southeast	7.47	23.47	22.93
Midwest	6.67	13.19	12.24
Northeast	7.17	18.69	17.92
North	5.28	17.28	17.4

IV. CONCLUSIONS

In this work, we investigate the ability of some machine learning models to predict precipitation with high accuracy. We use local and Atlantic Ocean temperatures as input characteristics, both in the southern and northern regions. We include the temperature of the Atlantic Ocean due to its recognized influence on the precipitation regime in Brazil.

We test three machine learning models, which are Random Forest (RF), LSTM, and bidirectional LSTM. Trained with data on climate variables, these models show excellent results, with emphasis on RF, which achieved the best performance. The RF model has an average absolute error with a value of about 10^{-4}, while the other models about 10^{-3}. We perform two different tests on each model. In one, we use local and ocean temperatures, while in the other, we also considered the delayed total precipitation data. Our results indicate that there is no significant improvement in prediction across all machine learning models.

The total precipitation data from the southern region exhibit a noisier and more complex behavior than other regions. However, even with this complexity, the RF model provides excellent accuracy in its predictions. On the other hand, the bidirectional LSTM performs worse compared to the other models.

Overall, our findings highlight the interplay between local and ocean temperatures in rainfall forecasts in Brazil. By using machine learning algorithms, our work provides a step toward connecting some climate data and meteorological prediction. Capturing complex patterns associated with climate data, it is possible to improve the rainfall forecast accuracy. Due to this fact, disaster impacts related to heavy rains can be reduced.

In future works, we plan to focus on rainfall forecasting in the southern region, where intense rainfall is frequent and has a significant impact. Our approach will involve using imagery to predict the amount of rainfall locally and exploring other model architectures to further improve forecast accuracy.

ACKNOWLEDGMENTS

This work was possible with partial financial support from the following Brazilian government agencies: CNPq, CAPES, Fundação Araucária, and São Paulo Research Foundation (FAPESP) (Nos. 2022/13761-9, 2024/14478-4, and 2024/05700-5). E.C.G. received partial financial support from Coordenação de Aperfeiçoamento

de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 88881.846051/2023-01. R.L.V. received partial financial support from CNPq (Nos. 403120/2021-7 and 301019/2019-3), CAPES (No. 88881.143103/2017-01), and FAPESP (No. 2022/04251-7). We acknowledge 105 Group Science (www.105groupscience.com).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Sidney T. da Silva: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Letícia C. Milani:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Enrique C. Gabrick:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Kelly C. Iarosz:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Ricardo L. Viana:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Iberê L. Caldas:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Antonio M. Batista:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

[1]L. A. F. Monteiro, F. I. C. do Nascimento, J. F. de Oliveira-Júnior, D. D. Nunes, D. Mendes, G. de Gois, F. O. Sanches, C. A. Wollmann, M. Watanabe, and J. P. A. Gobo, "Rainfall projections for the Brazilian legal Amazon: An artificial neural networks first approach," Climate **12**, 187 (2024).
[2]I. C. M. Tinôco, B. G. Bezerra, P. S. Lucio, and L. M. Barbosa, "Characterization of rainfall patterns in the semiarid Brazil," Anuário do Instituto de Geociências **41**, 397–409 (2018).
[3]F. Song, H. Dong, L. Wu, L. R. Leung, J. Lu, L. Dong, P. Wu, and T. Zhou, "Rainfall forecast in Brazil using machine learning," Nat. Commun. **16**, 2188 (2025).
[4]M. V. Young and N. S. Grahame, "The history of UK weather forecasting: The changing role of the central guidance forecaster. Part 1: The pre-computer era," Weather **77**, 344–348 (2022).
[5]N. R. Afshar and H. Fahmi, "Rainfall forecasting using Fourier series," J. Civ. Eng. Archit. **6**, 1258–1262 (2012).
[6]Z.-H. Zhou, *Machine Learning* (Springer, Singapore, 2021).
[7]F. R. Volkmar, *Encyclopedia of Autism Spectrum Disorders* (Springer, New York, 2013).
[8]M. A. Wani, F. A. Bhat, S. Afzal, and A. I. Khan, *Advances in Deep Learning* (Springer, Singapore, 2020).
[9]A. Cutler, D. R. Cutler, and J. R. Stevens, *Random Forests* (Springer, New York, 2012).
[10]P. Antonello, "Climate models," Rendiconti Lincei **25**, 49–58 (2014).
[11]D. P. Chowdhury, D. Roy, and U. Saha, "Incorporatin of weather parameters in MMRC-K model for rainfall disaggregation," Stoch. Environ. Res. Risk Assess. **39**, 289–308 (2025).
[12]B. Halder, B. Rana, L. Juneng, C. B. Pande, S. Alshehery, M. Elsahabi, K. K. Yadav, S. Sh. Sammen, and S. R. Naganna, "Cloud computing-based estimation of Peninsular India's long-term climate change impacts on rainfall, surface temperature, and geospatial indices," Geomat. Nat. Hazards Risk **15**, 2381635 (2024).
[13]S. Markuna, P. Kumar, R. Ali, D. K. Vishwkarma, K. S. Kushwaha, R. Kumar, V. K. Singh, S. Chaudhary, and A. Kuriqi, "Application of innovative machine learning techniques for long-term rainfall prediction," Pure Appl. Geophys. **180**, 335–363 (2023).
[14]U. Narang, K. Juneja, P. Upadhyaya, P. Salunke, T. Chakraborty, S. K. Behera, S. K. Mishra, and A. D. Suresh, "Artificial intelligence predicts normal summer monsoon rainfall for India in 2023," Sci. Rep. **14**, 1495 (2024).
[15]A. K. Sahai, M. K. Soman, and V. Satyan, "All India summer monsoon rainfall prediction using an artificial neural network," Clim. Dyn. **16**, 291–302 (2000).
[16]N. S. Philip and K. B. Joseph, "A neural network tool for analyzing trends in rainfall," Comput. Geosci. **29**, 215–223 (2003).
[17]S. Poornima and M. Pushpalatha, "Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units," Atmosphere **10**, 668 (2019).
[18]V. K. Somvanshi, O. P. Pandey, P. K. Agrawal, N. V. Kalanker, M. R. Prakash, and R. Chand, "Modeling and prediction of rainfall using artificial neural network and ARIMA techniques," J. Indian Geophys. Union **10**, 141–151 (2006).
[19]C. L. Wu, K. W. Chau, and C. Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques," J. Hydrol. **389**, 146–167 (2010).
[20]S. Aswin, P. Geetha, and R. Vinayakumar, "Deep learning models for the prediction of rainfall," in *International Conference on Communication and Signal Processing (ICCSP), Chennai, India* (IEEE, 2018), pp. 657–661.
[21]A. N. Caseri, L. B. L. Santos, and S. Stephany, "A convolutional recurrent neural network for strong convective rainfall nowcasting using weather radar data in Southeastern Brazil," Artif. Intell. Geosci. **3**, 8–13 (2022).
[22]A. de Araújo, A. R. Silva, and L. E. Zárate, "Extreme precipitation prediction based on neural network model—A case study for southeastern Brazil," J. Hydrol. **606**, 127454 (2022).
[23]J. T. Esteves, G. S. Rolim, and A. S. Ferraudo, "Rainfall prediction methodology with binary multilayer perceptron neural networks," Clim. Dyn. **52**, 2319–2331 (2019).

[24]G. A. Hounsou-gbo, M. Araujo, B. Bourlés, D. Veleda, and J. Servain, "Tropical Atlantic contributions to strong rainfall variability along the Northeast Brazilian coast," Adv. Meteorol. **2015**, 902084 (2014).
[25]Y. K. Kouadio, J. Servain, L. A. T. Machado, and C. A. D. Lentini, "Heavy rainfall episodes in the eastern northeast Brazil linked to large-scale ocean-atmosphere conditions in the tropical Atlantic," Adv. Meteorol. **2012**, 369567 (2012).
[26]V. Barros, M. Gonzalez, B. Liebmann B, and I. Camilloni, "Influence of the South Atlantic convergence zone and South Atlantic Sea surface temperature on interannual summer rainfall variability in Southeastern South America," Theor. Appl. Climatol. **67**, 123–133 (2000).
[27]J.-H. Yoon and N. Zeng, "An Atlantic influence on Amazon rainfall," Clim. Dyn. **34**, 249–264 (2010).
[28]M. A. I. Sunny, M. S. Maswood, and A. G. "Alharbi, "Deep learning-based stock price prediction using LSTM and bi-directional LSTM model," in *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)* (IEEE, 2020), pp. 87–92.
[29]U. Singh, S. Chauhan, A. Krishnamachari, and L. Vig, "Ensemble of deep long short term memory networks for labelling origin of replication sequences," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, 2015), pp. 1–7.
[30]S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput. **9**, 1735–1780 (1997).
[31]A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Physica D **404**, 132306 (2020).
[32]P. Chandra, "Sigmoidal function classes for feedforward artificial neural networks," Neural Process. Lett. **18**, 185–195 (2003).
[33]Z. Zijun, "Improved Adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), Banff, AB, Canada* (IEEE, 2018), pp. 1–2.
[34]S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative analysis of data mining techniques for Malaysian rainfall prediction," Int. J. Adv. Sci. Eng. Inf. Technol. **6**, 1148–1153 (2016).
[35]P. Nguyen-Duc, H. D. Nguyen, Q.-H. Ngyuen, T. Phan-Van, and H. Pham-Thanh, "Application of long short-term memory (LSTM) network for seasonal prediction of monthly rainfall across Vietnam," Earth Sci. Inform. **17**, 3925–3944 (2024).