UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA HIDRÁULICA E AMBIENTAL
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS HÍDRICOS

LARISSA ZAIRA RAFAEL ROLIM

ANALYSIS OF THE DYNAMICS AND COMPLEXITY OF RAINFALL AND STREAMFLOW TIME SERIES IN CEARÁ, BRAZIL: INTEGRATING CHAOS THEORY, INFORMATION THEORY, MACHINE LEARNING AND STOCHASTIC APPROACHES

FORTALEZA
2023

LARISSA ZAIRA RAFAEL ROLIM

ANALYSIS OF THE DYNAMICS AND COMPLEXITY OF RAINFALL AND STREAMFLOW TIME SERIES IN CEARÁ, BRAZIL: INTEGRATING CHAOS THEORY, INFORMATION THEORY, MACHINE LEARNING AND STOCHASTIC APPROACHES

Thesis defended at the Water Resources and Environmental Sanitation Post-Graduate Program of the Technology Center at the Federal University of Ceará, as partial requirement to obtain the doctor degree in Civil Engineering. Concentration Area: Water Resources.

Advisor: Prof. Dr. Francisco de Assis de Souza Filho.

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

R653a Rolim, Larissa Zaira Rafael.
Analysis of the dynamics and complexity of rainfall and streamflow time series in Ceará, Brazil: integrating chaos theory, information theory, machine learning and stochastic approaches / Larissa Zaira Rafael Rolim. – 2023.
186 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Civil: Recursos Hídricos, Fortaleza, 2023.
Orientação: Prof. Dr. Francisco de Assis de Souza Filho.

1. deterministic chaos. 2. complexity analysis. 3. hydrometeorological time series. 4. nonlinear methods. 5. machine learning. I. Título.

CDD 627

LARISSA ZAIRA RAFAEL ROLIM

ANALYSIS OF THE DYNAMICS AND COMPLEXITY OF RAINFALL AND STREAMFLOW TIME SERIES IN CEARÁ, BRAZIL: INTEGRATING CHAOS THEORY, INFORMATION THEORY, MACHINE LEARNING AND STOCHASTIC APPROACHES

Thesis defended at the Water Resources and Environmental Sanitation Post-Graduate Program of the Technology Center at the Federal University of Ceará, as partial requirement to obtain the doctor degree in Civil Engineering. Concentration Area: Water Resources.

Approved on: 11/07/2023.

EXAMINING COMMITTEE

Prof. Dr. Francisco de Assis de Souza Filho (Advisor)
Universidade Federal do Ceará (UFC)

Prof. Dr. Dirceu Silveira Reis Júnior
Universidade de Brasília (UNB)

Prof. Dr. Eduardo Sávio Passos Rodrigues Martins
Universidade Federal do Ceará (UFC)

Prof. Dr. Francisco das Chagas Vasconcelos Júnior
Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME)

Prof. Dra. Samiria Maria Oliveira da Silva
Universidade Federal do Ceará (UFC)

A Deus.

Aos meus pais, Edinaldo e Aurilene.

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos a todas as pessoas que me apoiaram durante minha jornada acadêmica e contribuíram para a realização desta pesquisa.

Em primeiro lugar, quero agradecer ao meu professor orientador, Prof. Dr. Francisco de Assis, por sua orientação valiosa, paciência, sabedoria e orientação impecável durante todo o processo de pesquisa. Agradeço por me apresentar esse tema onde pude entender que o caos é a ciência das surpresas, do não-linear e do imprevisível.

Aos meus pais, Edinaldo e Aurilene pela educação, apoio aos meus estudos, encorajamento e suporte que me deram ao longo deste caminho. Ao meu querido namorado, Konstantinos, por toda força, carinho, compreensão e companheirismo a mim dedicado. Sem o apoio inabalável deles, eu não teria chegado tão longe.

Gostaria de agradecer aos meus amigos por todo o apoio, incentivo e inspiração que me deram, particularmente aos colegas do laboratório de Gerenciamento do Risco Climático (GRC) nos diversos projetos e trabalhos que realizamos juntos. Também agradeço aos professores do Departamento de Engenharia Hidráulica e Ambiental pelo conhecimento adquirido e os membros da banca pelo tempo, pelas valiosas colaborações e sugestões.

Por fim, o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Sua contribuição foi fundamental para a realização desta pesquisa.

“Chaos often breeds life, when order breeds habit.” (Henry Adams).

RESUMO

Situações que são inicialmente simples e fáceis de entender em detalhes podem se tornar complicadas devido à presença do caos. Empregando uma nova combinação de métodos não lineares, teoria do caos e teoria da informação, desafia-se o paradigma determinístico/estocástico convencionalmente utilizado na dinâmica das variáveis hidrológicas. O objetivo é oferecer uma compreensão mais robusta da complexidade e do caos subjacentes a esses fenômenos e aprimorar a capacidade de previsão dessas séries temporais. A primeira fase da pesquisa concentrou-se na detecção do caos determinístico usando métodos não lineares e de teoria do caos. Os resultados revelaram que mais de 70% das series temporais de chuva e 80% das vazões apresentaram sinais de caos em escalas de tempo mensais, utilizando a dimensão de correlação. No entanto, a detecção de séries caóticas diminuiu à medida que as escalas de tempo aumentaram. A avaliação do maior expoente de Lyapunov indicou uma presença mais forte de caos nas vazões do que nas chuvas, sugerindo que as estações de chuva com caos determinístico têm períodos de previsibilidade mais longos do que seus equivalentes de vazões. Essas descobertas têm implicações cruciais para o gerenciamento de recursos hídricos e o desenvolvimento de planos integrados, especialmente considerando as limitações inerentes às previsões de longo prazo de vazão. A segunda análise utilizou métodos de teoria da informação, especificamente a entropia multiescala (MSE), para aprofundar o entendimento da complexidade das séries temporais. A análise de MSE indicou que as vazões exibem menor entropia (maior previsibilidade) em escalas de tempo menores, o que significa menor complexidade. Notavelmente, foi observada uma diminuição distinta na complexidade em metade das estações de precipitação, enquanto duas estações de vazão na região sudeste mostraram um aumento da entropia, sugerindo maior complexidade nessas séries temporais específicas. Essas descobertas destacam a importância de compreender a dinâmica hidrológica, pois a complexidade dessas séries varia espacial e temporalmente. Especificamente, a parte noroeste do estado, que é considerada mais complexa em termos de chuva e vazão. A complexidade e o comportamento caótico observados no regime hidrológico do Ceará desempenham um papel vital nos recursos hídricos. Aproveitando os resultados da detecção do caos, a terceira fase do estudo utilizou 20 séries temporais de precipitação que mostraram caos determinístico como dados de entrada para modelos de aprendizado de máquina. Os resultados mostraram que os modelos de Máquina de Vetor de Suporte e Floresta Aleatória se destacaram na previsão, entretanto, cada modelo foi adaptado a padrões de chuva únicos em diferentes locais. O desempenho bem-sucedido desses modelos demonstra o potencial de métodos

orientados por dados na previsão da dinâmica das chuvas sem a necessidade de informações físicas extensivas. A fase final do estudo aplicou um arcabouço multi-modelo, incorporando seis modelos de previsão, para prever a vazão média anual de curto e longo prazo. Os modelos híbridos superaram os modelos independentes, sugerindo a eficácia desse método para melhorar a precisão das previsões de longo prazo. No entanto, o estudo reconhece a limitação de excluir variáveis exógenas que podem influenciar a vazão, como chuva e índices climáticos. Em conclusão, esta tese de doutorado oferece um exame abrangente das complexidades inerentes às séries temporais de precipitação e vazão, além de fornecer uma metodologia inovadora para detectar, analisar e prever essas séries usando métodos não lineares, teoria do caos e teoria da informação. As descobertas da pesquisa têm um potencial significativo para melhorar a confiabilidade das previsões hidrológicas e aprimorar as estratégias de gerenciamento de recursos hídricos.

Palavras-chave: caos determinístico; análise de complexidade; séries temporais hidrometeorológicas; métodos não lineares; aprendizado de máquina; previsão de precipitação; previsão de vazão; gerenciamento de recursos hídricos.

ABSTRACT

Situations that are initially simple and easy to understand in detail can end up being complicated due to the presence of chaos. Employing a novel combination of nonlinear, chaos theory, and information theory methods, it challenges the conventional deterministic/stochastic paradigm commonly used in hydrological variable dynamics. It aims to offer a more robust understanding of the underlying complexity and chaos in these phenomena and enhance the prediction capacity of such time series. The first phase of the research focused on detecting deterministic chaos using nonlinear and chaos theory methods. The results revealed that over 70% of the rainfall and 80% of streamflow showed signs of chaos at monthly timescales, using the correlation dimension. However, the detection of chaotic series diminished as the timescales increased. The largest Lyapunov exponent assessment indicated a stronger presence of chaos in streamflow than in rainfall, suggesting that rainfall stations with deterministic chaos have longer predictability periods than their streamflow counterparts. These findings carry crucial implications for water resource management and the development of integrated plans, especially considering the inherent limitations in long-term streamflow predictions. The second analysis used information theory methods, specifically multiscale entropy (MSE), to delve into the complexity of the time series. The MSE analysis indicated that streamflow exhibits lower entropy (greater predictability) at smaller timescales, signifying less complexity over time. Notably, a distinct decrease in complexity was observed in half of the rainfall stations, while two streamflow stations in the southeastern region showed increased entropy, suggesting heightened complexity in those specific time series. These findings underline the importance of understanding hydrological dynamics, as the complexity of these series varies spatially and temporally. Specifically, the northwestern part of the state is found to be more complex in terms of both rainfall and streamflow. The complexity and chaotic behavior observed in the hydrological regime of Ceará play a vital role in water resources. Building upon the chaos detection results, the third phase of the study used 20 rainfall time series, which showed deterministic chaos, as input data for machine learning models. The results showed that Support Vector Machine and Random Forest models stood out in the prediction; however, each model was adapted to unique rainfall patterns in different locations. The successful performance of these models demonstrates the potential of data-driven methods in forecasting rainfall dynamics without extensive physical information. The final phase of the study applied a multi-model framework, incorporating six prediction models, for forecasting short- and long-term average annual streamflow. The hybrid models outperformed stand-alone models, suggesting the

efficacy of this approach for improving long-term prediction accuracy. However, the study recognizes its limitation in excluding exogenous variables that may influence streamflow, like rainfall and climatic indices. In conclusion, this doctoral thesis offers a comprehensive examination of the complexities inherent in rainfall and streamflow time series and provides an innovative methodology to detect, analyze, and predict these complexities using nonlinear, chaos theory, and information theory methods. The research findings carry significant potential for improving the reliability of hydrological forecasting and enhancing water resource management strategies.

Keywords: deterministic chaos; complexity analysis; hydrometeorological time series; nonlinear methods; machine learning; rainfall forecasting; streamflow forecasting; water resource management.

LIST OF FIGURES

Figure 1 – Complexity theory..... 24

Figure 2 – Overall methodology of this thesis..... 28

Figure 3 – Rainfall and streamflow stations location..... 32

Figure 4 – Violin plots of (a) rainfall data from the 31 analyzed stations and (b) streamflow data from the 30 analyzed stations. The maximum and minimum values are represented by points in the top and bottom of the plots, respectively, and the median is illustrated by the horizontal line on the boxplot inside the violin plot..... 35

Figure 5 – Schematic representation of the proposed chaos identification framework..... 36

Figure 6 – Attractor of reconstructed phase space for monthly rainfall (left) and streamflow (right) time series. The phase space was reconstructed with an embedding dimension of three and a one-month time delay..... 43

Figure 7 – The spatial distribution of time delay, determined by mutual information, is examined for rainfall at two different timescales: (a) monthly timescale in months and (b) seasonal timescale in terms of the number of seasons..... 44

Figure 8 – The spatial distribution of time delay, determined by mutual information, is examined for streamflow at two different timescales: (a) monthly timescale in months and (b) seasonal timescale in terms of the number of seasons..... 45

Figure 9 – Spatial distribution of time delay (years) determined by mutual information for (a) rainfall and (b) streamflow at annual timescale..... 46

Figure 10 – Spatial distribution of embedding dimension determined by the FNN method for rainfall at (a) monthly, (b) seasonal and (c) annual timescales..... 48

Figure 11 – Spatial distribution of embedding dimension determined by the FNN method for streamflow at (a) monthly, (b) seasonal and (c) annual timescales..... 48

Figure 12 – Log–log plot of the correlation integral function for (a) rainfall station #15 and (c) streamflow station #13. Correlation exponent versus embedding dimension for (b) rainfall station #15 and (d) streamflow station #13 50
Figure 13 – Spatial distribution of the chosen dimension of rainfall for which the correlation exponent versus embedding dimension plot saturates for (a) monthly, (b) seasonal and (c) annual timescale 52
Figure 14 – Spatial distribution of the chosen dimension of streamflow for which the correlation exponent versus embedding dimension plot saturates for (a) monthly and (b) seasonal timescales 52
Figure 15 – Evolution of the logarithm of the mean distance S(ν) as a function of the time step ν. The red dashed line is the regression line. Rainfall station (#10) (a) at monthly and (c) seasonal timescales; and streamflow station (#6) (b) at monthly and (d) seasonal timescales 53
Figure 16 – Spatial distribution of the largest Lyapunov exponent for the rainfall stations at (a) monthly, (b) seasonal and (c) annual timescale 55
Figure 17 – Spatial distribution of the largest Lyapunov exponent for the streamflow stations at (a) monthly and (b) seasonal timescale 55
Figure 18 – Examples of time series and recurrence plot (RP) based on monthly rainfall time series at selected stations (a) Time series of station #2 and (c) its RP. (b) Time series of station #6 and (d) its RP. (e) Time series of station #8 and (g) its RP. (f) Time series of station #15 and (h) its RP. (i) Time series of station #27 and (k) its RP. (j) Time series of station #28 and (l) its RP 56
Figure 19 – Examples of time series and recurrence plot (RP) based on monthly streamflow time series at selected stations (a) Time series of station #4 and (c) its RP. (b) Time series of station #12 and (d) its RP. (e) Time series of station #18 and (g) its RP. (f) Time series of station #23 and (h) its RP. (i) Time series of station #25 and (k) its RP. (j) Time series of station #29 and (l) its RP 59
Figure 20 – (a) El Niño 3.4 and (b) its recurrence plot 61
Figure 21 – Location of the meteorological stations 69

Figure 22 – MSE for monthly rainfall and monthly anomalies for stations from 1 to 16..... 76
Figure 23 – MSE for monthly rainfall and monthly anomalies for stations from 17 to 31.... 77
Figure 24 – MSE for monthly streamflow and monthly anomalies for stations from 1 to 15..... 78
Figure 25 – MSE for monthly streamflow and monthly anomalies for stations from 16 to 30..... 79
Figure 26 – Spatial distribution of MSE from $\tau = 1$ to $\tau = 12$ months over the state of Ceará for rainfall original series..... 80
Figure 27 – Spatial distribution of MSE from $\tau = 1$ to $\tau = 12$ months over the state of Ceará for streamflow original series..... 81
Figure 28 – The location of the hydrometeorological stations in Ceará State, Brazil..... 90
Figure 29 – LSTM-based neural network structure..... 94
Figure 30 – Mutual information with delay time for monthly rainfall stations (a) 1 through 10 and (b) 11 through 20..... 98
Figure 31 – Spatial distribution of the chosen dimension for the rainfall stations according to the prediction accuracy..... 100
Figure 32 – Comparison between measured and forecasted rainfall over the validation phase for stations 1 through 10..... 104
Figure 33 – Comparison between measured and forecasted rainfall over the validation phase for stations 11 through 20..... 105
Figure 34 – Comparison between measured and forecasted rainfall and best fit lines by the DT, LAP, LSTM, RF and SVM for the best accurate input combinations over the testing phase for the Maracanu, São Gonçalo do Amarante, Caucaia, Carire and Mucambo station..... 110
Figure 35 – Comparison between measured and forecasted rainfall and best fit lines by the DT, LAP, LSTM, RF and SVM for the best accurate input combinations over the testing phase for the Tianguá, Ubajara, Mulungu, Croatá and Catunda stations..... 111
Figure 36 – Comparison between measured and forecasted rainfall and best fit lines by the DT, LAP, LSTM, RF and SVM for the best accurate input combinations over the testing phase for the Ipueiras, Alto Santo, São João do Jaguaribe, Solonópole and Umari stations..... 112

Figure 37 – Comparison between measured and forecasted rainfall and best fit lines by the DT, LAP, LSTM, RF and SVM for the best accurate input combinations over the testing phase for the Aurora, Acopiara, Missão Velha, Jaguaruana and Santana do Cariri stations..... 113
Figure 38 – Taylor diagram of the forecasted rainfall over the testing period for all stations..... 114
Figure 39 – Location of the meteorological stations analyzed in this study..... 122
Figure 40 – The overall processes of the multi-model methodologies applied in this study. The original series was decomposed using the CEEMDAN into n IMFs and a residue term (r) and using the WT into n daughter wavelets..... 123
Figure 41 – (a) Original annual streamflow time series, (b) decomposed results for the Orós reservoir using CEEMDAN and (c) decomposed results for the Orós reservoir using WT..... 127
Figure 42 – Power spectrum XTC (a) streamflow and El Nino 3.4, (b) streamflow and PDO and (c) streamflow and AMO and Power spectrum WTC (d) streamflow and El Nino 3.4, (e) streamflow and PDO and (f) streamflow and AMO..... 129
Figure 43 – Forecasting ensembles (boxplot) and the historic flows (red) for the AR model..... 135
Figure 44 – Forecasting ensembles (boxplot) and the historic flows (red) for the HMM model..... 135
Figure 45 – Forecasting ensembles (boxplot) and the historic flows (red) for the WARM model..... 136
Figure 46 – Forecasting ensembles (boxplot) and the historic flows (red) for the CEEMDAN-AR model..... 137
Figure 47 – Forecasting ensembles (boxplot) and the historic flows (red) for the WHMM model..... 137
Figure 48 – Forecasting ensembles (boxplot) and the historic flows (red) for the CEEMDAN-HMM model..... 138
Figure 49 – Forecasted and observed average annual streamflow during the 3-, 5-, 10-, and 15-year period by WARM, WHMM, CEEMDAN-AR, CEEMDAN- 139

HMM, AR, HMM, and Multi-model at Iguatu station.....

LIST OF TABLES

Table 1 – Identification of the streamflow stations and its statistical characteristics..... 32

Table 2 – Selected dimension for which the correlation exponent versus embedding dimension plot saturates for all the analyzed stations and timescales..... 50

Table 3 – Summary of RQA measures at different rainfall stations..... 63

Table 4 – Summary of RQA measures at different streamflow stations..... 64

Table 5 – Geophysical and statistical information for rainfall time series (1962-2009) of 31 meteorological stations..... 70

Table 6 – Geophysical and statistical information for streamflow time series of 30 meteorological stations..... 71

Table 7 – Trend of rainfall complexity change of the analyzed rainfall stations..... 82

Table 8 – Trend of streamflow complexity change of the selected streamflow stations..... 83

Table 9 – Hurst exponent for analyzed rainfall and streamflow stations..... 84

Table 10 – Statistical summary of the monthly rainfall time series..... 90

Table 11 – Chosen embedding dimension for the training dataset according to the accuracy of the models..... 99

Table 12 – Performance of machine learning models for monthly rainfall in the validation phase..... 101

Table 13 – Performance of machine learning models for monthly rainfall in the testing phase..... 106

Table 14 – Statistical information of streamflow data observed at the hydrological station 122

Table 15 – Time period of model training, validation, and prediction for the multi-model..... 126

Table 16 – The optimal order of AR model and number of hidden states for the HMM in the hybrid models for 3-, 5-, 10- and 15-year forecasts..... 130

Table 17 – The optimal order of AR model and number of hidden states for the HMM in the models without pre-processing for 3-, 5-, 10- and 15-year forecasts..... 131
Table 18 – Evaluation of hybrid models in modeling the subseries of streamflow for the training and validation periods..... 131
Table 19 – Evaluation of AR and HMM models in modeling the subseries of streamflow for the training and validation periods..... 132
Table 20 – Regression coefficients for the LASSO regression for each forecast window for 3-, 5-, 10- and 15-year forecasts..... 132
Table 21 – The prediction performance indicators of WARM, WHMM, CEEMDAN-AR, CEEMDAN-HMM, AR, HMM, and Multi-model at Iguatu station..... 133

SUMMARY

1	**INTRODUCTION**	22
2	**OBJECTIVES**	26
2.1	**Main Objective**	26
2.2	**Specific Objectives**	26
3	**METHODOLOGICAL STRUCTURE**	27
4	**EXPLORING SPATIOTEMPORAL CHAOS IN HYDROLOGICAL DATA: EVIDENCE FROM CEARÁ, BRAZIL**	29
4.1	**Introduction**	29
4.2	**Study area and hydrological data**	31
4.3	**Methods**	35
4.3.1	***Phase space reconstruction (PSR)***	35
4.3.1.1	*Time delay selection*	37
4.3.1.2	*Embedding dimension*	38
4.3.2	***Lyapunov Exponent***	40
4.3.3	***Recurrence Plot (RP) and Recurrence Quantification Analysis (RQA)***	40
4.4	**Results and Discussion**	42
4.4.1	***Phase-space reconstruction***	42
4.4.2	***Estimation of Lyapunov exponent***	52
4.4.3	***RQA***	54
4.5	**Conclusion**	64
5	**COMPLEXITY ANALYSIS OF RAINFALL AND STREAMFLOW APPLYING MULTISCALE ENTROPY TO THE STATE OF CEARÁ, BRAZIL**	66
5.1	**Introduction**	66
5.2	**Study area and hydrological data**	68
5.3	**Methods**	71
5.3.1	***Sample Entropy***	71
5.3.2	***Multiscale Entropy***	72
5.3.3	***Trend analysis for complexity change***	73
5.3.4	***Hurst exponent***	73
5.4	**Results and discussion**	74

5.4 **Conclusions** ... 84
6 **PHASE-SPACE RECONSTRUCTION WITH MACHINE LEARNING METHODS: APPLICATION OF RAINFALL FORECASTING TO THE STATE OF CEARÁ, BRAZIL** ... 86
6.1 **Introduction** ... 86
6.2 **Study area and hydrological data** ... 88
6.3 **Methods** ... 90
6.3.1 ***Phase Space Reconstruction (PSR)*** ... 90
6.3.2 ***Nonlinear Local Approximation Prediction*** ... 91
6.3.3 ***Long Short-Term Memory (LSTM) Neural Network*** ... 91
6.3.4 ***Decision Tree (DT)*** ... 93
6.3.5 ***Random Forest (RF)*** ... 94
6.3.6 ***Support Vector Machine (SVM)*** ... 94
6.3.7 ***Cross-validation in time series*** ... 95
6.3.8 ***Chaos theory and machine learning*** ... 96
6.4 **Results and discussion** ... 97
6.4.1 ***Phase-space reconstruction*** ... 97
6.4.2 ***Comparisons of the results*** ... 99
6.5 **Conclusions** ... 115
7 **A MULTI-MODEL FRAMEWORK FOR STREAMFLOW FORECASTING BASED ON STOCHASTIC MODELS: AN APPLICATION TO THE STATE OF CEARÁ, BRAZIL** ... 117
7.1 **Introduction** ... 117
7.2 **Study area and hydrological data** ... 120
7.3 **Methods** ... 121
7.3.1 ***Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)*** ... 123
7.3.2 ***Wavelet Transform*** ... 123
7.3.3 ***Hidden Markov Model*** ... 124
7.3.4 ***Multi-Model Framework*** ... 124
7.4 **Results and discussion** ... 125
7.4.1 ***The original series decomposed*** ... 125
7.4.2 ***Method for parameter selection*** ... 129

7.4.3 ***Comparative analysis*** 132
7.5 **Conclusion** 139
8 **FINAL CONCLUSIONS AND REMARKS** 141
REFERENCES 145
APPENDIX A – SUPPLEMENTAL FILES OF CHAPTER 4 167
APPENDIX B – SUPPLEMENTAL FILES OF CHAPTER 6 168

1 INTRODUCTION

Water resources are considered dynamic systems because their configuration can change over time and space, and different processes regulate their variables at various spatial and temporal scales. Understanding these complex and highly nonlinear variables has been a challenging task and a major focus for water resources engineers and managers, particularly in arid regions. Arid and semi-arid regions are characterized by high temperatures, low annual precipitation, frequent droughts, and variations at inter-annual to higher scales. These factors contribute to the vulnerability of these areas (RAMARAO et al. 2019; SINGH & CHUDASAMA, 2021). Therefore, accurate predictions are essential for water resources planning, management, and operation.

Hydrological forecast models use observed time series as input variables. Over the past half a century, hydrologic time series analysis has become a vital part of hydrologic studies. Depending upon the hydrologic variable and the period of observation, a hydrologic time series may be composed of deterministic events, stochastic events, or a combination of the two. Although, for an extended period of observations, a hydrologic time series is usually a combination of stochastic components superimposed on deterministic components (SIVAKUMAR, 2016). Furthermore, the deterministic component can display inherent nonlinearity and sensitivity to initial conditions. The main focus of this doctoral thesis will be on examining the deterministic chaos component characterized by its sensitivity to initial conditions.

Numerous time series analysis methods have found applications in hydrology, such as stochastic and data-driven techniques. Among these methods, linear stochastic methods are much more popular and well-established, including Box & Jenkins (1970), Yevjevich (1972), and Salas (1980). This popularity can be attributed, in part, to earlier developments and the assumption that hydrologic processes are stochastic in nature. Lack of computational power to develop the nonlinear mathematical models was an important factor that contributed to the use of linear approaches. However, fast developments in data measurement and computer technologies have made other methods equally attractive, such as those based on nonlinear dynamic and chaos theories, which are exemplified in Tsonis (1992), Sivakumar (2000), and Kantz & Schreiber (2004).

In the nonlinear science literature, the term chaos refers to the occurrence of seemingly random and complex behaviors arising from simple, nonlinear deterministic systems that have a sensitive dependence on initial conditions (LORENZ, 1963). Due to its properties

(i.e., nonlinear inter-dependence, hidden determinism and order, and sensitivity to initial conditions), chaos theory has been relevant for hydrologic systems and processes, chaos theory has found a growing number of applications in hydrological time series (SIVAKUMAR, 2000, 2009, 2016).

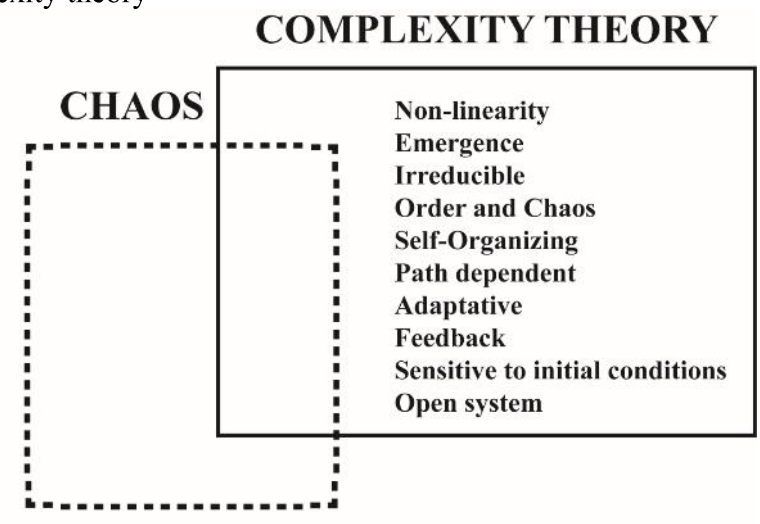
Another intriguing aspect in water resources is complexity. Numerous authors have deliberated on the concept of complexity and its significance in comprehending hydrological systems, although a consensus on its definition remains elusive due to its subjective nature. Nonetheless, a viable definition of complexity entails a system comprising inter-connected components that interact in a nonlinear fashion, and nonlinearity is a common feature shared by both chaotic and complex system (SIVAKUMAR, 2016).

A complex system is characterized by its intricate organization, inherent uncertainty, nonlinear dynamics, interactions between scales, feedback loops, and emergent behavior (RICKLES et al., 2007). The many elements that compose the complex system have different components interacting at multiple scales. At each scale, a different structure can be form, leading to emergence, which is another important property of complex systems. Emergence is a concept in the study of complex systems that refers to the property of a system in which collective behavior arises from the interactions of its individual components, resulting in properties or patterns that cannot be predicted from the individual components alone. Self-organization is a necessary condition for emergence to occur in complex systems as it provides the underlying structure for the interactions of the individual components that result in emergent behaviors and patterns (SIVAKUMAR, 2016).

The individual components of a complex system spontaneously arrange themselves into structured and organized forms, creating the conditions for emergence to occur and resulting in collective behaviors and patterns characteristic of the system as a whole. Complex systems have multiple scales, and while chaos may exist on one scale, self-organization may occur on a coarser scale above it (BARANGER, 2000). A chaotic system is characterized by the generation of complicated, aperiodic, and seemingly random behavior. This behavior arises from the iteration of a simple rule and is chaotic in a precise mathematical sense. The complicatedness in chaotic systems is not the same as complexity in the context of complex systems science. On the other hand, a complex system is defined by the generation of rich, collective dynamical behavior. This behavior emerges from simple interactions between many subunits. Complexity in complex systems is not necessarily chaotic, and chaotic systems are not necessarily complex. However, complex systems can exhibit chaotic behavior under certain conditions or values of variables or control parameters (RICKLES et al., 2007). In its basic

form, complexity theory includes the concepts of chaos and complex adaptive systems (CAS), along with path dependence, system history, nonlinearity, emergence, irreducibility, adaptiveness, operating between order and chaos, and self-organization (TURNER and BAKER, 2019). Chaos is supported by self-organization, feedback, and deterministic systems, while CAS are supported through self-organization, emergence, adaptation/evolution, feedback/history, and non-deterministic systems as described in Figure 1.

Figure 1 - Complexity theory



Source: Adapted from Turner and Baker (2019).

In hydrology, complexity can refer to the number of dominant variables and the nonlinearity of processes that govern watershed dynamics (Sivakumar, 2016). Further, hydrological complexity is closely connected to concepts such as model parsimony, parameter identifiability, and criteria for model selection (Ombadi et al., 2021). In this doctoral thesis, we will focus on the first concept. To quantitatively understand complexity, various tools can be used, including statistical measures like the coefficient of variation, nonlinear dynamic measures such as dimension, information-theoretic measures like entropy, and others. For instance, complexity can be assessed based on the number of variables that predominantly control the system, requiring a certain amount of information to describe the system accurately.

While complex systems may exhibit patterns, structures, and self-organizing behavior, their overall behavior can still be highly dynamic, difficult to anticipate, and characterized by multiple possible outcomes. On the other hand, chaotic systems exhibit deterministic behavior, meaning that their future states are completely determined by their initial conditions and the rules governing their dynamics. However, due to the complexity and nonlinearity of these systems, long-term prediction becomes challenging, as small errors or uncertainties in measuring or knowing the initial conditions can quickly amplify and lead to

substantial deviations from predicted outcomes (POLLARD et al., 2011). Therefore, predicting hydrological variables in the long-term presents a challenge due to the inherent characteristics of hydrological time series.

Generally, the water availability in reservoirs is estimated using inflow forecasts at various time scales such as daily, weekly, and monthly (SILVA et al., 2017). The commonly used models are stochastic models which may overlook important characteristics of the time series, leading to inaccurate forecasts in the face of climate variability. This raises questions about the possibility of accurately predicting hydrological variables using models that consider the inherent nonlinearity and complexity of hydrological series. Can the predicted time series provide a good database for optimizing the operation of a reservoir system in the presence of climate variability? The scientific justification of this work is given by the importance and innovative nature of improving hydrological forecasts by incorporating the characteristics of nonlinearity, complexity, and chaotic behavior.

Situations that are initially simple and easy to understand in detail can end up being complicated due to the presence of chaos. Further, the focus of this doctoral thesis was to propose a framework that identifies the presence of chaos in rainfall and streamflow time series using classical methods from chaos theory and nonlinear methods. Furthermore, it is also important to quantify the complexity of hydrological time series and how it evolves spatially and temporally. Thus, method based on information theory are applied to evaluate the spatiotemporal changes of the relative complexity of time series including series with multiple temporal scale characteristics. The latest state-of-the-art machine learning methods are applied to predict the chaotic time series, and a multi-model framework with stochastic methods is proposed to improve the modeling process.

2 OBJECTIVES

2.1 Main Objective

To develop a methodology for detecting deterministic chaos in rainfall and streamflow time series, and incorporating nonlinearity through the use of nonlinear, chaos theory, and information theory methods. Furthermore, based on the previous diagnosis, a methodology is developed to predict deterministic chaos using machine learning and a multi-model based on stochastic models.

2.2 Specific Objectives

1. To identify the underlying dynamics of monthly, seasonal, and annual rainfall and streamflow in the State of Ceará, Brazil.
2. To assess the spatiotemporal behavior of rainfall and streamflow and determine their level of complexity using the Multiscale Entropy method.
3. To investigate the relation between rainfall and streamflow complexity.
4. To evaluate Machine Learning techniques in forecasting hydrological time series with chaotic characteristics.
5. To assess the effectiveness of a multi-model approach incorporating hybrid models for forecasting hydrological time series with a stochastic approach.

3 METHODOLOGICAL STRUCTURES

To address the topics outlined in the introduction and meet the proposed objectives, this doctoral thesis is structured into four articles and a final section summarizing the key conclusions and recommendations for future work. The first focuses on detecting deterministic chaos in observed hydrometeorological variables (i.e., rainfall and streamflow), using different methodologies based on chaos theory and nonlinear methods. Additionally, this article uses Recurrence Plots to comprehend the relationship between these hydrometeorological variables and El Niño events.

The second article aims to explore the underlying spatiotemporal complexity of hydrological variables. Given the nonlinear and nonstationary behavior of the climate system, understanding its inherent complexity has both theoretical and practical significance in revealing the uncertainty and variability of the system. In this article, sample entropy (SampEn), which provides a measure of complexity obtained by examining the similarity of observed segments from a time series, and Multiscale entropy (MSE), which provides insights into the complexity of fluctuations over a range of time scales and extends standard sample entropy measures, are used. The article also evaluates the intra-annual distribution of rainfall complexity, and the relationship between rainfall and streamflow complexity.

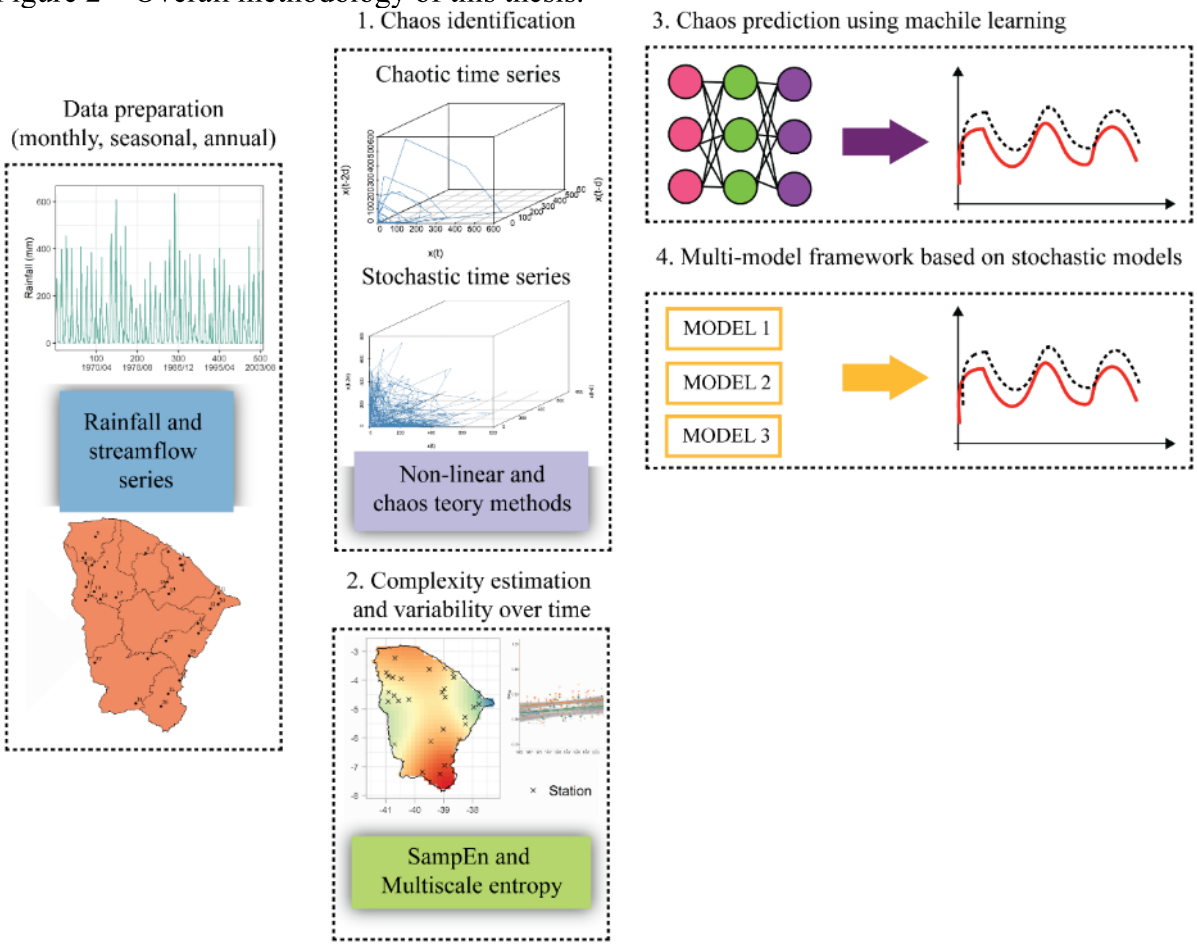
The results from the first and second articles diagnose the dimensionality, complexity, and variability of the time series using nonlinear dynamic approaches, providing a deeper understanding of rainfall and streamflow variability at various spatiotemporal scales. The third article incorporates the results from the previous articles, such as the time-delay, to predict chaotic time series using different machine learning methods. These methods can capture the nonlinearity present and improve streamflow modeling and forecasting performance.

The final article proposes a multi-model framework to predict short- and long-term streamflow time series, and a comparative analysis of different hybrid prediction models. The hybrid model was based on decomposition methods (i.e., wavelet and complete ensemble empirical mode decomposition with adaptive noise) along with stochastic models (i.e., autoregressive model and hidden Markov model). The forecast models were coupled using the least absolute shrinkage and selection operator (LASSO) regression method. This framework is adopted as preprocessing approaches can enhance the accuracy of streamflow forecasting, particularly for long-term forecasting. Additionally, the cross-wavelet transform, and the wavelet transform coherence are applied to analyze the effect of climatic indices such as El

Niño, PDO, and AMO on the analyzed streamflow.

In this context, the articles represent important steps, illustrated in Figure 2, when dealing with time series: diagnosing its dynamics, its complexity, and incorporating these important features of hydrological time series into the modeling framework. Thus, this methodological structure is not limited to rainfall and streamflow-related studies and can be applied to different fields of science dealing with nonlinear systems.

Figure 2 – Overall methodology of this thesis.



Source: Prepared by the author

4 EXPLORING SPATIOTEMPORAL CHAOS IN HYDROLOGICAL DATA: EVIDENCE FROM CEARÁ, BRAZIL[1].

4.1 Introduction

The study of hydrological systems' characteristics has long been a subject of interest due to the presence of complex, highly variable, random, and interdependent dynamics. Accurate information on hydrological variables at different timescales is crucial for monitoring and managing water resources, especially when dealing with extreme events such as droughts and floods (PHAM et al., 2020; SHU et al. 2021). The complexity of these systems, which this paper refers to as the number of dominant variables governing the time series dynamics, is often associated with a random process, leading to the use of stochastic methods. However, seemingly random processes can also arise from the sensitivity to initial conditions in deterministic systems.

Edward Lorenz first described chaos theory in the early 1960s, and since then, it has gained traction in several fields of natural sciences and engineering. Chaotic systems are characterized as random-looking and complex systems originating from simple deterministic systems sensitive to slight changes in initial conditions (LABAT et al., 2016; JIANG et al., 2021). This sensitivity has significant implications for system modeling and prediction. Variables considered to be random are irreproducible and unpredictable, whereas chaotic variables can be reproducible in the short term due to their deterministic nature. However, due to sensitivity to initial conditions, these variables are not predictable in the long-term. Consequently, chaos and randomness exhibit very different behaviors (SIVAKUMAR, 2016).

The application of chaos theory has demonstrated that chaotic models are often better suited for certain complex hydrological time series (SIVAKUMAR, 2000). An increasing number of studies have applied chaos theory to analyze hydrological process dynamics, starting from the late 1980s (RODRIGUEZ-ITURBE et al., 1989; HU et al., 2013; OUYANG et al., 2016; HONG et al., 2016; ZHOU et al., 2022). Applications of chaos theory in hydrology began with identifying chaos, primarily in rainfall data (RODRIGUEZ-ITURBE et al., 1989), and its development has significantly increased since then. Chaos theory has been used to address

[1] Reproduced with permission from Springer Nature. This version of the article has been accepted for publication and is available online at: https://doi.org/10.1007/s00477-023-02501-5.

different problems in hydrological systems, including characterization (ABARBANEL & LALL, 1996), noise reduction (ELSHORBAGY et al., 2002), missing data, prediction (TONGAL & BERNDTSSON, 2014; OUYANG et al., 2016; HONG et al., 2016; TONGAL, 2020; WANG et al., 2021; ZHOU et al., 2022), scaling, catchment classification (SIVAKUMAR & SINGH, 2012), and disaggregation, among others.

Different methods have been used to analyze and identify chaotic behavior. For example, Xu et al. (2009) applied the Lyapunov exponent and correlation dimension methods to recognize the chaotic features in the annual runoff of the Tarim River, finding that it exhibits complex nonlinear characteristics with chaotic dynamics. Kedra (2013) used several independent methods and tools to analyze daily discharge from gauging stations in southern Poland, including the surrogate and the determinism tests. Labat et al. (2016) applied the correlation dimension method to streamflow data from karstic watersheds in France, detecting the presence of chaos with attractor dimension values below three. Jiang et al. (2020) applied the Lyapunov exponent, nonlinear prediction, and correlation dimension methods to analyze monthly streamflow data from the Daiying hydrological station in northern China, concluding that the monthly streamflow is chaotic. Shu et al. (2021) identified features of chaos in daily rainfall data from the UK and conducted an extended complexity analysis using recurrence quantification analysis.

Furthermore, several studies have employed chaos theory to properly understand the underlying dynamics that control the behavior of hydrological time series across different timescales. However, there is a lack of quantitative analysis of chaotic characteristics in semi-arid regions. This shortfall might be related to the idea that chaos characterization and prediction methods require long, noise-free time series, while data from these regions often consist of short series with missing values. Nevertheless, some studies have argued that data size is not a significant issue for identifying and predicting chaos and that available methods can provide reliable results even with small time series (SIVAKUMAR 2005; SIEK & SOLOMATINE, 2010).

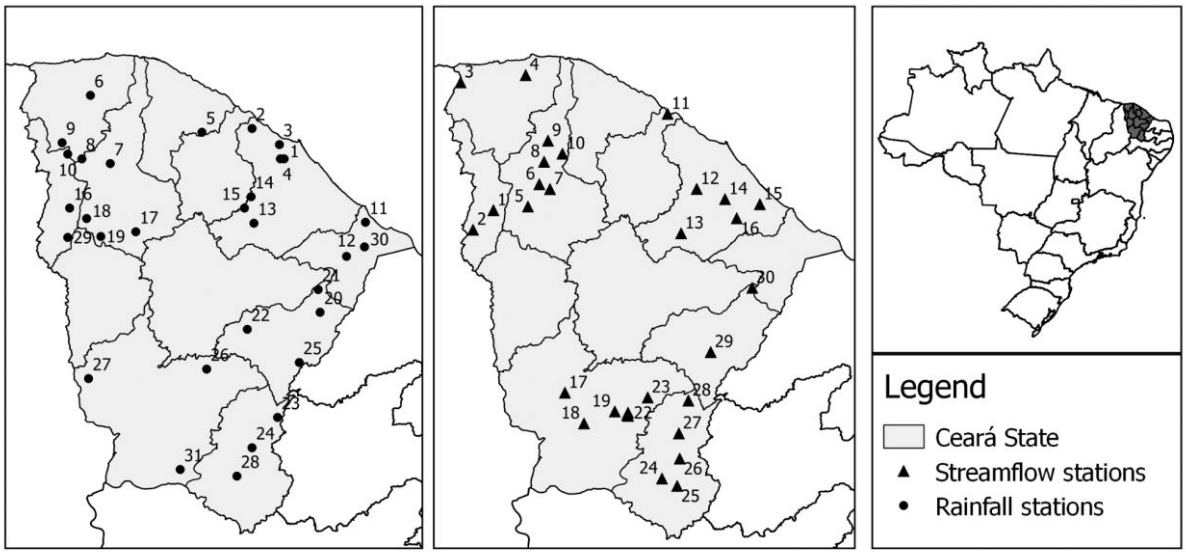
Understanding the spatiotemporal dynamics of rainfall and streamflow is critical for effective water resource management in regions with a highly variable climate, such as Ceará, Brazil. Traditional linear models often fail to capture the complex, nonlinear relationships between them, potentially leading to inaccurate predictions and inefficient water management practices. Nonlinear dynamic analysis and chaos theory can serve as reliable alternatives for modeling the complex behavior of hydrological data. Thus, this paper aims to

employ these methodologies to identify the underlying spatiotemporal dynamics of monthly, seasonal, and annual rainfall and streamflow in Ceará, Brazil. Well-established methods, such as correlation dimension, Lyapunov exponent, and recurrence quantification analysis, are applied in this study. Through the application of these methodologies, this study seeks to improve our understanding of the complex interactions and feedback between these variables. Additionally, comprehending the spatial variability of rainfall dynamics is crucial for the interpolation/extrapolation of hydrological variables and classification of catchments, which can inform more effective water management practices in the region.

4.2 Study area and hydrological data

Ceará, as shown in Figure 3, is located in the Northeast region of Brazil (NEB). The state's climate is predominantly semi-arid, with the economy heavily dependent on rainfall due to its agricultural base. The rainfall regime is mainly controlled by the southward movement of the Intertropical Convergence Zone (ITCZ), with rainfall concentrated from February to May, which accounts for approximately 70% of the annual rainfall. The mean annual rainfall is around 810 mm, with higher values in the coastal area and lower values in the central portion, where the semi-arid landscape is more accentuated (COSTA et al., 2021). Natural fluctuations in the sea surface temperature also modulate rainfall patterns. Several studies identify the influence of climate indices such as the Atlantic Multidecadal Oscillation (AMO), El Niño Southern Oscillation (ENSO) phenomenon, and Pacific Decadal Oscillation (PDO). These indices have been associated with the rainfall regime in the NEB (ANDREOLI & KAYANO, 2005; KAYANO & ANDREOLI, 2006; GARREAUD et al., 2009; KAYANO et al., 2020). The shallow soils with a crystalline basement in most of the state result in low base flows, and the rivers in the region are mainly naturally ephemeral or intermittent due to the lack of sustained groundwater recharge.

Figure 3 - Rainfall and streamflow stations location



Source: Prepared by the author

The rainfall and streamflow data used in the spatiotemporal analysis were obtained from the Brazilian National Water Agency (ANA). The monthly rainfall series ranges from January 1962 to December 2006. The length of the monthly streamflow series varies, and so do its start and end dates, depending on geographical location (refer to Table 1). This study includes data from 31 rain gauges and 30 streamflow gauges located in the state of Ceará, as illustrated in Figure 3.

Table 1 - Identification of the streamflow stations and its statistical characteristics.

Rainfall Data					
Station Number	Station ID	Station Name	Start – end of the series (Month-Year)	Mean Annual Rainfall (mm)	Coefficient of variation (%)
1	338005	Maracanau	01-1962/12-2006	1333.32	38.78
2	338008	São Gonçalo do Amarante	01-1962/12-2006	1076.51	39.19
3	338009	Caucaia	01-1962/12-2006	1340.72	34.87
4	338016	Maranguape	01-1962/12-2006	1254.94	34.71
5	339034	Uruburetama	01-1962/12-2006	1173.36	33.62
6	340008	Martinopole	01-1962/12-2006	1104.34	36.87
7	340014	Carire	01-1962/12-2006	951.93	37.02
8	340023	Mucambo	01-1962/12-2006	1040.49	36.14
9	340030	Tianguá	01-1962/12-2006	1270.84	33.27

Station Number	Station ID	Station Name	Start – end of the series (Month-Year)	Mean Annual Rainfall (mm)	Coefficient of variation (%)
10	340031	Ubajara	01-1962/12-2006	1528.52	33.85
11	437000	Aracati	01-1962/12-2006	991.02	51.03
12	437010	Russas	01-1962/12-2006	771.42	40.51
13	438032	Itapiuna	01-1962/12-2006	783.11	35.49
14	439008	Mulungu	01-1962/12-2006	1187.90	33.52
15	439018	Aratuba	01-1962/12-2006	1449.65	44.45
16	440005	Croatá	01-1962/12-2006	579.52	39.76
17	440009	Catunda	01-1962/12-2006	720.25	46.07
18	440014	Ipueiras	01-1962/12-2006	972.67	39.45
19	440017	Nova russas	01-1962/12-2006	860.34	39.11
20	538003	Alto santo	01-1962/12-2006	775.31	37.5
21	538010	São João do Jaguaribe	01-1962/12-2006	765.93	41.27
22	539023	Solonópole	01-1962/12-2006	782.61	31.53
23	638008	Umari	01-1962/12-2006	792.64	30.23
24	638010	Aurora	01-1962/12-2006	934.20	30.64
25	638011	Pereiro	01-1962/12-2006	1024.07	35.4
26	639021	Acopiara	01-1962/12-2006	791.03	31.52
27	640015	Parambu	01-1962/12-2006	697.91	47.8
28	739007	Missão Velha	01-1962/12-2006	1006.12	26.61
29	440018	Poranga	01-1962/12-2006	657.56	42.58
30	437006	Jaguaruana	01-1962/12-2006	824.63	50.51
31	739005	Santana do Cariri	01-1962/12-2006	930.15	30.29

Streamflow data

Station Number	Station ID	Station Name	Start – end of the series (Month-Year)	Mean Annual Rainfall (mm)	Coefficient of variation (%)
1	34730000	Croatá	01-1998/12-2006	3.47	57.96
2	34740000	Saudoso	01-1998/12-2010	3.64	83.98
3	35050000	Chaval Retiro	01-2001/12-2008	4.79	48.92
4	35205000	Pesqueiro	01-2001/12-2015	4.93	112.85
5	35210000	Fazenda Cajazeiras	01-1963/12-1972	8.02	99.67
6	35235000	Várzea do grosso	01-1987/12-2005	7.96	85.16
7	35240000	Trapiá	01-1988/12-2010	6.56	134.86
8	35260000	Groairas	01-1991/12-2010	4.48	170.93
9	35275000	Sobral	01-2006/12-2017	21.70	184.53
10	35279000	Fazenda bela Vista	01-1985/12-2003	0.23	147.13

Station Number	Station ID	Station Name	Start – end of the series (Month-Year)	Mean Annual Rainfall (mm)	Coefficient of variation (%)
11	35570000	São Luíz do Curu	01-1973/12-1984	19.94	198.68
12	35740000	Barra Nova	01-1999/12-2017	1.43	92.44
13	35830000	Caio Prado	01-1985/12-2005	4.97	128.22
14	35880000	Chorozinho	01-1970/12-1983	12.68	149.61
15	35900000	Itapeim	01-2001/12-2008	4.24	55.5
16	35950000	Cristais	01-1970/12-1981	6.12	180.99
17	36020000	Arneiroz	01-1976/12-2003	8.70	199.8
18	36045000	Malhada	01-1988/12-2010	1.62	169.75
19	36070000	Sítio Patos	01-2000/12-2017	7.40	158.49
20	36110000	Sítio Conceição	01-1968/12-1980	5.37	91.79
21	36125000	Sítio Poço Dantas	01-1991/12-2011	3.99	114.44
22	36130000	Cariús	01-1992/12-2006	8.45	135.71
23	36160000	Iguatu	01-1974/12-1996	33.93	141.22
24	36210000	Sítio lapinha	01-1985/12-2006	4.48	82.43
25	36250000	Podimirim	01-1973/12-2017	3.52	133.9
26	36260000	Sítio santa Cruz	01-2001/12-2017	6.70	102.52
27	36270000	Lavras da Mangabeira	01-1980/12-1996	26.27	119.45
28	36290000	Icó	01-1984/12-2015	25.91	117.33
29	36320000	Jaguaribe	01-1987/12-2017	41.70	112.08
30	36390000	Peixe gordo	01-1998/12-2014	40.97	135.44

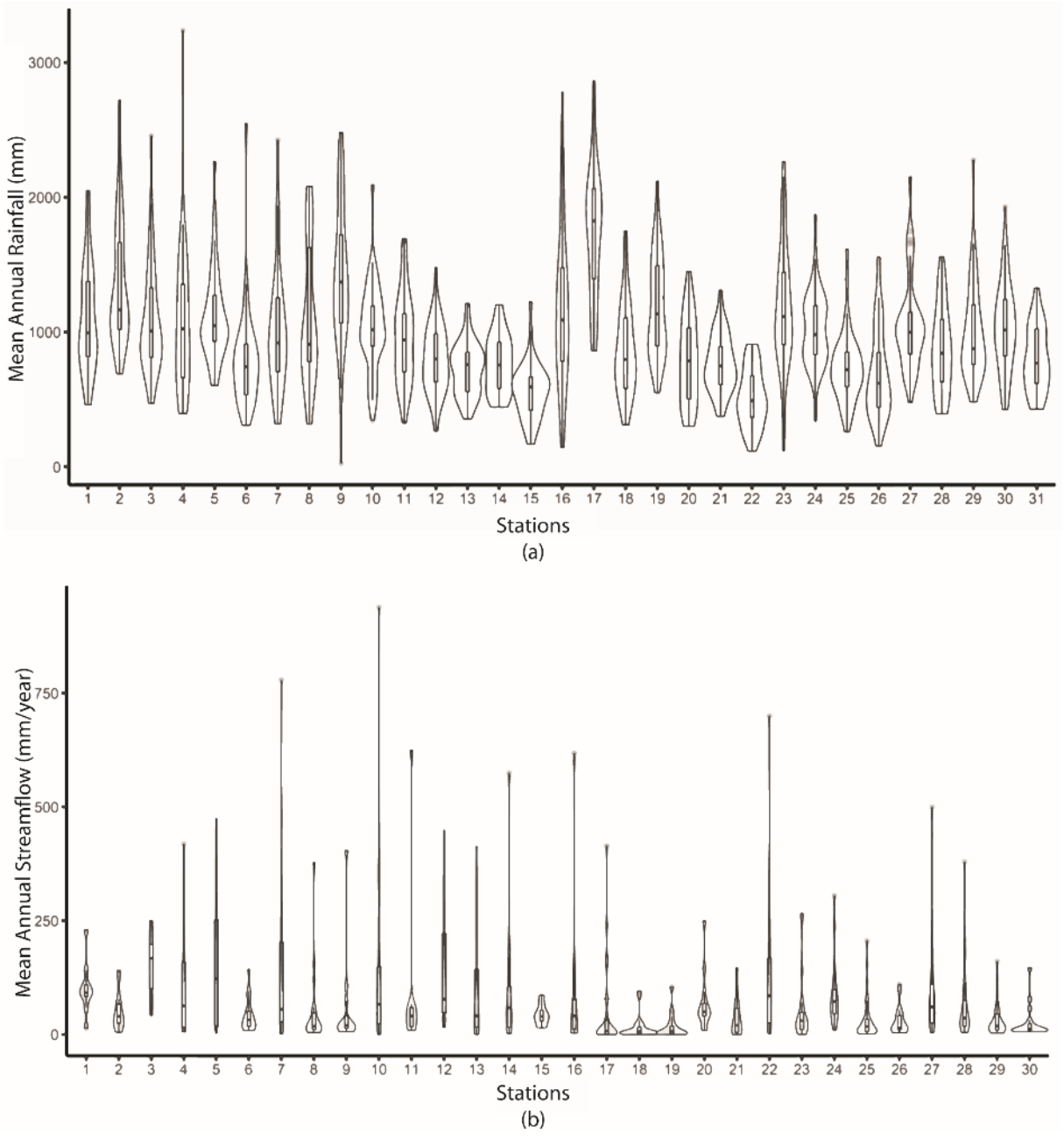
Source: Prepared by the author

The streamflow and rainfall time series each have very distinct statistical distributions. As seen in Figure 4a for rainfall analysis, the stations in the northern part of the State show the highest variability in rainfall, while those in the central and northeastern areas exhibit the narrowest range of values. The stations located in the northern and western regions have higher coefficients of variation (CV) (Table 1), indicating a large amount of variability in their datasets. Some stations display a violin plot that is thick at the center, indicating a relatively narrow range of values.

For streamflow (Figure 4b), the stations in the western area exhibit very high variability, as observed for rainfall. The stations in the eastern part of the state have the highest

mean annual streamflow, around 25m³/s. Further, the streamflow shows higher CV values compared to rainfall, with the higher values located in the northern of the state. However, the characterization of streamflow may be affected by the different time window analyzed due to the amount of missing data in the streamflow time series.

Figure 4 - Violin plots of (a) rainfall data from the 31 analyzed stations and (b) streamflow data from the 30 analyzed stations. The maximum and minimum values are represented by points in the top and bottom of the plots, respectively, and the median is illustrated by the horizontal line on the boxplot inside the violin plot.

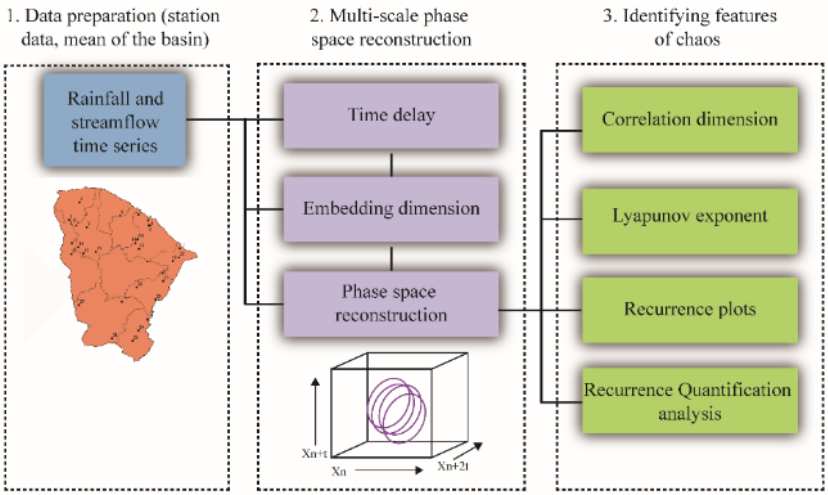


Source: Prepared by the author

4.3 Methods

Several methods have been developed to detect the presence of chaos in time series. These include the correlation dimension method (GRASSBERGER & PROCACCIA, 1983a), the Kolmogorov entropy method (BENETTIN et al., 1979), the Lyapunov exponent method (WOLF et al., 1985), the false nearest neighbor algorithm (KENNEL et al., 1992), the nonlinear prediction method (FARMER & SIDOROWICH, 1987), and recurrence quantification analysis (ZBILUT & WEBBER, 1992). To avoid false results, it is typically common practice to use multiple techniques to distinguish between a stochastic and a chaotic process. Also, most of these methods require the reconstruction of the time series' phase space (TAKENS, 1981). The correlation dimension, Lyapunov exponent, and recurrence quantification analysis were applied to analyze the chaotic dynamics of rainfall and streamflow series. Each method focuses on distinct aspects of the series to identify chaos. For instance, recurrence plots can capture the nonlinear structure of the dynamics, while Lyapunov exponents can measure the rate of separation between close trajectories, indicating sensitivity to initial conditions. Fig. 3 illustrates the framework of the paper.

Figure 5 - Schematic representation of the proposed chaos identification framework.



Source: Prepared by the author

4.3.1 Phase space reconstruction (PSR)

Nonlinear dynamic system analysis frequently entails the reconstruction of the

phase space, which consists of a representation of the governing variables at a given moment. The time-delay embedding theorem (TAKENS, 1981) facilitated the reconstruction of a multidimensional phase space employing a single time series, and that led to several applications of chaos theory in hydrology (LIANG et al., 2019), particularly because all the actual governing variables of a hydrological system are often not known a priori (SHU et al., 2021; YAN et al., 2021; OMBADI et al., 2021). The phase space diagram tracks the evolution of the system from its initial conditions, and the overlap of the variables' trajectories provides insight into the complexity or variability of the system's behaviors. According to the Takens theorem (TAKENS, 1981), a one-dimensional time series of length n can be expressed in terms of phase space vectors as described in Equation 1.

$$\boldsymbol{Y}_j = \left(X_j, X_{j+\tau}, X_{j+2\tau}, \dots, X_{j+(m-1)\tau}\right), \quad j = 1, 2, \dots, N \qquad \text{(Eq. 01)}$$

where τ is the time delay, m is the embedding dimension, and $N = n - (m-1)$ (TAKENS, 1981). Even though the phase space is a powerful method to gain qualitative information regarding how the systems evolve, the choice of a proper τ is significant in representing the optimal separation between trajectories within the minimum embedding dimension (MA & HAN, 2006; LI et al., 2010). For example, if the value of the time delay is very small, there will be no independence among phase-space coordinates, resulting in a lack of information about the systems' evolution. On the other hand, if the time delay is too large, the trajectories will diverge, causing the loss of relevant information (DHANYA & KUMAR, 2010; LABAT et al., 2016).

4.3.1.1 Time delay selection

In the PSR process, two variables are required: the time delay (τ) and the embedding dimension (m). Several guidelines have been proposed to determine the appropriate time delay, many of which are based on series correlation. Among the well-known approaches for identifying the time delay are autocorrelation and Mutual Information Functions (MIF). However, relying solely on the first zero of the autocorrelation function to select τ may lead to inaccurate results because it only measures linear dependence between consecutive points (FRASER & SWINNEY, 1986).

The MIF is a measure that is not affected by the nonlinearity of the series, and it is

commonly used to analyze time series with inherent nonlinearity (STROZZI et al., 2002). The time delay is often chosen as the local minimum of the MIF (FRASER & SWINNEY, 1986). This choice allows for the assessment of both the linear dependence and general dependence between successive points (SIVAKUMAR, 2016). For consecutive values of a time series, X_i and $X_{i-\tau}$, the MIF, I_τ, can be expressed by Equation 2.

$$I_\tau = \sum_{j,j-\tau} P(X_j, X_{j-\tau}) \log_2 \left[\frac{P(X_j, X_{j-\tau})}{P(X_j), P(X_{j-\tau})} \right] \quad \text{(Eq. 02)}$$

where $P(X_i, X_{i-\tau})$ is the joint probability density, and $P(X_{j-\tau})$ and $P(X_j)$ are the probabilities of $X_{j-\tau}$ and X_j, respectively. The local minimum of the MIF was adopted in this study, as it can reflect that consecutive points are independent enough of each other but not so independent as to have no connection. Further details can be found in Fraser and Swinney.

4.3.1.2 Embedding dimension

The two most common approaches for establishing the minimum necessary embedding dimension to represent the system's dynamics are the false nearest neighbors (FNN) and the correlation dimension (CD). The FNN method (KENNEL et al., 1992) assumes that points on the attractor's trajectory are close to each other in the phase space, i.e., they are neighbors. If the embedding dimension (m) is increased and a previously detected nearest neighbor moves away from the vector $\boldsymbol{Y}_i$, it is considered a false nearest neighbor, as it reaches the neighborhood of $\boldsymbol{Y}_i$ in the higher embedding dimension by projecting itself from a distant region of the attractor (KHATIBI et al., 2012).

The minimum embedding dimension (m) is found when almost all the neighbors are true. In other words, the portion of points for which $[R_{m+1}(i) - R_m(i)] > \varepsilon R_m(i)$ is very small, zero, or below a 10% falseness of nearest neighbors (HUANG et al., 2010). ε is a heuristic threshold factor. The distance $R_{m+1}(i) - R_m$ is calculated to the same neighbors that were identified with embedding m, but with the $(m + 1)th$ coordinate.

For many real-world time series contaminated by noise, the percentage of FNN generally does not fall to zero. Also, the percentage of FNN may not remain at this minimum when the embedding dimension is increased. Instead, it often increases as the embedding dimension increases due to the effects of noise propagation to higher dimensions. A low

percentage of FNN aids in visualizing the geometric structure of the attractor as the system's orbits are now separated and do not cross (SIVAKUMAR, 2016).

The correlation dimension (CD) is a measure of the relationship between the positions of points in the reconstructed phase space, and it is a popular approach for identifying chaotic behavior (GHORBANI et al., 2018). In the context of identifying the presence of chaos, dynamics governed by chaotic processes have a finite value of dimensions, while those governed by stochastic dynamics have an infinite value of dimensions (KHATIBI et al., 2012; SIVAKUMAR, 2016).

The CD, proposed by Grassberger and Procaccia (1983a), applies the correlation integral to distinguish between stochastic and chaotic systems. The correlation integral measures the probability that two randomly selected points will be within a certain radius (r) of each other. The correlation function, $C(r)$ is expressed by Equation 3 (GRASSBERGER & PROCACCIA, 1983a; THEILER, 1986).

$$C(r) = \lim_{n\to\infty} \frac{2}{N(N-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \boldsymbol{H}(r - |Y_i - Y_j|) \quad \text{(Eq. 03)}$$

N is the number of points on the reconstructed attractor, $\boldsymbol{H}()$ is the Heaviside function with $\boldsymbol{H}(u) = 0$ for $u \leq 0$ and $\boldsymbol{H}(u) = 1$ for $u \geq 0$. The variable u is defined as $u = r - |Y_i - Y_j|$, where r is the radius of the sphere centred on $Y_i - Y_j$. Therefore, the correlation integral approximates the number of points that are closer than a radius r in the data (LABAT et al., 2016). For chaotic time series, the correlation function scales with r as $C(r) \propto r^{D_2}$, where D_2 is termed the correlation exponent. In contrast, for stochastic time series, $C(r) \propto r^m$ is true. The correlation exponent can be expressed by Equation 4.

$$D_2 = \lim_{r\to 0} \lim_{n\to\infty} \frac{logC(r)}{log(r)} \quad \text{(Eq. 04)}$$

The correlation exponent, which measures the degree of nonlinear interdependence among points on the attractor, can be approximated by finding the slope of $\log C(r)$ versus $\log(r)$ plot. If the correlation exponent increases with increasing m, it suggests the absence of chaotic behavior. However, if the correlation exponent reaches a constant value despite the increase in m, it indicates the presence of chaotic behavior. The dimension at which the curve saturates can be taken as the optimal m (LIANG et al., 2019). The nearest integer to the

calculated correlation exponent is commonly considered the number of degrees of freedom of the system, thus providing significant insights for the modeling process (GHORBANI et al., 2018). Although the CD method has been widely used, some remarks have been raised concerning its application to real-world time series, particularly for noisy data, short series, and data with the presence of zeros. However, Sivakumar et al. (2002) described that the CD can still be a reliable indicator of low-dimensional chaos in hydrological time series with limited data.

4.3.2 Lyapunov Exponent

The Lyapunov exponent (LE) is a popular method for identifying the presence of chaos in a time series. This method determines whether the phase trajectory has any diffusion motion features. The LE measures the average exponential rate of separation of nearby orbits in the phase space, serving as an indicator of the sensitivity of the attractor to initial conditions. To be characterized as a chaotic series, the presence of a positive Lyapunov exponent is required (KHATIBI et al., 2012). Among the existing methods to estimate the LE of a time series, the approach proposed by Wolf et al. (1985) and later complemented by Rosenstein et al. (1993) and by Kantz (1994) is extensively applied. This method involves computing the distance between the reference point (Y_i) and the nearest neighbor ($Y_{\hat{\imath}}$) of different trajectories in the reconstructed phase space. The distance between them is $d_i(0) = \left|\left|Y_i - Y_{\hat{\imath}}\right|\right|$. However, the nearest neighbor is not necessarily the closest one in the time domain. The distance after j discrete time steps is computed in Equation 5.

$$d_i(i) = \left|\left|Y_{i+j} - Y_{\widehat{i+j}}\right|\right| \quad \text{(Eq. 05)}$$

The evolution of this separation is observed, and the divergence between the trajectories based on the initial separation, $d_i(0)$, can be described by the exponential function $d_i(j) \cong d_i(0)e^{\lambda(j \cdot \Delta t)}$, where Δt is the considered period, and λ is the largest LE. Therefore, the largest Lyapunov exponent (LLE) can easily be achieved by calculating the slope of the curve in the exponential plot (ROSENSTEIN et al., 1993; JIANG et al., 2020).

4.3.3 Recurrence Plot (RP) and Recurrence Quantification Analysis (RQA)

In deterministic systems, the recurrence of a state implies that if any deviation occurs during the trajectory, that state will be returned in the future (Shu et al. 2021). The

recurrence plot (RP) was first introduced by Eckmann et al. (1987), and it is applied to visualize the recurring patterns in dynamic systems and their trajectories (MARWAN et al., 2007). This well-establish nonlinear analysis technique illustrates how likely a recurrent behavior observed in m-dimensional phase space is to recur. Further, this method is known for detecting the dynamics of non-stationary or relatively short data (SHU et al., 2021). A system's dynamics are characterized by several vectors, which represent the trajectory of the attractor in a mathematical space. The RP is established on the recurrence matrix, which can be expressed by Equation 6.

$$\boldsymbol{R}_{i,j}(\varepsilon) = \boldsymbol{H}\left(\varepsilon - \left|\left|\vec{x}_i - \vec{x}_j\right|\right|\right), \quad i,j = 1,\dots,N. \tag{Eq. 06}$$

where N is the number of measured points $\vec{x}$, ε is a threshold distance, and $\boldsymbol{H()}$ is a Heaviside function. The recurrence of state $\vec{x}_i$ at time j is represented by ones (graphically depicted by a black dot), while zeros reflect how the states are different at times i and j (MARWAN et al., 2002). The corresponding matrix compares the systems' states at times i and j, with $\boldsymbol{R}_{i,j} = 1$ if the states are similar and $\boldsymbol{R}_{i,j} = 0$ if the states are rather different. The matrix also illustrates whether similar states occur in the underlying system (MARWAN, 2011). Several rules for selecting the threshold ε have been proposed in previous studies. In this paper, we apply the recommendation by Marwan et al. (2002) to use 10% of the maximum phase space diameter.

The RP is a graphical representation of a system's dynamics that can reveal different patterns and behaviors of the system. Marwan et al. (2002) have described the different meanings that the patterns in the RP can contain. For example, periodic processes are generally characterized by diagonal lines or checkboard patterns. Single points can indicate a short persistence in the state of system dynamics, whereas short diagonal lines mean that the trajectory in phase space recurs to the same point, but at different times. This suggests the possible presence of deterministic characteristics. Further explanation of RP can be found in Marwan et al. (2002).

Recurrence Quantification Analysis (RQA) is a heuristic method of quantifying RPs based on their patterns and line structures (MARWAN et al., 2002). Although RPs are widely used, they are a visual tool that can provide only a qualitative view of the dynamic system characteristics (SHU et al., 2021). To overcome this aspect, RQA proposes a quantitative assessment to describe the systems' underlying dynamics (ZBILUT et al., 2002). RQA measures, such as determinism (DET), maximal length of diagonal structures (LMAX),

recurrence rate (RR), and entropy (ENT) can identify and diagnose small-scale diagonal structures in RP, which indicate the presence of chaos. The RR is described by Equation 7.

$$RR = \frac{1}{N^2}\sum_{j,i=1}^{N} R_{ji} \quad \text{(Eq. 07)}$$

RR measures the portion of recurrence points with regard to the total number of possible recurrences (MOCENNI et al. 2011). Normally, high recurrence indicates lower system variability. The entropy (ENT) is described by Equation 8.

$$ENT = -\sum_{L=L_{min}}^{N} P(L) \log P(L) \quad \text{(Eq. 08)}$$

where L is the length of the diagonal line structure, and $P(L)$ is the frequency distribution of the diagonal line lengths. ENT indicates the complexity of the RP regarding the diagonal lines; for example, small values of ENT reflect low complexity. A diagonal line of length l means that two segments of the trajectory, at different times, are rather close during l time steps. Hence, these lines are related to the divergence of the trajectory. The maximal length of diagonal structures (LMAX) is described by Equation 09.

$$L_{MAX} = \max\left(\{L_j, j = 1, \dots, N\}\right) \quad \text{(Eq. 09)}$$

where L_j is the length of the j-th diagonal line in the RP. Marwan et al. (2002) observed that deterministic dynamics result in longer diagonals and fewer isolated recurrence points, while dynamics with uncorrelated or weakly correlated, stochastic, or chaotic behavior produce none or very short diagonals. Therefore, one can use the ratio of recurrence points that form diagonal structures (of at least length L_{min}) to all recurrence points as a measure of determinism of the system. DET is described by Equation 10.

to all recurrence points as a measure of determinism of the system. DET is described by Equation 10.

$$DET = \frac{\sum_{L=L_{min}}^{N} LP(L)}{\sum_{L=1}^{N} LP(L)} \quad \text{(Eq. 10)}$$

The threshold L_{min} eliminates diagonal lines which are formed by the tangential motion of the phase space trajectory. High values of DET are an indication of determinism in the system, however, they are not sufficient to confirm that the system is deterministic. For

example, if a system has DET ≈ 1 for minor recurrence density (i.e., RR < 0.1), the underlying process will be deterministic (i.e., a periodic or chaotic system).

4.4 Results and Discussion

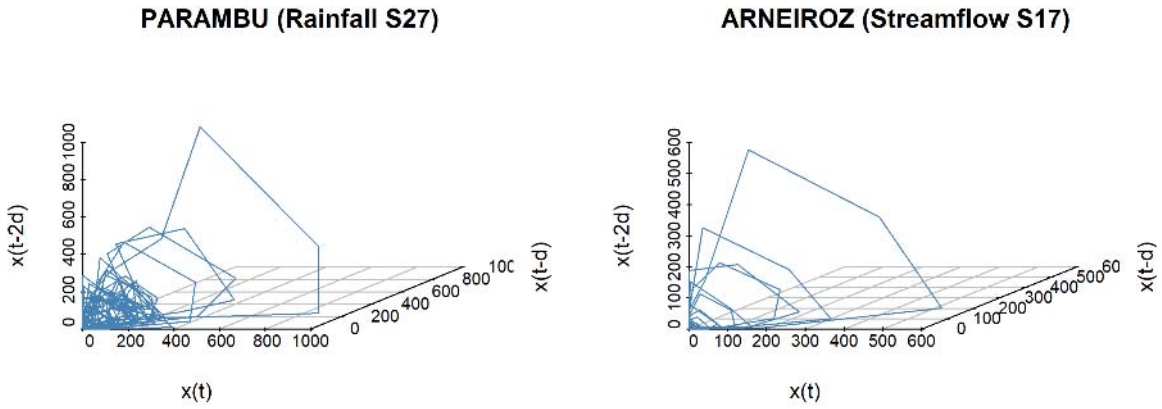
4.4.1 Phase-space reconstruction (PSR)

This paper seeks to investigate the existence and inherent nature of chaos in hydrological processes, such as rainfall and streamflow. It also explores how data aggregation affects the detection of the deterministic chaos in the system's dynamics, as well as the spatial distributions of chaotic time series and the relationship between these two hydrological variables. The identification of deterministic chaos in the data is key for comprehending the behavior of hydrological variables, and PSR is the first step for applying several methods in chaos theory. PSR can graphically represent the variables needed to describe the system's state and can characterize the entire dynamics of the process with a single time series (SHU et al., 2021; YAN et al., 2021).

The phase space for monthly rainfall and streamflow stations, geographically close to each other, were reconstructed with a time delay equal to one and an embedding dimension equal to three (see Figure 6), using time series from 1976 to 2006. Fig 4 shows a region of attraction for these trajectories in the phase space, which provides possible evidence of attractors. Although there is no clearly defined pattern, the trajectories are not dispersed throughout the phase space, indicating an intermediate level of complexity between deterministic and stochastic processes. Therefore, there is a need to characterize rainfall and streamflow dynamics using chaos theory.

Figure 6 - Attractor of reconstructed phase space for monthly rainfall (left) and streamflow (right) time series. The phase space was reconstructed with an embedding dimension of three

and a one-month time delay.

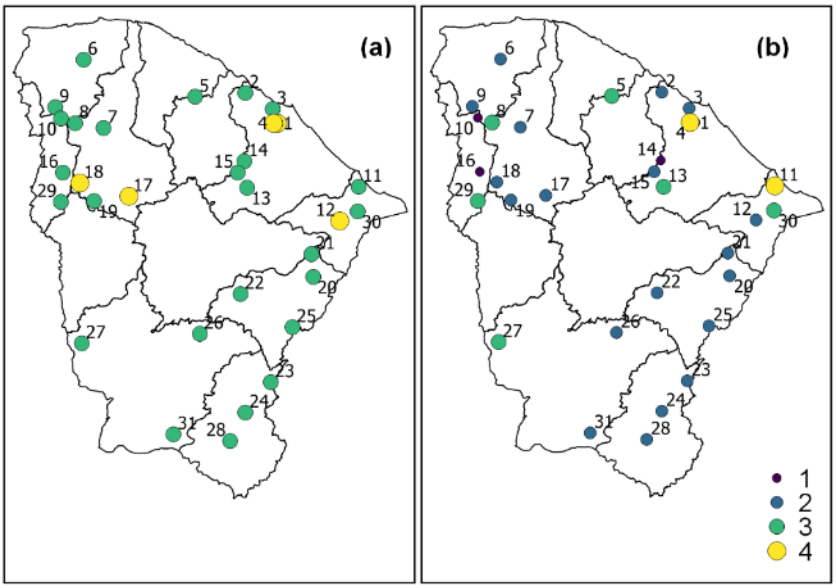


Source: Prepared by the author

Two parameters are needed for the PSR: the time delay (τ) and the embedding dimension (m). These can be determined using methods such as the MIF, the FNN, and the CD. The time delay is determined as the first minimum of the MIF. Figure 7 shows the results for time delay identification using the MIF for the rainfall stations on both monthly and seasonal timescales. The time delay varies among the stations and the timescales, except for station #27, which had the same numerical value of τ across all timescales, including the annual timescale (Figure 9). On the monthly timescale, τ ranges from three to four months, coinciding with the State's rainy season. In the seasonal analysis, which considered data from January to June, τ ranges from one to four seasons, with higher values mainly concentrated in the State's northeastern area.

Figure 7 - The spatial distribution of time delay, determined by mutual information, is examined for rainfall at two different timescales: (a) monthly timescale in months and (b) seasonal

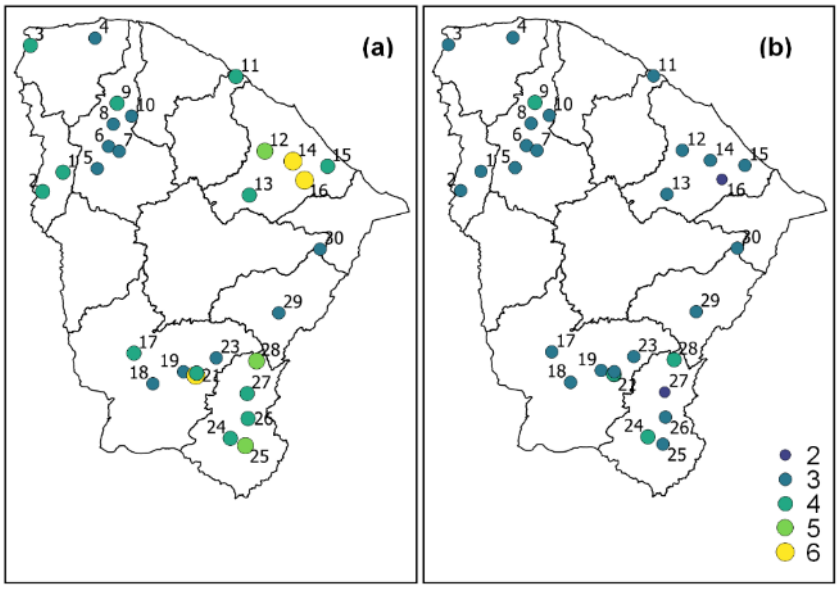
timescale in terms of the number of seasons.



Source: Prepared by the author

For streamflow, the time delay ranged from two to four seasons on the seasonal timescale. The stations with a τ of four were mainly concentrated in the northwestern and southeastern parts of the State (Figure 8). On a monthly timescale, the time delay varies from three to six months. The southeastern and northern areas present the largest time delay for the monthly time series. Further, the results show that a higher fluctuation of variability is observed at the seasonal scale for rainfall compared to that on the monthly timescale, mainly due to rainfall variability throughout the State. While for streamflow, most of the rivers are intermittent. Thus, the time delay does not vary much throughout the region.

Figure 8 - The spatial distribution of time delay, determined by mutual information, is examined for streamflow at two different timescales: (a) monthly timescale in months and (b) seasonal timescale in terms of the number of seasons.

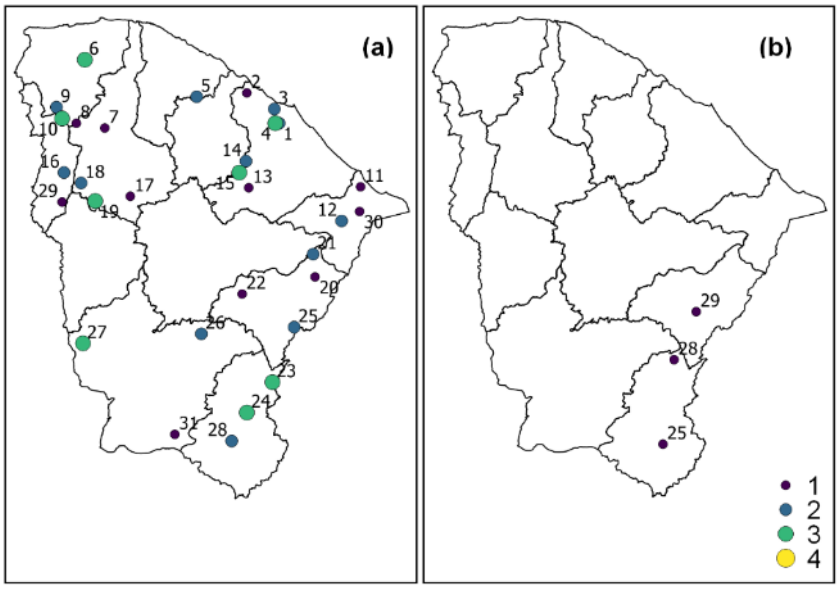


Source: Prepared by the author

In the annual analysis (Figure 9), the time delay for rainfall ranges from one to three years, indicating that rainfall exhibits interannual variability. Due to missing values, the annual analysis of the streamflow series was only applied to series with over 30 years of data. These series correspond to stations #25, #28, and #29, located in the southeastern region of the State. On an annual timescale, the time delay was one for all stations. In comparison between the time delays of rainfall and streamflow at stations in close geographical proximity, it is observed that the streamflow value is typically higher. The interdependency between rainfall and streamflow, which varies across all analyzed timescales, is often associated with changes in land cover, climate, human water usage, or alterations in the characteristics of rainfall events themselves, which are major drivers of streamflow. However, stations located near the central region of the State present the same time delay for both variables.

Figure 9 - Spatial distribution of time delay (years) determined by mutual information for (a)

rainfall and (b) streamflow at annual timescale.



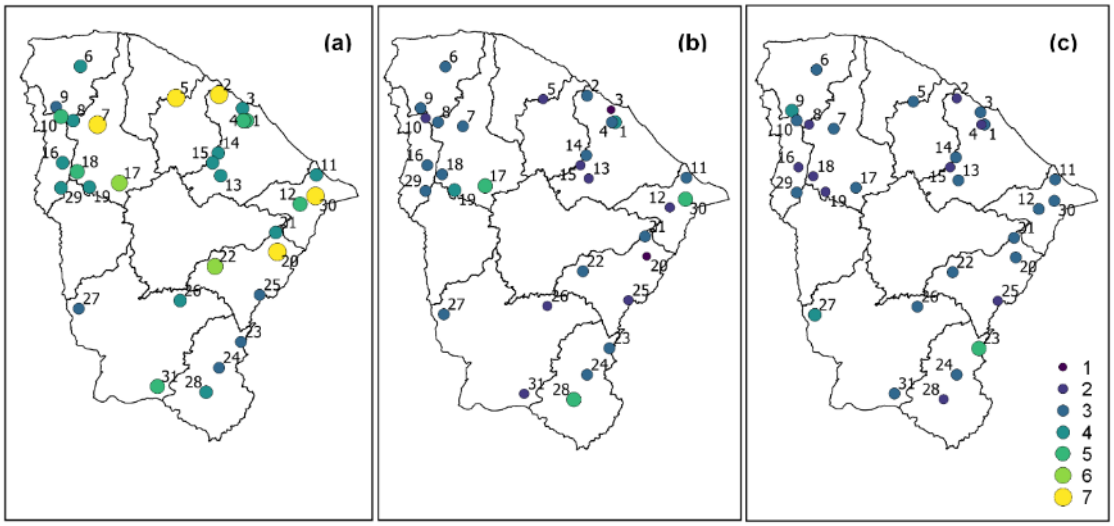
Source: Prepared by the author

When analyzing the embedding dimension using the FNN method for rainfall (Figure 10) on a monthly timescale, the number of dimensions ranges from three to seven, with higher embedding dimensions observed in the northern and eastern regions of the State. When analyzing only the seasonal timescale, the number of dimensions decreases by at least one for most stations. Furthermore, 32% of the stations maintained the same dimension value when changing from seasonal to annual timescales, with most of these stations situated in the central part of the State where rainfall is scarce. The optimal number of dimensions can be defined as the number of variables that effectively describe the dynamics of a given variable. As the timescales become coarser, the complexity of rainfall decreases (i.e., a decrease in the embedding dimension), and higher complexity is seen in the northern part of the State, where rainfall is more abundant.

On a monthly timescale, streamflow analysis (Figure 11) showed a range of three to 15 dimensions, with higher dimensions concentrated in the western region of the State. When shifting the timescale to a seasonal analysis, there is a decrease in the number of dimensions as previously observed for rainfall. Only 13% of the stations showed the same number of dimensions with the change in timescales, likely due to the many factors that influence streamflow compared to rainfall. Notably, the streamflow series showed higher embedding dimensions than the rainfall series, which are primarily determined by meteorological factors such as temperature, air pressure, humidity, and topography. In contrast, streamflow is affected

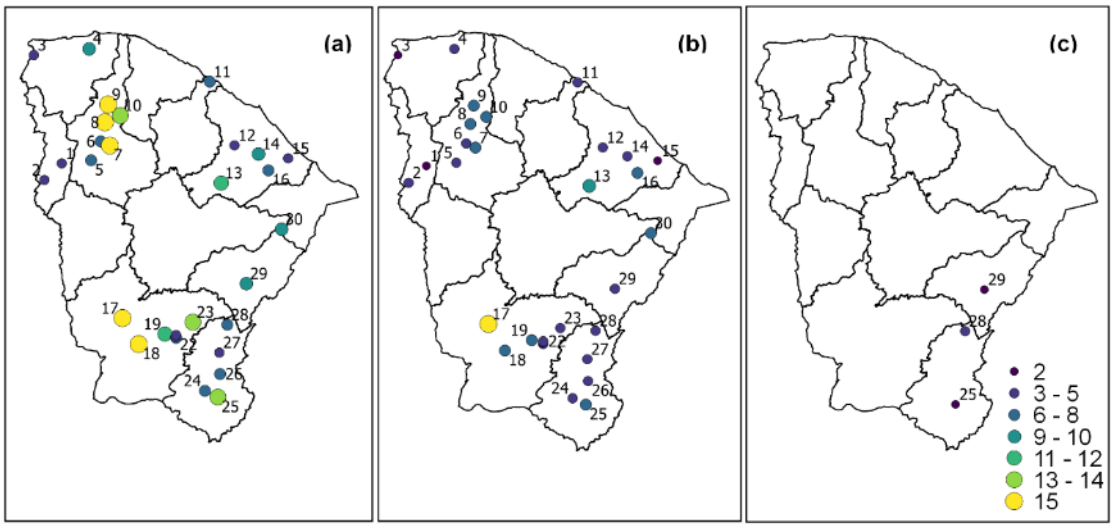
by a variety of factors, including rainfall, evaporation, infiltration, soil moisture, and land use changes such as deforestation or urbanization, which alter the water input/output.

Figure 10 - Spatial distribution of embedding dimension determined by the FNN method for rainfall at (a) monthly, (b) seasonal and (c) annual timescales.



Source: Prepared by the author

Figure 11 - Spatial distribution of embedding dimension determined by the FNN method for streamflow at (a) monthly, (b) seasonal and (c) annual timescales.



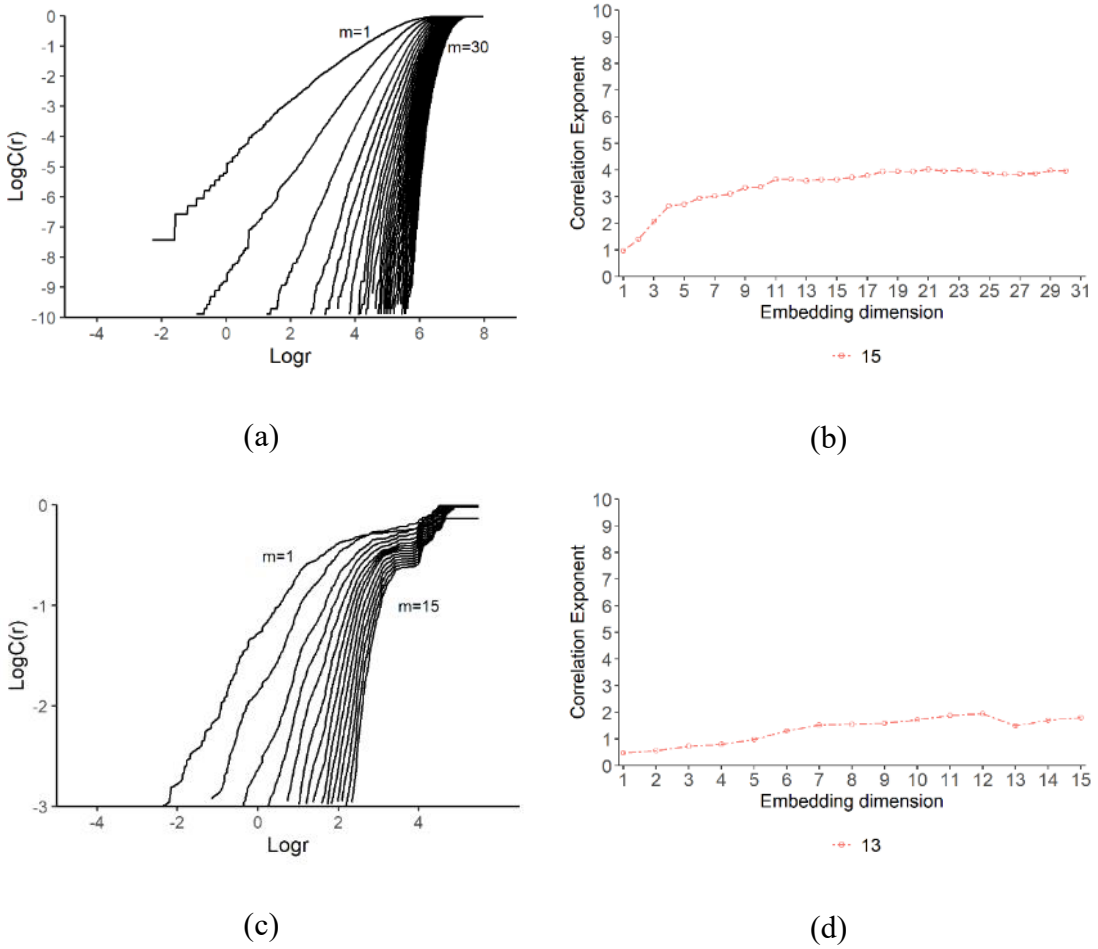
Source: Prepared by the author

The CD with an increasing dimension plot was also applied to determine the optimal embedding dimension. In many practical applications, the point on the plot where the correlation dimension saturates is often chosen as the optimal dimension (LABAT et al., 2016;

FUWAPE et al., 2017; LIANG et al., 2019; JIANG et al., 2020). The relationship between InCm(r)~Inr was was obtained for the monthly timescale, with the value of m ranging from 1 to 30 for all stations (see Figure 12a for an example). Figure 12b exhibits the correlation function for station #15 (rainfall) with an increasing embedding dimension. The plot suggests the presence of a chaotic process as the correlation function saturates around m=11. Figure 12c and d illustrate the same relations for streamflow station #13, which is geographically close to the analyzed rainfall station. The correlation function reaches saturation around m=8 for streamflow. Table 2 shows the chosen embedding dimension for the 31 rainfall and 30 streamflow stations at all the analyzed timescales based on their saturation point. 78% of the stations at monthly, 68% at the seasonal, and 26% of the annual rainfall stations presented a saturation point. 83% and 76% of the monthly and seasonal streamflow stations, respectively, showed a saturation point. Most stations showed a decrease in the embedding dimension with increasing timescale. However, rainfall station #3 maintained the same embedding dimension for seasonal and monthly timescales. Some stations showed an increase in m with the increase of the timescale (e.g., stations #6, #14, #21, and #28). Although it is generally believed that temporal aggregation results in some form of smoothing and hence, less complex (and more predictable) behavior, previous results, as discussed in Sivakumar (2016), also found an increase in the dimensionality (or complexity) of the flow dynamics with the scale of aggregation.

The nearest integer to the calculated correlation exponent is often adopted as the number of degrees of freedom of the system. In general, a correlation dimension of 1 indicates a periodic system, whereas a quasi-periodic system is characterized by a correlation dimension of 2 (REITERER et al., 1998). However, a non-integer correlation dimension value suggests the presence of chaos in the system. For monthly rainfall, the correlation exponent values range from 3.47 to 5.67, confirming the presence of deterministic chaos in time series with non-integer values. The results indicate that the dynamics of monthly rainfall require between four to six equations to describe them, which is consistent with the findings of Sivakumar et al. (2014), with values ranging from 4.82 to 8.87. For streamflow, the correlation exponent values range from 0.52 to 3.64, indicating that the dynamics require one to four equations. These results are consistent with those found in other studies for streamflow, such as 0.46 (JAYAWARDENA & LAI, 1994), 2.4 (KHATIBI et al., 2012), and 0.9–2.3 (BABOVIC & KEIJZER, 2000).

Figure 12 - Log–log plot of the correlation integral function for (a) rainfall station #15 and (c) streamflow station #13. Correlation exponent versus embedding dimension for (b) rainfall station #15 and (d) streamflow station #13.



Source: Prepared by the author

Table 2 – Selected dimension for which the correlation exponent versus embedding dimension plot saturates for all the analyzed stations and timescales.

Rainfall station number	Monthly	Seasonal	Annual	Streamflow station number	Monthly	Seasonal
1	19	-	-	1	-	-
2	12	-	10	2	-	5
3	15	15	-	3	14	-
4	-	-	-	4	8	4
5	13	12	8	5	12	6
6	12	14	-	6	12	4
7	23	18	-	7	-	-
8	21	-	11	8	-	6

9	19	-	-	9	2	2
10	20	17	-	10	9	5
11	-	-	-	11	2	3
12	-	-	-	12	2	2
13	-	-	-	13	8	5
14	13	16	-	14	9	-
15	11	5	-	15	9	6
16	25	16	-	16	9	6
17	22	18	-	17	4	5
18	22	17	-	18	12	5
19	-	18	-	19	13	6
20	21	17	-	20	10	-
21	15	18	-	21	10	6
22	19	-	-	22	11	6
23	25	16	-	23	9	-
24	19	9	-	24	11	-
25	-	16	12	25	12	4
26	25	16	-	26	10	6
27	25	11	-	27	5	4
28	7	10	-	28	-	6
29	-	18	13	29	8	4
30	20	6	13	30	12	3
31	25	-	11			

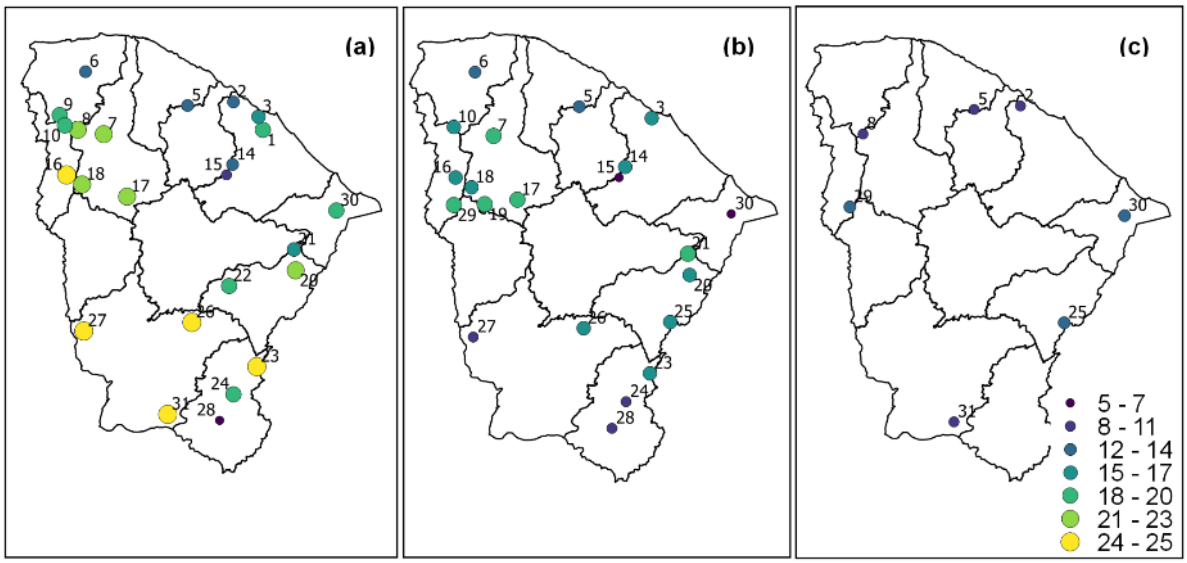
Source: Prepared by the author

Analyzing the spatial distribution of the embedding dimension chosen by the CD method for rainfall stations (Figure 13), it was observed that the southern and western regions of the State had higher embedding dimensions for the monthly analysis, in contrast to the coastal region of the State, which experiences higher rainfall rates. The seasonal analysis showed similar behavior but with smaller embedding dimensions. Furthermore, only two stations (Stations #5 and #30) showed a saturation point across all analyzed timescales. The monthly and seasonal timescales (Figure 13) show that the embedding dimension appears to form three clusters in the State: one in the coastal area, one in the central part, and one in the south.

In the streamflow analysis, higher values of m were found in the southern and northwestern regions of the State (Figure 14). As seen in the rainfall analysis, there was a reduction in the embedding dimension when moving from the monthly to seasonal timescales, as well as with the FNN method, suggesting that larger timescales require fewer dimensions. The clusters formed by the streamflow time series slightly differ from those observed in rainfall, with the cluster in the coastal part of the State being divided into a northwestern and a central part. Although Vignesh et al. (2015) stated that the FNN dimension could serve as a measure of similarity for identifying stations with similar levels of variability in time series dynamics,

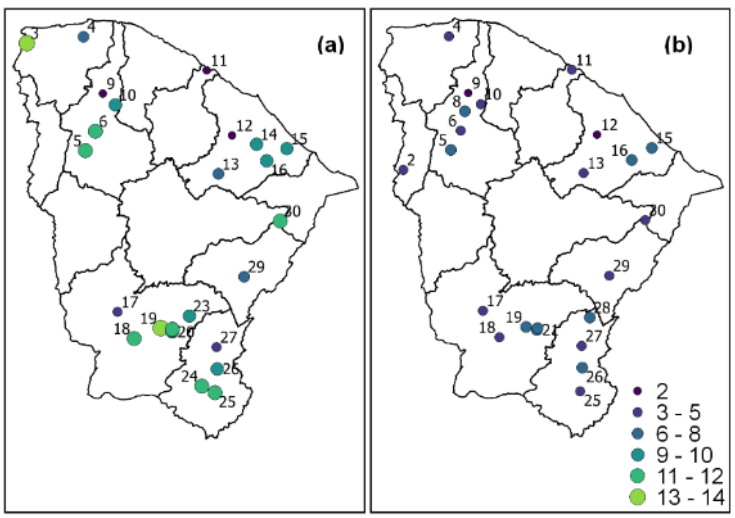
no consistent pattern was observed in this study with this method. This inconsistency may be associated with the size of the analyzed area. When comparing the CD and FNN methods to estimate the optimum embedding dimension, it was found that the CD method yielded higher values for rainfall, while the FNN method produced lower values. The CD method also showed a different pattern compared to the FNN method, with higher values of the embedding dimension found in the northern part of the State. Both methods demonstrated similar results for streamflow, with higher embedding dimensions observed in the western region of the State.

Figure 13 – Spatial distribution of the chosen dimension of rainfall for which the correlation exponent versus embedding dimension plot saturates for (a) monthly, (b) seasonal and (c) annual timescale.



Source: Prepared by the author

Figure 14 – Spatial distribution of the chosen dimension of streamflow for which the correlation exponent versus embedding dimension plot saturates for (a) monthly and (b) seasonal timescales.

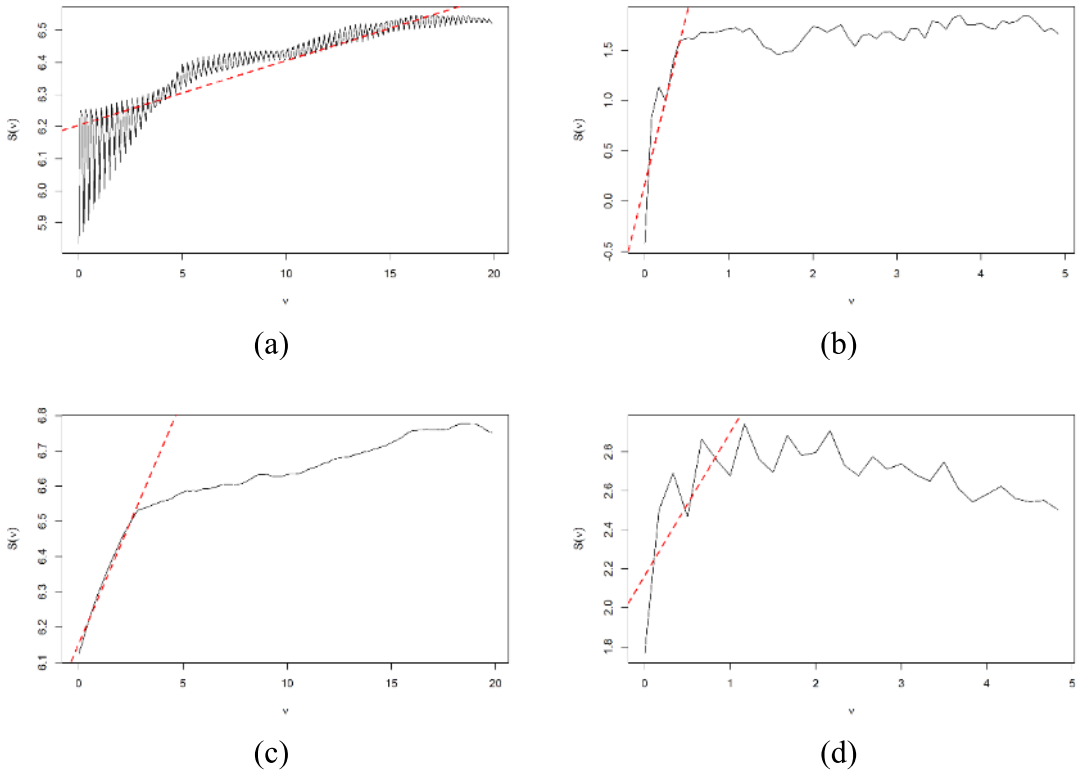


Source: Prepared by the author

4.4.2 Estimation of Lyapunov exponent

The Lyapunov exponent plots are exponential graphs that show the divergence of the data over time. If the separation between two points grows exponentially, the plot will appear as a straight line. However, sometimes two points move away from each other more closely together. Thus, the plot will present a straight line with oscillations. Figure 15 illustrates a rainfall (#10) and a streamflow (#6) station at monthly and seasonal timescales.

Figure 15 - Evolution of the logarithm of the mean distance S(ν) as a function of the time step ν. The red dashed line is the regression line. Rainfall station (#10) (a) at monthly and (c) seasonal timescales; and streamflow station (#6) (b) at monthly and (d) seasonal timescales.



Source: Prepared by the author

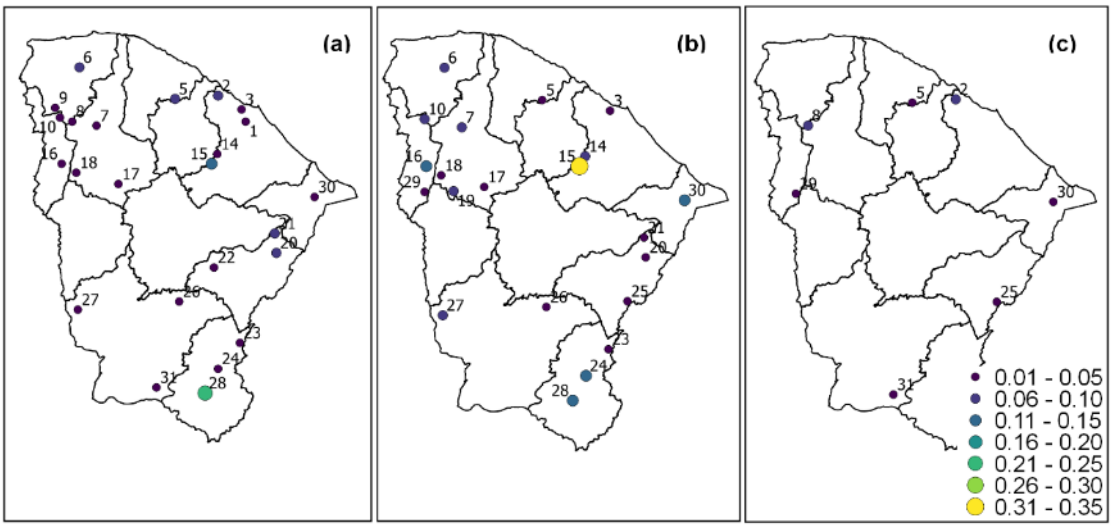
The largest Lyapunov exponent (LLE) was calculated only for stations that presented a saturation point in the CD analysis. For monthly rainfall, positive values of LLE ranged from 0.02 to 0.24, indicating a signature of chaos (Figure 16). At the seasonal timescale, the LLE values were similar, ranging from 0.13 to 0.32, while at the annual timescale, they increased, ranging from 0.06 to 0.5. The LLE for the monthly timescale is a bit smaller than the

values found by Echi et al. (2013), who found a value of 0.0632 for the daily rainfall time series in Nigeria. Fuwape et al. (2017) found LLE values ranging from -10 to 2 for daily rainfall over Nigeria. Falayi et al. (2022) analyzed monthly rainfall data of West African stations and observed positive LLE ranging from 0.13 to 0.36. The higher LLE values for rainfall were concentrated in the northern part of the State (Figure 16), where the embedding dimensions were smaller according to the CD method.

The LLE values for streamflow showed negative values for some stations (e.g., Station #5, #14, and #17 for the monthly timescale, and Station #5 at the seasonal timescale), indicating periodic orbit (Figure 17). Also, the values of LLE show a wider range from -0.32 to 3.4 and -0.04 to 11 for monthly and seasonal streamflow, respectively. The higher LLE values for streamflow were in the northwestern part of the State. Zhou et al. (2018) found an LLE of 4.142 for a monthly streamflow time series in the Jinsha River Basin, China. Mihailović et al. (2019) found values that range from 0.018 to 0.39 for daily streamflow at the Brazos River in Texas (USA). Larger values of LLE indicate stronger chaos and greater sensitivity to initial conditions in a dynamical system. Notably, the monthly rainfall analysis shows similar LLE values across the whole State, while streamflow showed higher LLE compared to rainfall.

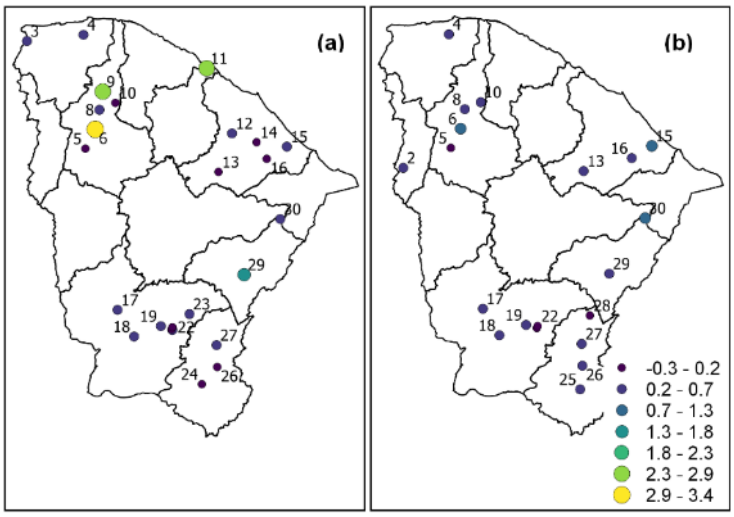
Alfaro et al. (2018) stated that the inverse of the LLE can be used to predict the boundaries of a chaotic time series. Based on this, an estimate of the forecast horizon for rainfall stations is in the range of 4 to 76 months into the future, while for streamflow, it ranges from 1 to 10 months. Ogunjo et al. (2022) found predictability of 40–58 days into the future for daily streamflow. Our results indicate that rainfall stations that show the presence of deterministic chaos are more predictable for longer periods compared to the streamflow in the same region.

Figure 16 - Spatial distribution of the largest Lyapunov exponent for the rainfall stations at (a) monthly, (b) seasonal and (c) annual timescale.



Source: Prepared by the author

Figure 17 - Spatial distribution of the largest Lyapunov exponent for the streamflow stations at (a) monthly and (b) seasonal timescale.



Source: Prepared by the author

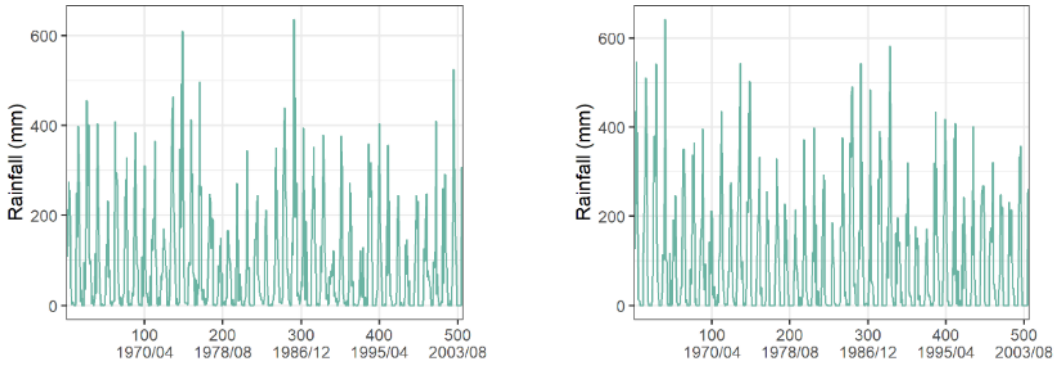
4.4.3 RQA

The presence of chaos in rainfall and streamflow dynamics is illustrated in the RP (Figure 18 and Figure 19). One advantage of using a recurrence plot (RP) is that it transforms a phase space of dimension m into a binary 2-dimensional space, with a threshold that represents the level of recurrence, making it easier to visualize any existing patterns. The figures illustrate

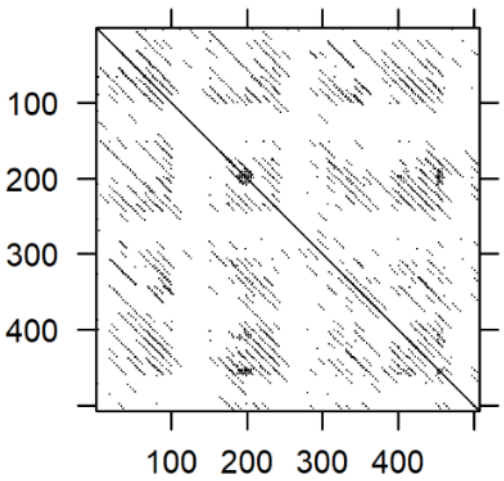
that the recurrence points do not form a periodic pattern and are not homogeneously distributed. The RPs of the stations across the State present different behaviors. For example, in Figure 18c, blocks with white stripes between them can be observed. Figure 18a, which illustrates the time series, shows a higher peak of rainfall in the period of the white stripes, indicating that the RP can be an accurate method to detect a transition in the dynamic structure. Also, for all the analyzed stations, the recurrence points are primarily concentrated in the low rainfall and streamflow periods, as noted by Shu et al. (2021), which is indicative of lower system variability and a persistent dynamic structure during dry spells.

In some of the other stations analyzed in this study, the same white ribbons as in Station #2 can be seen for the same period, such as in Stations #1 and #3 around 1971 and 1983. The cyclic pattern observed in Station #8 is also present in Stations #9, #10, #18, and #24. Other stations (i.e., Stations #7, #14, #23, #26, and #31) display a homogeneous RP with numerous single points and some short diagonal lines (compared to the length of the largest observed interval), which indicates a developed chaotic state.

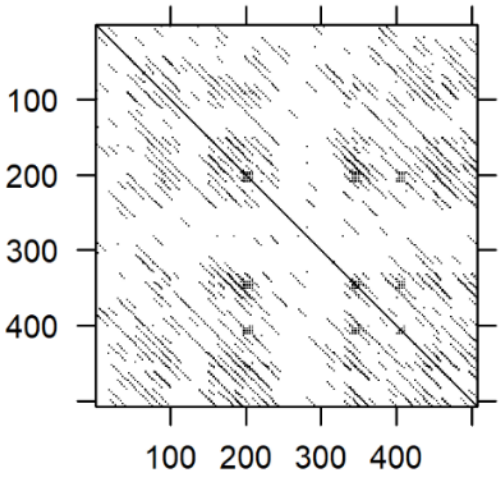
Figure 18 - Examples of time series and recurrence plot (RP) based on monthly rainfall time series at selected stations (a) Time series of station #2 and (c) its RP. (b) Time series of station #6 and (d) its RP. (e) Time series of station #8 and (g) its RP. (f) Time series of station #15 and (h) its RP. (i) Time series of station #27 and (k) its RP. (j) Time series of station #28 and (l) its RP.



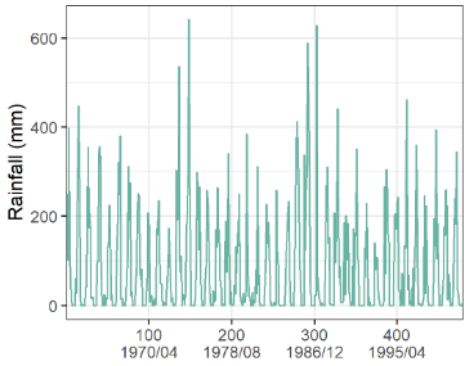
(a) (b)


100
200
300
400
100 200 300 400

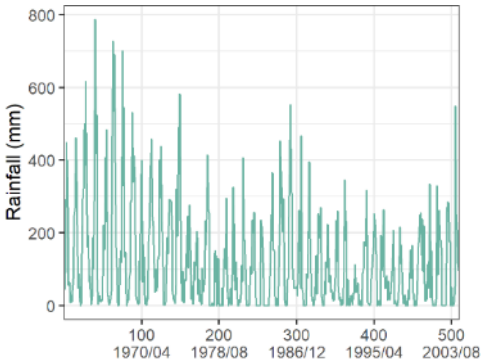
(c)


100
200
300
400
100 200 300 400

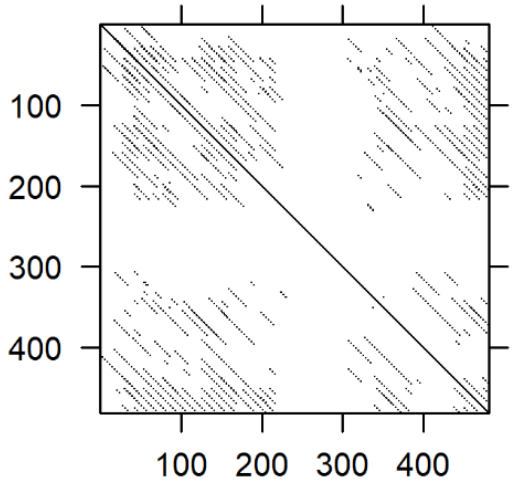
(d)


Rainfall (mm)
600
400
200
0
100
1970/04
200
1978/08
300
1986/12
400
1995/04

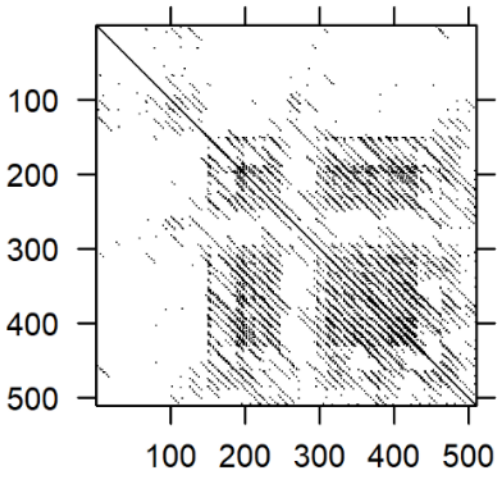
(e)


Rainfall (mm)
800
600
400
200
0
100
1970/04
200
1978/08
300
1986/12
400
1995/04
500
2003/08

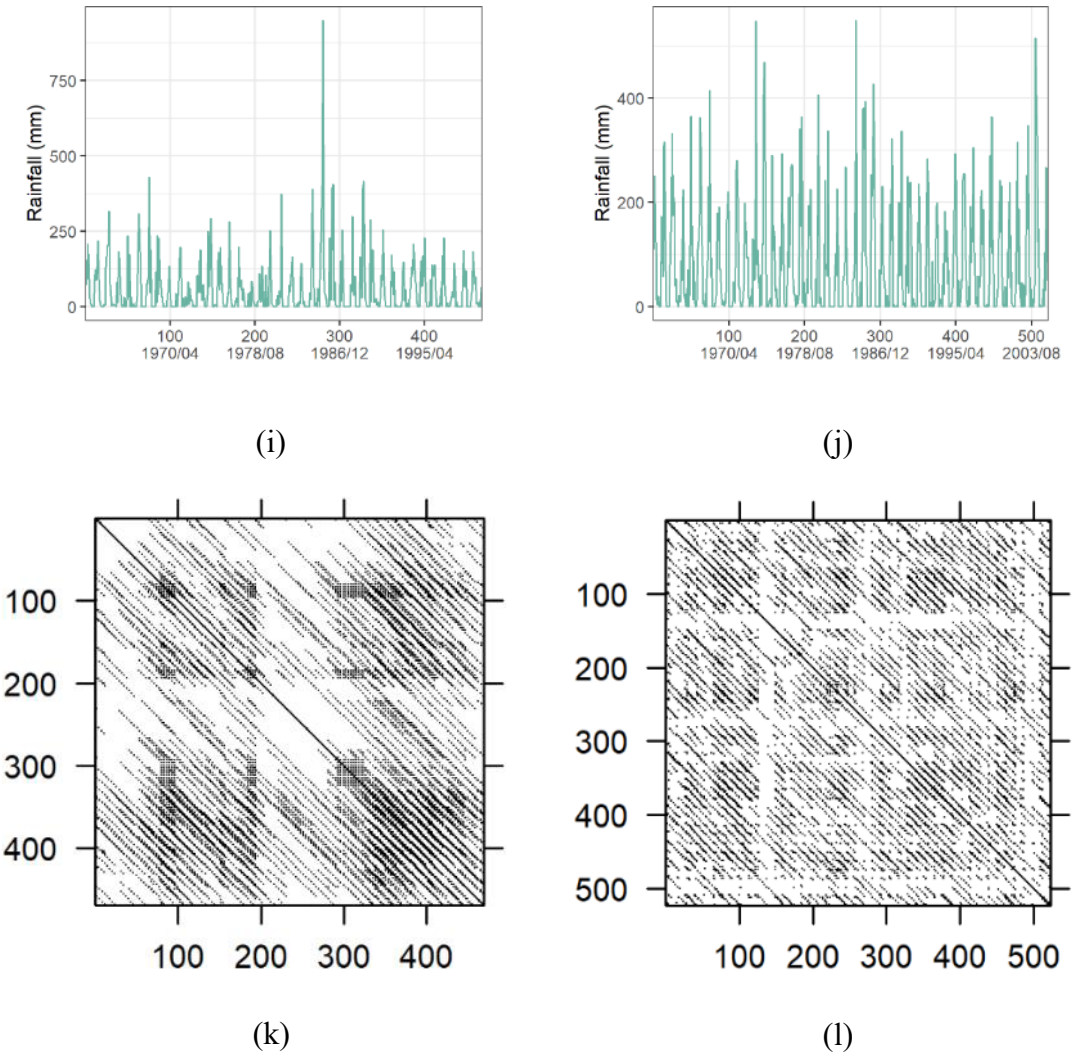
(f)


100
200
300
400
100 200 300 400

(g)


100
200
300
400
500
100 200 300 400 500

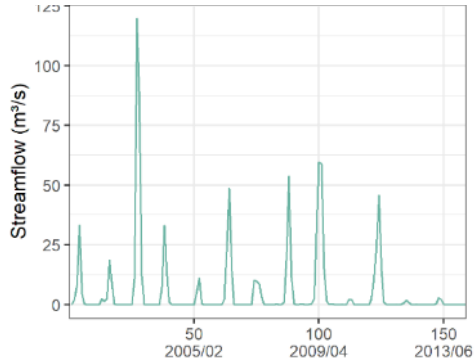
(h)



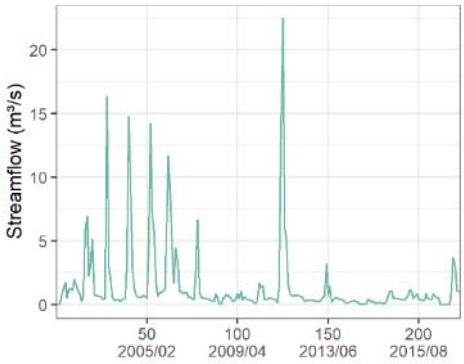
Source: Prepared by the author

The RPs for streamflow time series mostly consist of several rectangular blocks separated by horizontal and vertical white ribbons. It is noticeable that the distribution of these rectangular blocks and the density of recurrence points within them change. For example, Station #4 has only one such block from 2011, while Station #19 has a structure similar to that in the same period. Other stations, such as Stations #5 and #14, show the presence of only one block from 1968 and Station #16 from 1974. Some stations present multiple blocks from different periods. In the streamflow series, it is possible to observe a common change in dynamics between the analyzed stations. For example, Stations #6, #12, #18, #29, and #30 show a shift between 2000/2001. Stations #19, #29, and #30 show a shift in 2005, and Stations #23, #25, and #27 show a shift around 1981/1982.

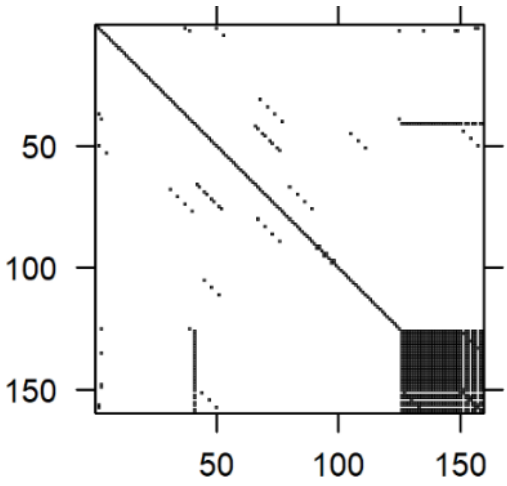
Figure 19 – Examples of time series and recurrence plot (RP) based on monthly streamflow time series at selected stations (a) Time series of station #4 and (c) its RP. (b) Time series of station #12 and (d) its RP. (e) Time series of station #18 and (g) its RP. (f) Time series of station #23 and (h) its RP. (i) Time series of station #25 and (k) its RP. (j) Time series of station #29 and (l) its RP.



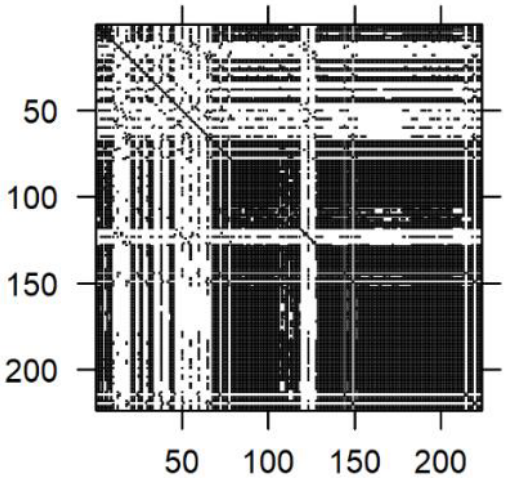
(a)



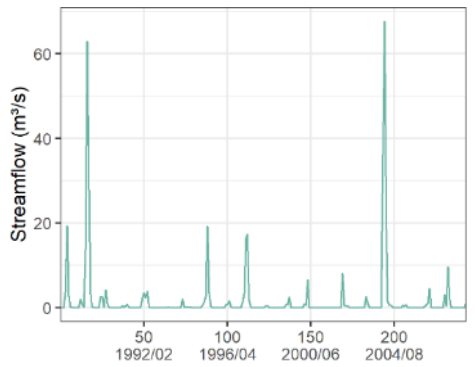
(b)



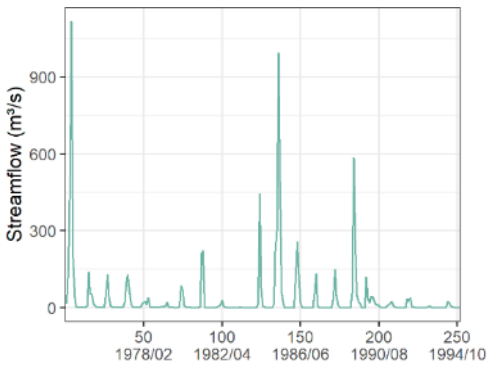
(c)



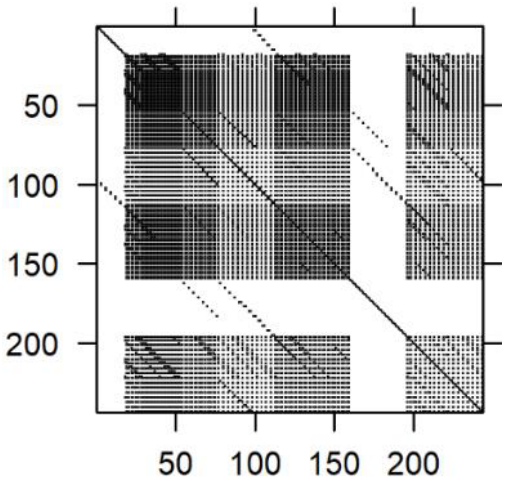
(d)



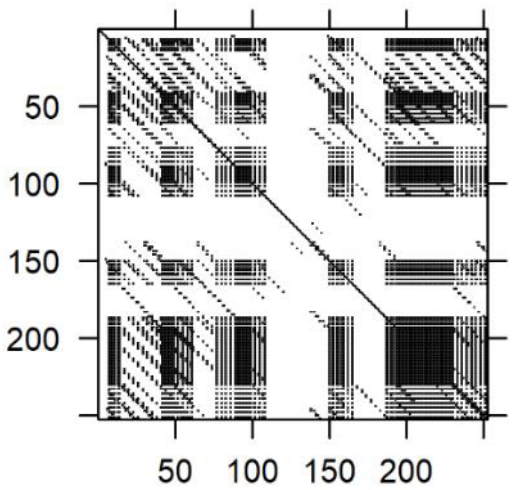
(e)



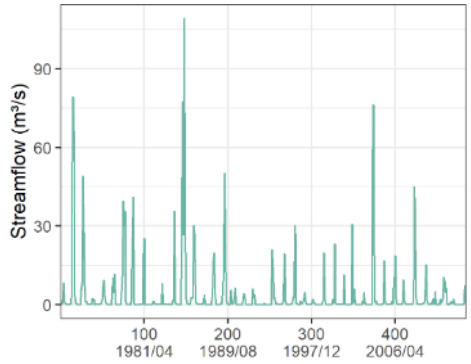
(f)



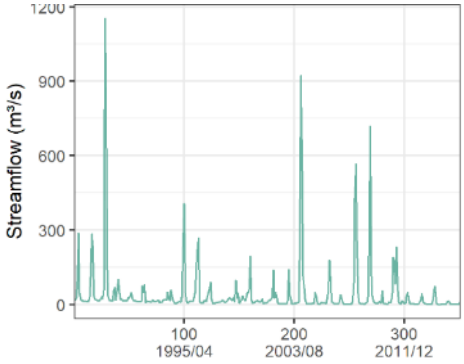
(g)



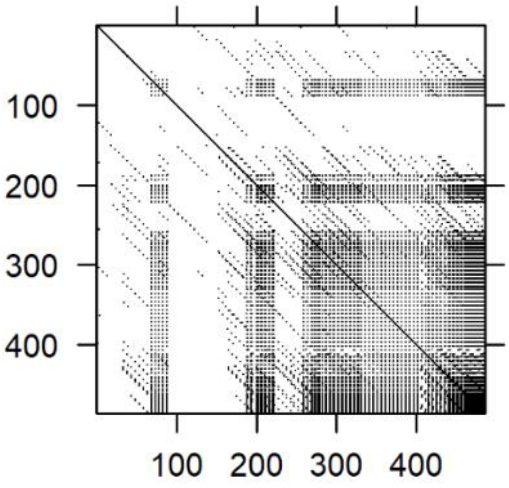
(h)



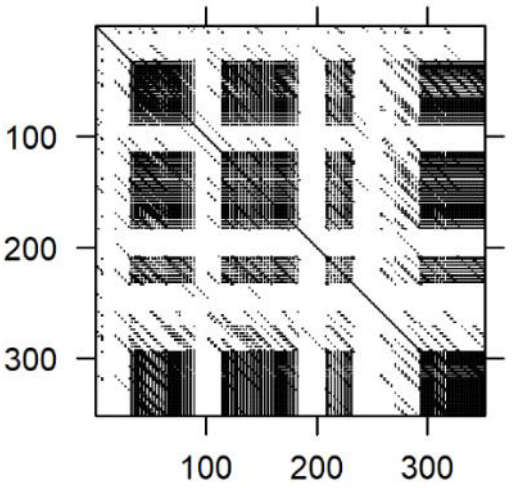
(i)



(j)



(k)



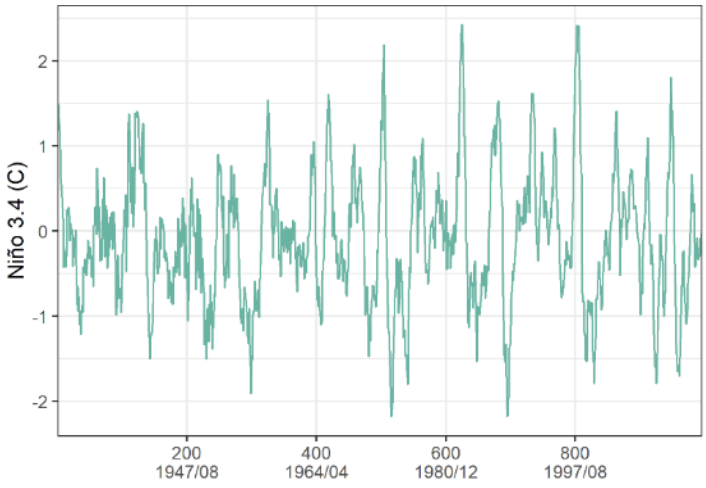
(l)

Source: Prepared by the author

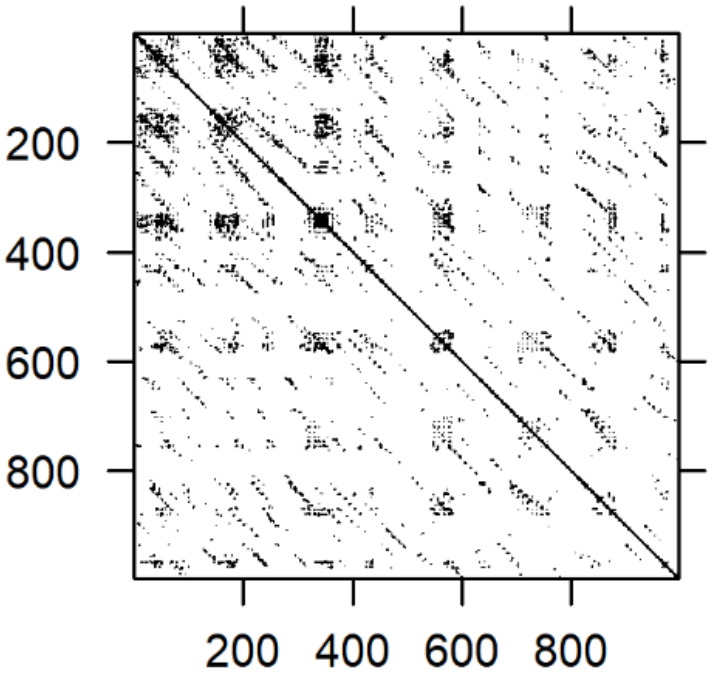
Climate, as discussed, is a main driver of the hydrological cycle, and indices such

as the PDO and El Niño are known to influence rainfall patterns in the region (KAYANO et al. 2020). Numerous studies have explored the presence of chaos and nonlinearity in El Niño dynamics (NEELIN et al., 1998; STONE et al., 1998; MAJUMDER & KANJILAL, 2019; OGUNJO & FUWAPE, 2020). These studies establish nonlinear structures and even indications of chaoticity. Given the identification of a chaotic signal in El Niño events, the RP was applied to the monthly time series to investigate the potential influence of this index on rainfall and streamflow patterns. The RP of El Niño (Figure 20) displays white stripes around 1972, 1981, and 1997. These are consistent with the breakpoints observed in the rainfall time series around 1971 and 1983 and in the streamflow around 1982. Rolim et al. (2021) explored the influence of the Oceanic Niño Index (ONI) in the same region using information theory metrics. They found that during El Niño years (e.g., 1963, 1965, 1982, 1987, 1991, 1992, 1997, 2002), rainfall is scarce, resulting in higher variability. As observed in the RPs, these years showed changes in rainfall and streamflow patterns. These alterations in the time series provide further evidence of the influence of climate indices in the region under study. The authors also analyzed the PDO, a well-known index that affects rainfall and streamflow in the region. They found that the PDO exhibits more recurrence points compared to the El Niño 3.4 index. However, both indices demonstrate similar transition periods (APPENDIX A).

Figure 20 - (a) El Niño 3.4 and (b) its recurrence plot.



(a)



(b)

Source: Prepared by the author

In addition to visualizing rainfall and streamflow behavior using RPs, RQA is also applied to describe the dynamics by quantifying the patterns in the RP's structure. The irregular and heterogeneous patterns in the RP suggest chaotic behavior, justifying its quantification. Four parameters in the RQA (RR, DET, LMAX, ENT) were computed for the monthly rainfall and streamflow time series that showed a saturation point in the CD method.

The first recurrence variable, RR, quantifies the percentage of recurrent points within a specified radius, with values ranging from 0 to 1. For rainfall, the RR values range from 0.003 to 0.06 (Table 3), showing that the recurrence matrix is sparse for rainfall, with fewer recurrent points. This result is consistent with the findings of Santana et al. (2020a), who also found low RR values for monthly rainfall in Pernambuco, a state also located in the NEB. For streamflow, the values range from 0.01 to 0.77, with lower values located in the southeastern and coastal parts of the State. The other areas show high recurrence, suggesting that the system is revisiting many states over time.

The second recurrence variable, DET, measures the proportion of recurrent points that form diagonal line structures reflecting the occurrence of regular patterns (predictability) in a time series. Periodic signals result in very long diagonal lines, chaotic signals result in short diagonal lines, and stochastic signals result in no diagonal lines at all. For the rainfall, the values

of DET range from 0.23 to 0.76, indicating that some stations have recurrent points in deterministic structures. The stations with the highest DET (#23 and #31) are located in the southeastern region of the State, while low DET values can be found in the north of the State. Lower values suggest a more complex and less predictable system. Furthermore, 46% of the rainfall stations had above-average DET and below-average RR, indicating a high degree of complexity. For streamflow, the DET value range from 0.3 to 1 (Table 4). Santana et al. (2020b) found high DET values (0.98 and 0.79 – values before and after the dam construction) for a streamflow time series in the NEB. For the streamflow, only 16% of the stations had above-average DET and below-average RR, with these stations located in the northwestern and southeastern areas of the State.

The third recurrence variable, linemax (LMAX), is inversely proportional to the LLE (ECKMANN et al., 1987). The shorter the LMAX, the more chaotic (less stable) the signal is. Stations #5, #6, #15, #27, and #28 presented high values of LMAX, while the LMAX values for stations #17, #21, and #31 equaled 0. For streamflow, only Stations #3, #5, #14, #15, #16, #20, #22, #24, and #25 presented low values of LMAX, indicating a more chaotic behavior. These stations are located mainly in the southeastern and coastal parts of the State. The fourth recurrence variable is ENT, a measure of signal complexity. For the rainfall, the ENT ranges from 0 to 0.64. Most of the State has low ENT values for rainfall, with high values concentrated in the western part of the State, indicating higher complexity in this region. The coastal area of the State showed low values of ENT, as also seen by Rolim et al. (2021). For the streamflow, the ENT varies from 0 to 3.37. Stations #3, #15, #22, #24, and #26, located in the coastal and southeastern regions of the State, have low ENT values.

Table 3 - Summary of RQA measures at different rainfall stations

Rainfall Station	REC	DET	LMAX	ENT	TT
1	0,00	0,48	2,00	0,16	2,14
2	0,01	0,24	3,00	0,09	2,04
3	0,01	0,24	2,00	0,05	2,03
5	0,01	0,28	9,00	0,15	2,03
6	0,02	0,23	15,00	0,12	2,03
7	0,01	0,39	2,00	0,11	2,00
8	0,01	0,25	2,00	0,09	2,00
9	0,01	0,32	2,00	0,15	2,05
10	0,01	0,29	2,00	0,12	2,00
14	0,01	0,39	2,00	0,16	2,09
15	0,03	0,34	19,00	0,18	2,24
16	0,01	0,40	2,00	0,35	2,00

17	0,01	0,38	0,00	0,00	2,00
18	0,01	0,26	3,00	0,28	2,00
20	0,00	0,46	2,00	0,64	2,00
21	0,01	0,24	2,00	0,07	2,30
22	0,01	0,36	2,00	0,27	2,00
23	0,00	0,61	0,00	0,00	2,00
24	0,01	0,30	2,00	0,08	2,00
26	0,00	0,45	2,00	0,30	2,00
27	0,06	0,27	69,00	0,17	2,32
28	0,05	0,29	25,00	0,45	2,17
30	0,01	0,31	2,00	0,07	2,03
31	0,00	0,76	0,00	0,00	0,00

Source: Prepared by the author

Table 4 - Summary of RQA measures at different streamflow stations

Streamflow Station	REC	DET	LMAX	ENT	TT
3	0,03	1,00	0,00	0,00	0,00
4	0,05	0,88	25,00	1,25	7,67
5	0,02	0,53	2,00	0,41	2,25
6	0,27	0,94	48,00	3,37	32,51
9	0,71	0,94	69,00	2,64	17,27
10	0,27	0,53	22,00	0,38	3,24
11	0,77	0,99	101,00	2,21	17,28
12	0,50	0,92	62,00	1,95	9,13
13	0,18	0,53	28,00	0,84	3,44
14	0,03	0,50	3,00	0,31	2,36
15	0,02	1,00	0,00	0,00	0,00
16	0,08	0,49	4,00	0,38	2,45
17	0,67	0,93	155,00	1,61	16,93
18	0,21	0,45	27,00	0,23	3,01
19	0,06	0,73	18,00	0,52	3,74
20	0,02	0,53	2,00	0,64	2,27
21	0,10	0,65	5,00	0,76	3,03
22	0,06	0,39	3,00	0,14	2,13
23	0,17	0,57	39,00	1,00	5,26
24	0,01	0,46	0,00	0,00	2,07
25	0,05	0,30	4,00	0,55	2,69
26	0,03	0,49	7,00	0,14	2,16
27	0,19	0,74	18,00	1,01	3,95
29	0,19	0,64	57,00	0,48	4,18
30	0,17	0,64	24,00	0,81	4,50

Source: Prepared by the author

4.5 Conclusion

Reliable modeling and forecasting of rainfall and streamflow are important for water resource applications. However, the complex dynamics inherited in hydrological data present a challenge for analysis. This study used nonlinear and chaos theory methods to explore temporal and spatial variability of rainfall and streamflow dynamics throughout a Brazilian state. Phase space reconstruction, correlation dimension, Lyapunov exponent, recurrence plot, and recurrence quantification analysis were applied to detect the presence of deterministic chaos in the data and provide an understanding of hydrological variable modeling.

With regard to time delay estimation, we found that the rainy season plays a crucial role in the reconstruction of the attractor for rainfall. On the other hand, streamflow exhibits a longer time delay. This extended time delay can be attributed to the iterations that this variable maintains with other processes in the hydrological cycle and land use. Moreover, we found the spatial variability of rainfall to be higher than that of streamflow, likely due to differences in soil types and characteristics of intermittent rivers.

We also investigated the impact of timescale on the reconstruction of the phase space for rainfall and streamflow. With an increase in timescale, the required dimension for reconstruction decreased for most stations, using both the FNN and the CD methods. Moreover, for streamflow, both methods showed higher embedding dimensions in the State's western region. This suggests that these systems require more variables to describe their behavior, thus implying increased complexity.

The saturation of the correlation exponent versus the embedding dimension is indicative of chaotic behavior. More than 70% of rainfall and 80% of streamflow stations presented a saturation for the monthly timescale. However, as the timescales increased, the detection of chaotic series decreased. Another metric to detect the presence of deterministic chaos is the LLE, which also has implications for predictability. The rainfall data showed positive LLE values, indicating the presence of deterministic chaos, while only a few streamflow stations had negative values. Higher values of LLE were found in the northern area for rainfall and in the northwestern region for streamflow. In addition, the streamflow data presented higher LLE values, suggesting a stronger presence of chaos. This finding has important implications for forecasting, as predictability is inversely related to the LLE. Thus, rainfall stations exhibiting deterministic chaos are more predictable over longer periods

compared to streamflow stations. It is worth mentioning that this low predictability of streamflow in the region can be related to the length of the time series, which is a limitation in this paper, and that only three time series could be analyzed due to missing values in the series.

Both variables presented a shift in their dynamics, as identified by the RPs, which displayed common shift periods. Further, climate dynamics impacted the rainfall and streamflow time series, with the Niño index presenting shifts similar to those in the rainfall and streamflow. Our results also highlighted the spatial heterogeneity of hydrological processes, with stations in the northwestern part of the State showing high values of ENT, indicating greater complexity in that region. Additionally, streamflow exhibited higher ENT values than rainfall, indicating higher complexity, as observed in the coefficient of variation in the statistical analysis.

The presence of chaos in rainfall and streamflow time series can be identified, even with changes in the timescale. However, the complexity of these series changes both spatially and temporally. Most methods indicated that the northwestern part of the State is more complex for both rainfall and streamflow. Furthermore, this study revealed that, due to the chaotic nature of the streamflow time series, long-term predictions may not be effective for water resources management in the region. This finding has significant implications for the development of an integrated water management plan for the region. While it may seem plausible to predict streamflow for water resource management within the study area, such attempts are likely to be highly flawed. However, rainfall results show that it might be possible to predict the chaotic time series for up to six years. The results presented here have potential relevance to several applications, as they provide insight into how the dynamics of the hydrological variables evolve and space in the State. This study found that climate dynamics have a significant influence on rainfall and streamflow time series. Therefore, future studies should consider a more in-depth analysis of how specific climatic indices influence these hydrological variables.

5 COMPLEXITY ANALYSIS OF RAINFALL AND STREAMFLOW APPLYING MULTISCALE ENTROPY TO THE STATE OF CEARÁ, BRAZIL.

5.1 Introduction

The hydrological constraints in arid and semi-arid regions, such as increasing temperatures, infrequent precipitation, and drought conditions, along with variations at the inter-annual and higher scales, render these areas susceptible to the impacts of climate variability (RAMARAO et al., 2019; SINGH & CHUDASAMA, 2021). Climate variability and change can lead to alterations in the hydrological cycle, resulting in changes in the quantity and distribution of hydrological variables, which in turn can negatively impact water availability and water quality (JEMAI et al., 2017; ZHENG et al., 2017).

Ceará State is situated in the Northeastern region of Brazil (NEB), which is the world's most populated semi-arid area. According to the IPCC (2014), the NEB is among the planet's regions that are most susceptible to the impacts of climate change in the upcoming century. Climate change and human activities are expected to result in significant temperature increases and reduced rainfall in the NEB (WU et al., 2016; MARENGO et al., 2017; DE JONG et al., 2018). The analysis of hydrological variability in drought-prone areas plays a crucial role in water resource planning and management as emphasized by Cirilo et al. (2017). Hence, it is imperative to conduct climate risk assessments in order to adapt to the increased risk of extreme events. However, comprehending the spatiotemporal complexity of hydrological variables remains a major challenge (MISHRA & SINGH, 2010; TONGAL & SIVAKUMAR, 2019; GUNTU et al., 2020).

Understanding the complexity of hydrological variables, such as precipitation, streamflow, and soil moisture can provide valuable information about the behavior of a catchment and the processes driving water movement within it. Further, comprehension of these complex hydrological variables can aid in the development of better water resource management strategies. For example, it can help in predicting droughts, floods, and water shortages, as well as improving water resource allocation. In the literature, several methods and statistical analyses are used to assess the variability in a time series, such as variance, regression analysis, nonparametric tests, diversity indices, and measures based on information theory (ZHANG et al., 2019). Nonetheless, techniques like regression methods, mostly capture the regularity of precipitation and may not accurately quantify the extent of irregular change. As a result, these techniques are not well-suited to characterizing the complex, intrinsic fluctuations

in precipitation.

In contrast, over the past few decades, innovative concepts from complex systems science have been applied more frequently time series analysis, improving our understanding of hydrological processes. These studies employ techniques such as fractal and multifractal methods (TAN & GAN, 2017; ADARSH et al., 2020), chaos theory (SIVAKUMAR, 2009; FUWAPE et al., 2017), and information measures (DA SILVA et al., 2016; TONGAL & SIVAKUMAR, 2019; ROLIM et al., 2022), and complex networks (FANG et al., 2017; BRAGA et al., 2016) to evaluate the nonlinearity and complexity of rainfall dynamics.

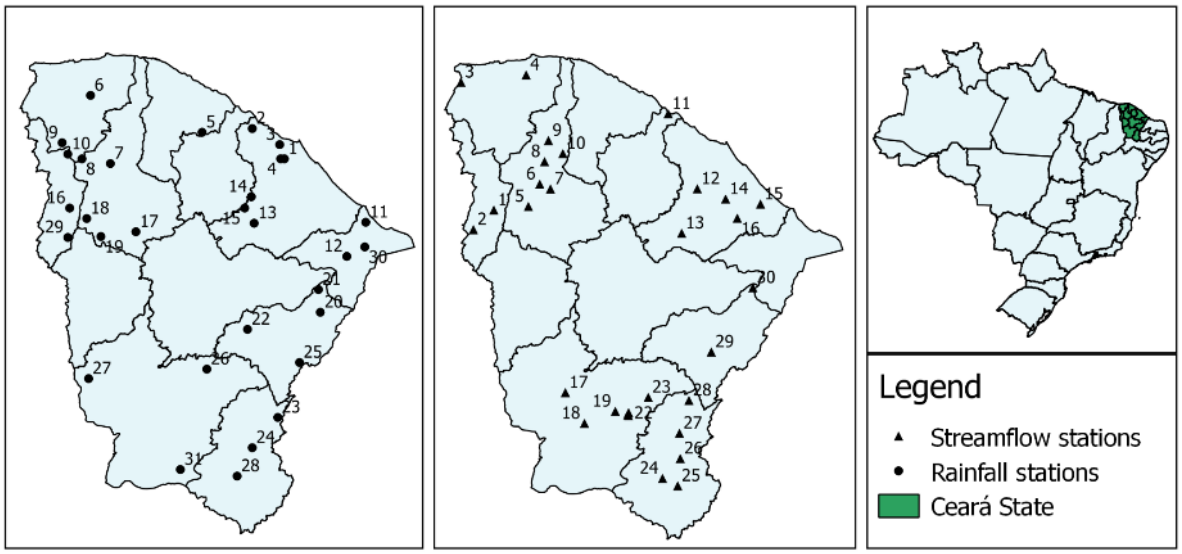
Among the information theory methods, entropy (SHANNON, 1948) has garnered significant attention in recent years in the field of water resources studies (ZHANG et al. 2019). Variations in entropy values reflect the amount of information transmitted or gained, enabling insights into a given variable. Additionally, hydrological time series exhibit variability at various temporal scales, and a thorough analysis requires a multiscale approach. Costa et al. (2002) introduced Multiscale Sample Entropy (MSE) as an extension of Sample entropy (RICHMAN & MOORMAN, 2000), which is calculated at multiple time scales and can reveal the structural complexity of the hydrological processes. In hydrology, MSE has been utilized to analyze streamflow and has been valuable in exploring alterations related to human activities (LI & ZHANG, 2008; ZHANG et al., 2012; BARRETO et al., 2020). However, less is known about the multiscale complexity of rainfall processes (DA SILVA et al., 2021).

The nonlinear and nonstationary nature of the climate system highlights the importance of exploring its inherent complexity in order to understand its uncertainty and variability. This information can also be useful in developing strategies for adaptation and mitigation measures aimed at enhancing the resilience of water systems. The study was conducted in the State of Ceará, located in the Northeast of Brazil, due to its history of severe and sometimes multi-year droughts. These long-term events pose challenges and weaknesses in the water resource system, emphasizing the significance of investigating the uncertainties in hydrological variables for improving water resource planning. This study aims to make a contribution in this area by showing that an entropy analysis can provide a new perspective on this phenomenon. The main objective of this study is to assess the spatiotemporal rainfall and streamflow variability using a multiscale approach and analyzed the relationship between these two processes. Additionally, a 10-years moving-window was applied to time series and the SampEn was calculated in order to assess the change in complexity for these hydrological variables over the years.

5.2 Study area and hydrological data

Located in the Northeast region of Brazil (NEB), Ceará (Figure 21) is predominantly semi-arid, with an economy heavily reliant on agricultural production and rainfall. The Intertropical Convergence Zone (ITCZ) controls the rainfall regime, with the majority of rainfall occurring from February to May, accounting for about 70% of the annual total. Although the mean annual rainfall is approximately 810 mm, values are higher in the coastal area and lower in the central region where the semi-arid landscape is more prominent (COSTA et al., 2021). The shallow soils with crystalline basement prevalent in most parts of the state lead to low base flows, resulting in predominantly naturally ephemeral or intermittent rivers in the area.

Figure 21 - Location of the meteorological stations



Source: Prepared by the author

The spatiotemporal complexity analysis utilized rainfall and streamflow data sourced from the Brazilian National Water Agency (ANA). The monthly rainfall data spans from January 1962 to December 2006. Table 5 describes the statistical characteristics of the rainfall timeseries. The monthly streamflow data, on the other hand, has varying start and end dates depending on the location, as detailed in Table 6. For this study, 31 rain gauges and 30 streamflow gauges were chosen, all situated in the State of Ceará, and depicted in Figure 21.

Table 5 - Geophysical and statistical information for rainfall time series (1962-2009) of 31 meteorological stations

Station Number	Station Name	Latitude	Longitude	Minimum Annual Rainfall (mm)	Maximum Annual Rainfall (mm)	Mean Annual Rainfall (mm)
1	Maracanau	-3.90	-38.63	595.6	2778.5	1333.32
2	São Gonçalo do Amarante	-3.58	-38.96	434.1	2200.3	1076.51
3	Caucaia	-3.75	-38.68	675.5	2418.6	1340.72
4	Maranguape	-3.90	-38.66	465.2	2311.7	1254.94
5	Uruburetama	-3.61	-39.50	419	2061.8	1173.36
6	Martinopole	-3.23	-40.68	399.8	2186.7	1104.34
7	Carire	-3.95	-40.46	260	2081	951.93
8	Mucambo	-3.90	-40.76	476.6	2120.8	1040.49
9	Tianguá	-3.73	-40.98	469.5	2409.3	1270.84
10	Ubajara	-3.85	-40.91	616	2864.5	1528.52
11	Aracati	-4.56	-37.76	220	2654.1	991.02
12	Russas	-4.93	-37.96	173.4	1929.1	771.42
13	Itapiuna	-4.58	-38.95	284.2	1573	783.11
14	Mulungu	-4.30	-38.98	479.1	2149.9	1187.90
15	Aratuba	-4.41	-39.05	416.3	3238.2	1449.65
16	Croatá	-4.41	-40.90	154	1116.8	579.52
17	Catunda	-4.66	-40.20	27.1	1987.7	720.25
18	Ipueiras	-4.53	-40.71	280.5	2004.2	972.67
19	Nova russas	-4.71	-40.56	351	1703.4	860.34
20	Alto santo	-5.51	-38.25	116.7	1823.9	775.31
21	São João do Jaguaribe	-5.28	-38.26	189.3	1852.5	765.93
22	Solonópole	-5.70	-39.01	237.2	1375.8	782.61
23	Umari	-6.63	-38.70	298	1702.6	792.64
24	Aurora	-6.95	-38.96	517.8	1951	934.20
25	Pereiro	-6.05	-38.46	294.6	2065.7	1024.07
26	Acopiara	-6.11	-39.45	366	1713	791.03
27	Parambu	-6.21	-40.70	229.4	2237.1	697.91
28	Missão Velha	-7.25	-39.13	594	1885.6	1006.12
29	Poranga	-4.73	-40.91	144	1262.8	657.56
30	Jaguaruana	-4.83	-37.78	118	1887.7	824.63
31	Santana do Cariri	-7.18	-39.73	304.4	1612.1	930.15

Source: Prepared by the author

Table 6 - Geophysical and statistical information for streamflow time series of 30 meteorological stations

Station Number	Station Name	Latitude	Longitude	Start/End date of the time series	Minimum Annual Streamflow (m^3/s)	Maximum Annual Streamflow (m^3/s)	Mean Annual Streamflow (m^3/s)
1	Croatá	-4.42	-40.91	01-01-1998/01-12-2006	0.47	7.79	3.47
2	Saudoso	-4.62	-41.13	01-01-1998/01-12-2010	0.36	9.91	3.64
3	Chaval Retiro	-3.08	-41.26	01-01-2001/01-12-2008	1.33	7.94	4.79
4	Pesqueiro	-3.00	-40.57	01-01-2001/01-12-2015	0.24	19.33	4.93
5	Fazenda Cajazeiras	-4.38	-40.55	01-01-1963/01-12-1972	0.14	23.35	8.02
6	Várzea do grosso	-4.14	-40.43	01-01-1987/01-12-2005	1.62	26.77	7.96
7	Trapiá	-4.20	-40.32	01-01-1988/01-12-2010	0.06	37.90	6.56
8	Groairas	-3.91	-40.38	01-01-1991/01-12-2010	0.35	33.50	4.48
9	Sobral	-3.69	-40.34	01-01-2006/01-12-2017	1.99	144.15	21.70
10	Fazenda bela Vista	-3.82	-40.19	01-01-1985/01-12-2003	0.00	1.37	0.23
11	São Luíz do Curu	-3.41	-39.08	01-01-1973/01-12-1984	2.16	144.86	19.94
12	Barra Nova	-4.19	-38.78	01-01-1999/01-12-2017	0.16	4.43	1.43
13	Caio Prado	-4.66	-38.94	01-01-1985/01-12-2005	0.02	21.13	4.97
14	Chorozinho	-4.30	-38.48	01-01-1970/01-12-1983	0.25	73.65	12.68
15	Itapeim	-4.35	-38.12	01-01-2001/01-12-2008	1.45	8.47	4.24
16	Cristais	-4.50	-38.36	01-01-1970/01-12-1981	0.19	39.87	6.12
17	Arneiroz	-6.32	-40.16	01-01-1976/01-12-2003	0.00	77.06	8.70
18	Malhada	-6.65	-39.96	01-01-1988/01-12-2010	0.00	10.43	1.62
19	Sítio Patos	-6.52	-39.64	01-01-2000/01-12-2017	0.03	45.07	7.40
20	Sítio Conceição	-6.57	-39.50	01-01-1968/01-12-1980	0.64	19.12	5.37
21	Sítio Poço Dantas	-6.56	-39.51	01-01-1991/01-12-2011	0.00	16.35	3.99
22	Cariús	-6.53	-39.50	01-01-1992/01-12-2006	0.13	45.44	8.45
23	Iguatu	-6.37	-39.29	01-01-1974/01-12-1996	0.06	173.96	33.93
24	Sítio lapinha	-7.22	-39.14	01-01-1985/01-12-2006	0.50	16.08	4.48
25	Podimirim	-7.30	-38.98	01-01-1973/01-12-2017	0.06	23.09	3.52

26	Sítio santa Cruz	-7.01	-38.96	01-01-2001/01-12-2017	0.85	24.04	6.70
27	Lavras da Mangabeira	-6.75	-38.97	01-01-1980/01-12-1996	1.10	132.65	26.27
28	Icó	-6.41	-38.87	01-01-1984/01-12-2015	1.62	148.53	25.91
29	Jaguaribe	-5.90	-38.63	01-01-1987/01-12-2017	3.66	203.14	41.70
30	Peixe gordo	-5.23	-38.20	01-01-1998/01-12-2014	8.23	222.58	40.97

Source: Prepared by the author

5.3 Methods

5.3.1 Sample Entropy

Sample entropy (SampEn) is a statistical measure that quantifies the complexity or randomness of a time series signal (RICHMAN & MOORMAN, 2000). It is often used in biomedical signal processing and analysis to evaluate the regularity and unpredictability of physiological signals such as heart rate variability or EEG signals. The SampEn is a modification of the approximate entropy (ApEn) method (PINCUS, 1991) by excluding the segment self-matches. The SampEn (m, r, N) is calculated as the negative natural logarithm of the conditional probability that two sequences of a given length remain similar at m as the length of the sequence increases m+1, assuming that self-matches are excluded from the probability calculation. Lower values of sample entropy indicate a higher degree of regularity and predictable behavior, while higher values suggest more complex and random behavior (BARRETO et al., 2020). SampEn has been largely applied in many fields, including hydrology (CHOU et al., 2014; HU et al., 2019; XAVIER et al., 2019; ZHOU & LEI, 2020).

The algorithm for calculating sample entropy involves the following steps (RICHMAN & MOORMAN, 2000):

1. Given a time series of length N, $u(i), i = 1, \dots N$, a m-dimensional set of vectors $x_m(i) = \{u(i+j): j = 0, \dots, m-1\}$, starting at position $i = 1, \dots N - m + 1$;
2. The distance between two vectors $x_m(i)$ and $x_m(k)$ is calculated as the maximum difference of their corresponding scalar components: $d[x_m(i), x_m(k)] = \max\{|u(i+j) - (k+j)|: k = 0, \dots, m-1\}$
3. Next, count the number of B_i of vectors $x_m(k)$, such as the calculated distance is less than or equal to r, which is the tolerance level of accepting matches, $k = 1, \dots N - m$, and $k \neq i$, to exclude self-matches;

4. One then defines $B_i^m(r) = \frac{B_i}{N-m-1}$ and $B^m(r) = \frac{\sum_{i=1}^{N-m} B_i^m(r)}{N-m}$, where $B^m(r)$ is the probability that two vectors will match for m points;
5. Repeat steps 1 to 4 for $m+1$, $m+2$, and so on, until $m = N-1$, to find $A_i^m(r) = \frac{A_i}{N-m-1}$ and $A^m(r) = \frac{\sum_{i=1}^{N-m} A_i^m(r)}{N-m}$, where A_i is the number of vector $x_{m+1}(k)$, which are within r of $x_{m+1}(i)$, excluding self-matches, and $A^m(r)$ is the probability that two vector will match for m points;
6. The sample entropy is defined as the negative natural logarithm of the average conditional probability that two sequences of a given length remain similar as the length of the sequence increases, as describe in Equation 11.

$$S_E(m,r) = \lim_{N\to\infty}\left[-\ln\frac{A^m(r)}{B^m(r)}\right] \quad \text{(Eq. 11)}$$

And can be estimated by the statistics in Equation 12.

$$S_E(m,r) = -\ln\frac{A^m(r)}{B^m(r)} \quad \text{(Eq. 12)}$$

5.3.2 Multiscale Entropy

Multiscale entropy (MSE) is a statistical measure that extends the concept of SampEn to assess the complexity or randomness of a time series signal at multiple time scales (COSTA et al., 2002). Traditional entropy-based measures (e.g., Shannon entropy (SHANNON, 1948), Kolmogorov entropy (GRASSBERGER & PROCACCIA, 1983b), approximate entropy (PINCUS, 1991) tend to grow monotonically with the degree of randomness fail to quantify the complexity as a meaningful structural richness (BARRETO et al., 2020).

Costa et al. (2002) introduced the concept of multiscale sample entropy (MSE) by calculating sample entropy for consecutive coarse-grained time series $y_\tau(k) = \frac{1}{\tau}\sum_{i=(k-1)\tau+1}^{k\tau} u(i)$ where $k = 1, \dots N/\tau$, $u(i)$, $i = 1, \dots, N$ represents the initial time series and τ is a scale factor. MSE is obtained by plotting SampEn values for each scale factor τ, and is more suitable to adequately describe complexity that should find both fully random (white noise) and completely regular (e.g., periodic) series to be less complex than a series

stemming from a structurally complex process, such as $1/f$ noise (COSTA et al., 2002). MSE method was successfully used in different fields, including hydrology (LI & ZHANG, 2008; ZHANG et al., 2012; CHOU, 2012; BARRETO et al., 2020; DA SILVA et al., 2021).

The MSE provides a measure of the signal's complexity at different time scales, with higher values indicating higher complexity and randomness, and lower values indicating more regular and predictable behavior. The MSE can also be used to identify changes in the signal's complexity at different time scales, which may be indicative of changes in the underlying physical processes.

5.3.3 Trend analysis for complexity change

To analyze the change in complexity of rainfall and streamflow, we employed the moving window method. As suggested by Huang et al. (2011), the hydrological regime is more likely to change gradually than abruptly due to human activities and climate change, which makes the moving window a more effective approach to analyzing time series. We selected a moving window of 10 years to divide the original time series of rainfall and streamflow into subseries, for which we calculated the sample entropy separately to form a SampEn time series. Finally, we used the Mann-Kendall test to analyze the trend in the SampEn series, which could indicate a change in the complexity trend of rainfall or streamflow.

The Mann-Kendall (Mann 1945) is a nonparametric test widely used in environmental and hydrological time series. In this study, the test is applied to detect trends in the SampEn calculated from the data. The null hypothesis employed in the test assumes that the data came from a population with independent and identically distributed realizations. The test's null hypothesis was rejected if the p-value was less than the significance level of 0.05, detecting either increasing or decreasing monotonic trends. Sen's slope estimator was used to quantify the magnitude of the trend according to Sen's method (Sen 1968).

5.3.4 Hurst exponent

To evaluate the performance of information theory metrics in comparison to other analysis techniques, we computed the Hurst exponent for each precipitation station. The Hurst exponent serves as a measure of process persistence, capturing the degree of data clustering and

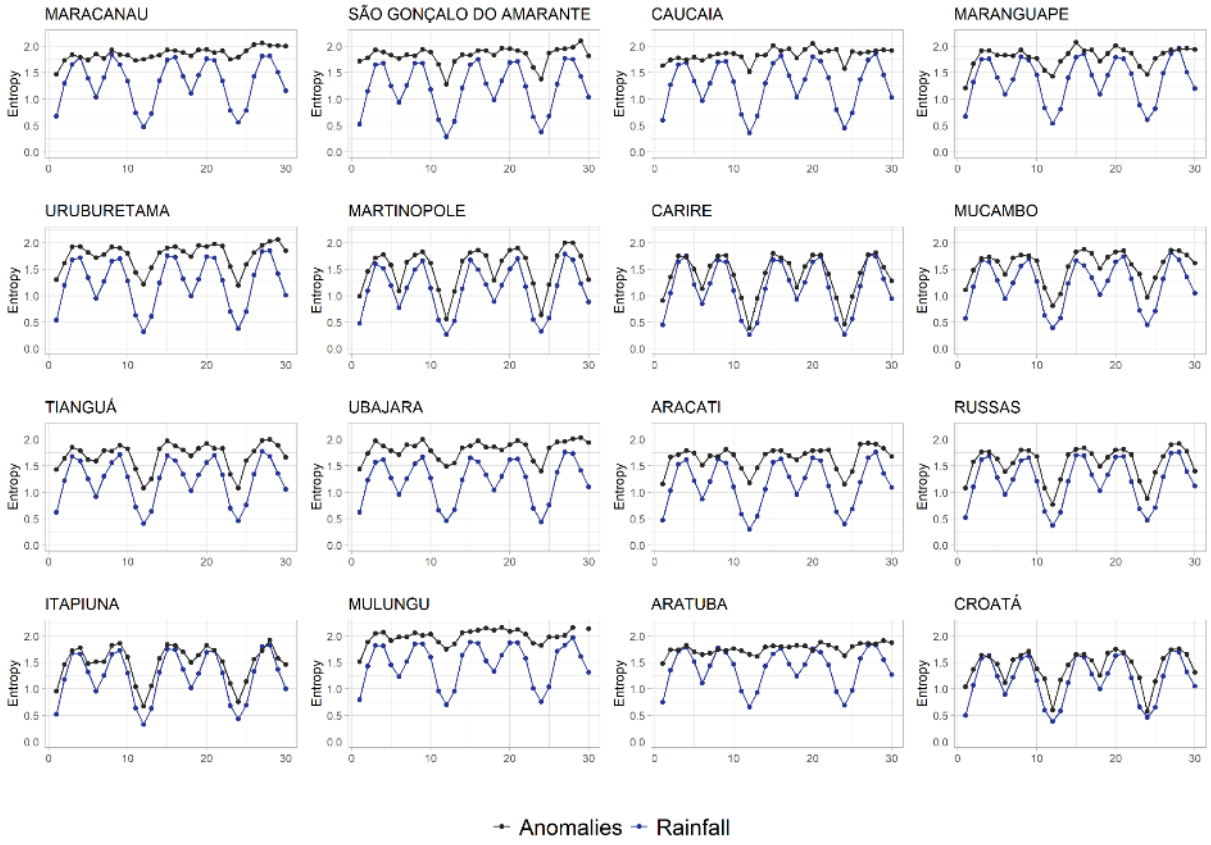
commonly referred to as a measure of system 'memory' or persistence (KOUTSOYIANNIS, 2002).

The Hurst exponent was estimated through the analysis of rescaled range (R/S). Among the methods applied to calculate the Hurst exponent, this is one of the most popular in many fields (KARMAKAR et al., 2019). This method involves dividing the time series into various subseries and calculating the mean and standard deviation for each subseries. Then, a demeaned series is calculated by subtracting the corresponding sub-means from the original time series. A cumulative series is calculated, and the widest range is generated (difference between maximum and minimum values). The range is then rescaled by dividing it by the standard deviation of the original time series. By plotting the rescaled range against the corresponding sub-series, the Hurst exponent can be determined as the slope of the resulting log-log plot. A value closer to 0.5 indicates a random or uncorrelated time series, while values greater than 0.5 indicate long-term memory or persistence, and values less than 0.5 suggest anti-persistence or anti-correlation. In this study, the Hurst exponent was calculated using the R/S analysis technique to gain insights into the persistence characteristics of the analyzed time series.

5.4 Results and Discussion

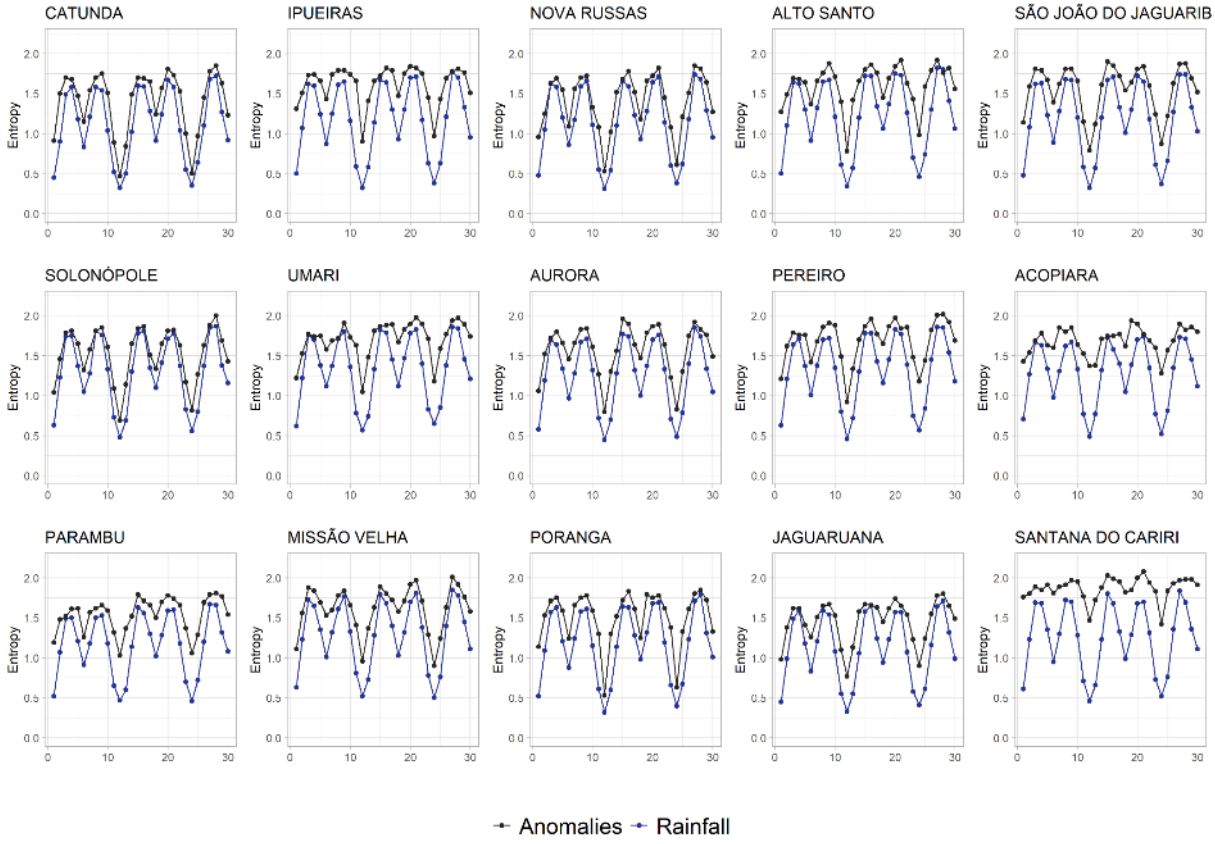
Figures 22 and 23 display the results of MSE values for both the original time series and anomalies at the 31 analyzed rainfall stations, with τ ranging from 1 to 30 months. Figures 24 and 25 show the same for the 30 analyzed streamflow stations. A homogeneous pattern is observed in the rainfall data, with well-defined intra-annual and inter-annual variability for most stations. The original series exhibit lower entropy values, suggesting greater regularity and predictability compared to the anomalous series. However, this difference becomes less pronounced for certain periods, particularly for the Maracanaú and Aratuba stations at τ values of 12 and 24. It is observed that some stations show inter-annual variability despite the calculation of anomalies, which is supposed to remove seasonality, revealing lower entropy values for τ values of 12 and 24. In addition to these two scales, a decrease in entropy is consistently observed at $\tau = 6$ in many stations. This suggests that these regions exhibit a form of synchronization in their rainfall patterns, showing patterns of intra-variability in their time series when the average monthly rainfall amount over a 6-month period is considered. The same behavior was observed in the rainfall patterns in Pernambuco State, also located in the NEB, by Da Silva (2021).

Figure 22 - MSE for monthly rainfall and monthly anomalies for stations from 1 to 16.



Source: Prepared by the author

Figure 23 - MSE for monthly rainfall and monthly anomalies for stations from 17 to 31.

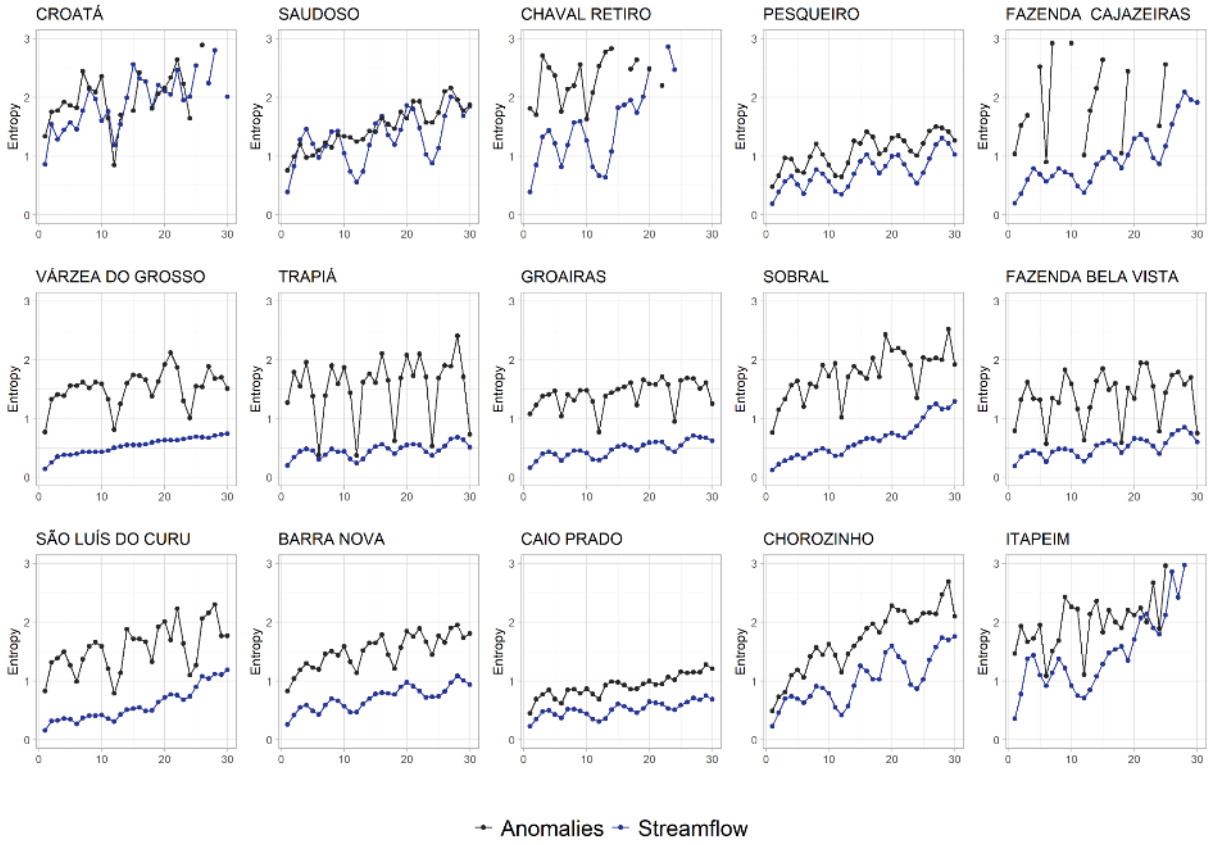


Source: Prepared by the author

Unlike the rainfall results, which show a cyclic pattern, the MSE values for streamflow results showed an increasing slope for most stations. This behavior indicates that for smaller time scales, the streamflow data shows greater regularity or predictability in the analyzed data, while for larger timescales, the series shows patterns of variation that are less predictable or more random. For most stations at τ=6, the streamflow time series showed a decrease in entropy, but some stations only showed that decrease at τ=12, such as Várzea do Grosso station. In addition, due to the length of some stations, the MSE was not calculated for longer timescales.

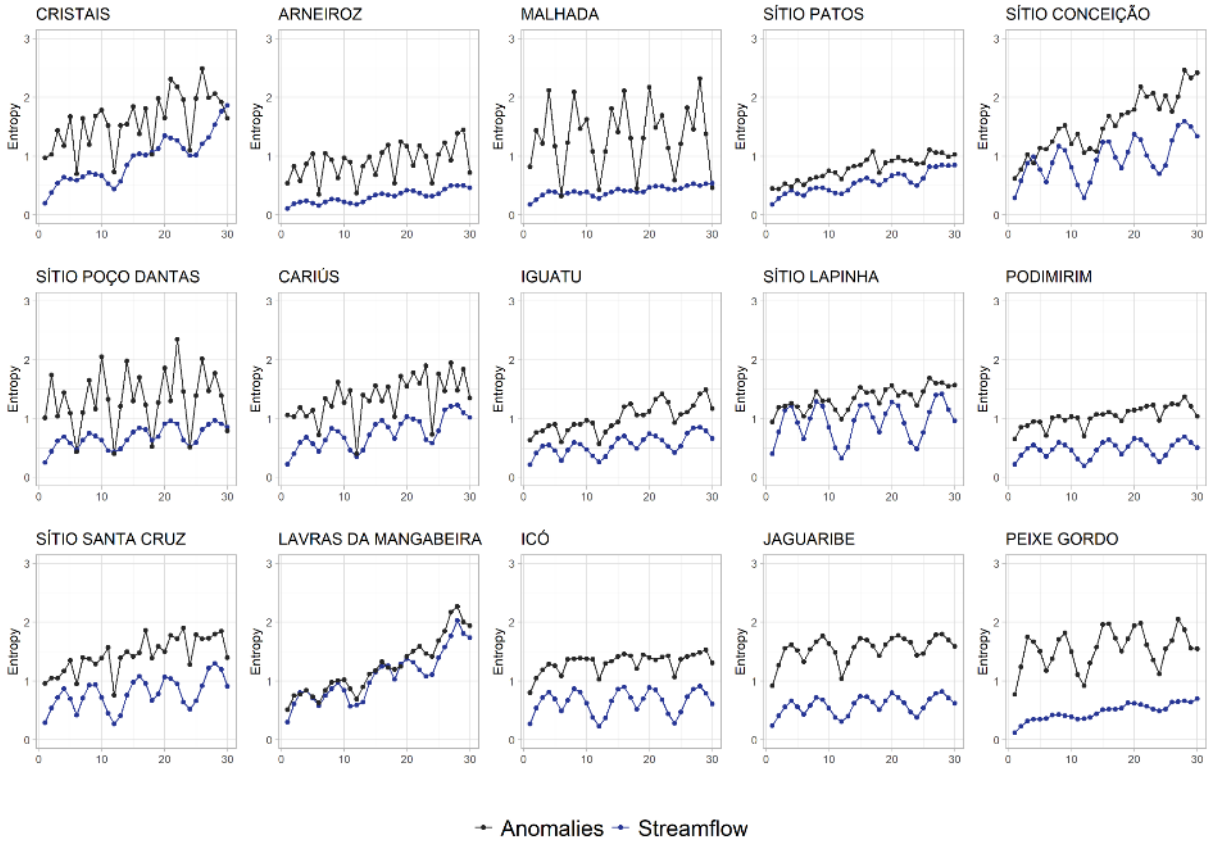
When comparing the results of the streamflow with the calculated anomaly, the original series showed lower entropy values. For most stations, there was a considerable difference between these two series. However, this difference is less pronounced in Lavras da Mangabeira, Sitio Patos, and Pesqueiro Stations.

Figure 24 - MSE for monthly streamflow and monthly anomalies for stations from 1 to 15.



Source: Prepared by the author

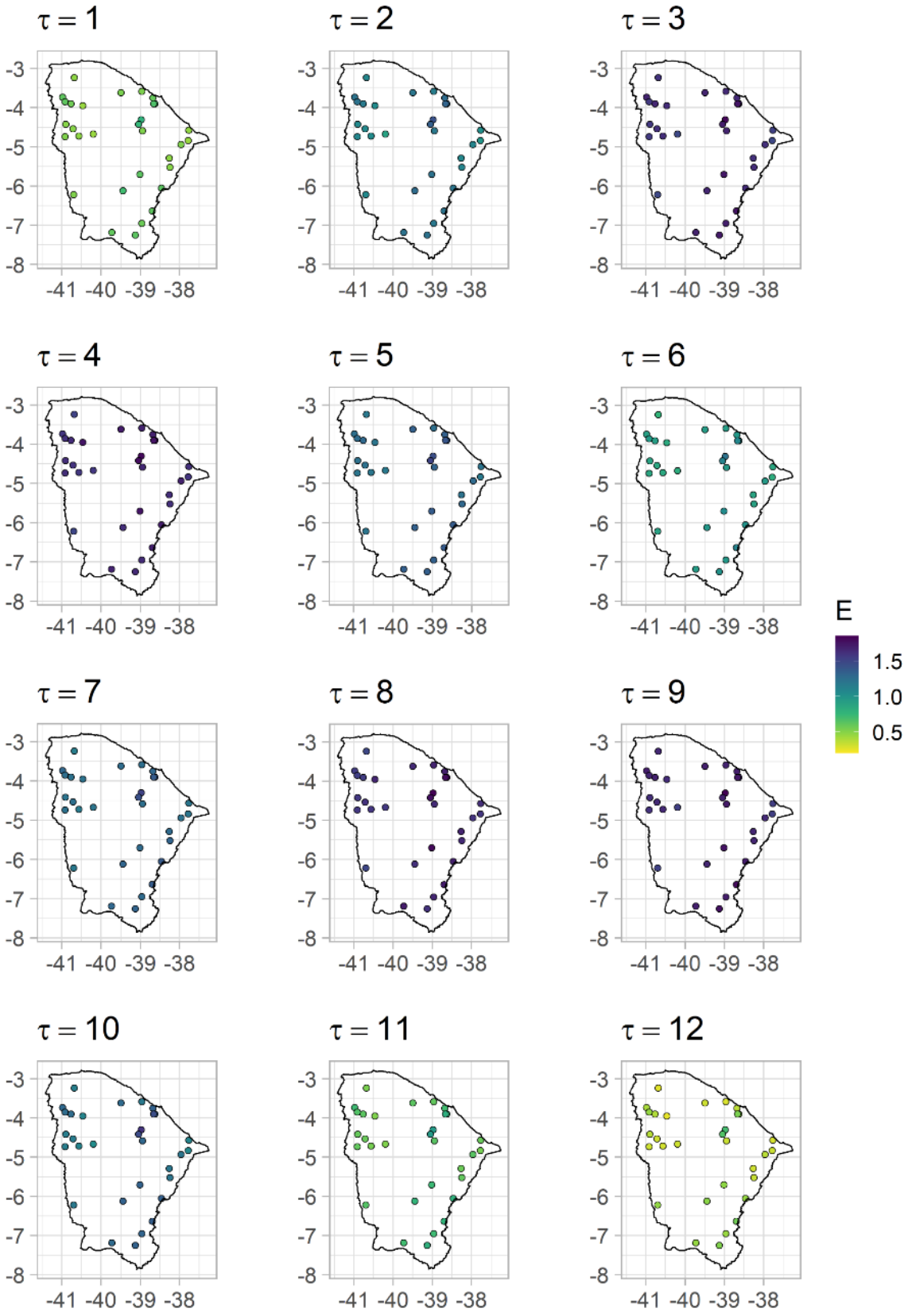
Figure 25 - MSE for monthly streamflow and monthly anomalies for stations from 16 to 30.



Source: Prepared by the author

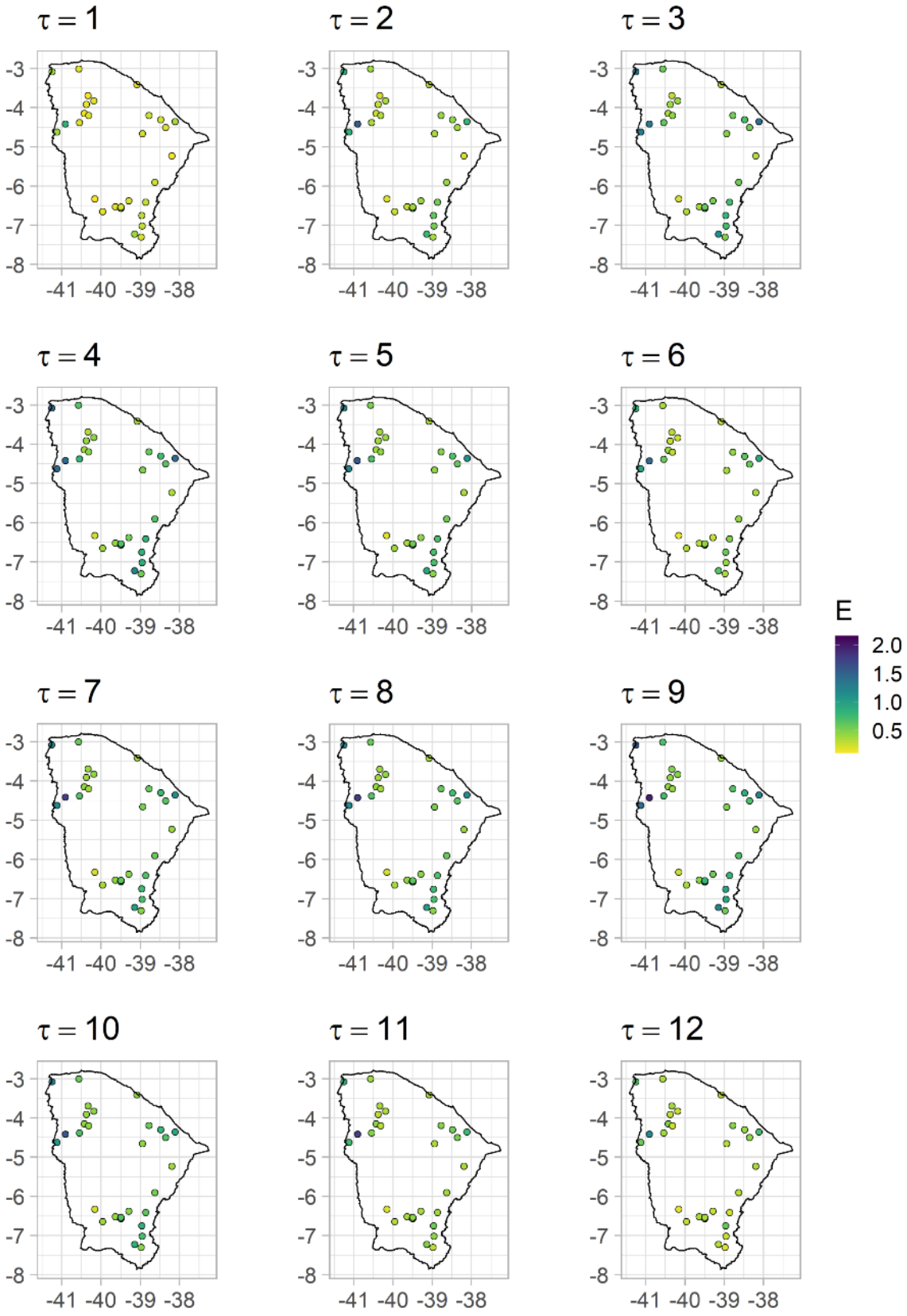
The MSE values for the rainfall and streamflow series of the analyzed stations across the state of Ceará were obtained, and their spatial distribution is displayed in Figure 26 and Figure 27, respectively. These figures reveal more detailed spatial patterns for each temporal scale. In particular, for the rainfall stations, lower entropy values are observed for $\tau = 1$ and $\tau = 12$, and the spatial distribution reflects the behavior seen in Figures 22 and 23. On the other hand, the streamflow time series displays lower entropy values compared to the rainfall series. Additionally, the stations located in the northwestern and southeastern regions of Ceará exhibit higher entropy values in the streamflow data.

Figure 26 - Spatial distribution of MSE from τ = 1 to τ = 12 months over the state of Ceará for rainfall original series



Source: Prepared by the author

Figure 27 - Spatial distribution of MSE from τ = 1 to τ = 12 months over the state of Ceará for streamflow original series



Source: Prepared by the author

To evaluate changes in the complexity of rainfall and streamflow time series, we applied a 10-year moving window to the rainfall time series and calculated SampEn for each sub-series. We then applied the Mann-Kendall test to the SampEn time series and present the trend analysis results in Tables 7 and 8. Due to the length of the streamflow time series, we only analyzed time series that had more than 10 sub-series, which totaled 14 stations. Of the 31 rainfall stations analyzed, we found no increase in flow complexity. However, 15 stations showed a decrease in rainfall complexity. The majority of stations that showed a decrease in rainfall complexity are located in the coastal region of Ceará State. Rolim et al. (2022) also applied a trend test to evaluate a standardized entropy measure of variability of rainy days within a given year and found a positive trend in five stations. Comparing both results, the rainfall in the region, which is concentrated in a few months of the year, tends to increase. However, for larger timescales the complexity is actually decreasing. When analyzing changes in streamflow complexity, we found that seven stations showed a significant trend, with five showing a decreasing trend, mainly located in the northern region of the State, consistent with the rainfall behavior in that region. The two stations that showed positive trends are located in the southeastern part of the State.

Table 7 - Trend of rainfall complexity change of the analyzed rainfall stations.

Station	Z value	p-value	Sen's slope	Trend
Maracanau	-6.048	0.000	-0.008	↓
São Gonçalo do Amarante	-4.127	0.000	-0.003	↓
Caucaia	-2.956	0.003	-0.003	↓
Maranguape	-2.166	0.030	-0.002	↓
Uruburetama	-3.827	0.000	-0.004	↓
Martinopole	-4.100	0.000	-0.004	↓
Carire	0.967	0.334	0.001	−
Mucambo	-1.430	0.153	-0.002	−
Tianguá	1.594	0.111	0.004	−
Ubajara	-0.885	0.376	-0.001	−
Aracati	-4.563	0.000	-0.006	↓
Russas	0.940	0.347	0.002	−
Itapiuna	-2.683	0.007	-0.003	↓
Mulungu	-6.757	0.000	-0.022	↓
Aratuba	-1.131	0.258	-0.003	−
Croatá	-1.403	0.161	-0.001	−
Catunda	-3.582	0.000	-0.004	↓
Ipueiras	-1.103	0.270	-0.001	−

Nova Russas	-1.784	0.074	-0.003	–
Alto Santo	-3.473	0.001	-0.004	↓
São João do Jaguaribe	-1.076	0.282	-0.001	–
Solonópole	-0.531	0.595	-0.001	–
Umari	-4.536	0.000	-0.011	↓
Aurora	0.586	0.558	0.001	–
Pereiro	-1.757	0.079	-0.004	–
Acopiara	-5.625	0.000	-0.008	↓
Parambu	-2.820	0.005	-0.006	↓
Missão Velha	0.994	0.320	0.002	–
Poranga	-3.092	0.002	-0.005	↓
Jaguaruana	-0.123	0.902	0.000	–
Santana do Cariri	-1.621	0.105	-0.001	–

Source: Prepared by the author

Table 8 - Trend of streamflow complexity change of the selected streamflow stations.

Station	Z value	p-value	Sen's slope	Trend
Várzea do Grosso	-2.862	0.004	-0.009	↓
Trapiá	1.095	0.274	0.002	–
Groairas	-2.647	0.008	-0.005	↓
Fazenda Bela Vista	-3.220	0.001	-0.007	↓
Barra Nova	-0.716	0.474	-0.005	–
Caio Prado	0.617	0.537	0.002	–
Arneiroz	-1.929	0.054	-0.002	–
Malhada	-0.657	0.511	-0.002	–
Sítio Poço Dantas	3.223	0.001	0.013	↑
Iguatu	1.700	0.089	0.005	–
Sítio lapinha	-3.722	0.000	-0.014	↓
Podimirim	2.275	0.023	0.002	↑
Icó	-1.690	0.091	-0.001	–
Jaguaribe	-2.651	0.008	-0.004	↓

Source: Prepared by the author

For comparison purposes, the Hurst exponent was calculated for rainfall and streamflow stations, and the results are presented in Table 9. Among the rainfall stations, the Aratuba station exhibited the highest Hurst exponent with a value of 0.77, implying a long-term memory of the system. This suggests that the current condition of the system will influence its subsequent state (Giri & Devercelli, 2023). A Hurst exponent greater than 0.5 signifies a higher level of complexity and long-term memory in the series. It implies that past patterns or trends persist into the future, resulting in a more intricate, structured, and complex behavior.

Conversely, Stations Carire e Santana do Cariri displayed low Hurst exponent values, indicating anti-persistence in the time series. This suggests reduced complexity and variability, with extreme events or high values less likely to persist over time. The series reverts to the mean more quickly, and the occurrence of high or low values becomes less predictable or persistent. Approximately 70% of the stations exhibited a Hurst exponent close to 0.5 (considering a 10% margin for more or less), indicating a state of randomness or uncorrelated behavior in the time series. This kind of time series is tough to predict, and hence it is consistent with the random behavior.

In the streamflow analysis, 44% of the Stations showed a Hurst exponent around 0.5, with the Vázea do Grosso station having an exact value of 0.5. Meanwhile, 47% of the streamflow stations, showed a higher Hurst exponent for streamflow, with Station Sítio Conceição which has a Hurst value of 1. In the MSE analysis, streamflow demonstrated lower entropy values compared to rainfall, indicating greater regularity or predictability, despite several stations having high Hurst values, indicating long-term persistence. This persistence allows for the identification of patterns or dependencies that can be used to make predictions. However, it's important to note that the relationship between the Hurst exponent and predictability is not linear or absolute. Other factors, such as the specific characteristics of the time series, the presence of external factors or trends, and the quality of the data, can also influence the predictability.

Table 9 - Hurst exponent for analyzed rainfall and streamflow stations.

Rainfall Station	H	Rainfall Station	H
Maracanau	0.58	Catunda	0.47
São Gonçalo do Amarante	0.60	Ipueiras	0.58
Caucaia	0.50	Nova Russas	0.48
Maranguape	0.54	Alto Santo	0.41
Uruburetama	0.52	São João do Jaguaribe	0.53
Martinopole	0.61	Solonópole	0.46
Carire	0.30	Umari	0.37
Mucambo	0.53	Aurora	0.35
Tianguá	0.45	Pereiro	0.47
Ubajara	0.54	Acopiara	0.43
Aracati	0.62	Parambu	0.57
Russas	0.51	Missão Velha	0.34
Itapiuna	0.50	Poranga	0.67
Mulungu	0.61	Jaguaruana	0.57
Aratuba	0.77	Santana do Cariri	0.31
Croatá	0.46		
Streamflow Station	H	Streamflow Station	H

Croatá	0.63	Cristais	0.59
Saudoso	0.85	Arneiroz	0.56
Chaval Retiro	0.56	Malhada	0.56
Pesqueiro	0.57	Sítio Patos	0.61
Fazenda Cajazeiras	0.56	Sítio Conceição	1
Várzea do grosso	0.50	Sítio Poço Dantas	0.58
Trapiá	0.38	Cariús	0.83
Groairas	0.63	Iguatu	0.79
Sobral	0.57	Sítio lapinha	0.69
Fazenda bela Vista	0.76	Podimirim	0.55
São Luíz do Curu	0.59	Sítio santa Cruz	0.40
Barra Nova	0.72	Lavras da Mangabeira	0.80
Caio Prado	0.81	Icó	0.67
Chorozinho	0.87	Jaguaribe	0.64
Itapeim	0.54	Peixe gordo	0.56

Source: Prepared by the author

5.5 Conclusion

The analysis of rainfall and streamflow data across various stations in Ceará, Brazil, revealed interesting patterns and insights into the complexity and predictability of these hydrometeorological variables. The main objective of this study is to investigate the spatiotemporal variability of rainfall and streamflow, using a multiscale entropy approach. Furthermore, a 10-year moving-window analysis was applied to the time series data of rainfall and streamflow to calculate SampEn and investigate changes in the complexity of these hydrometeorological variables over time.

We calculated MSE values for each rainfall and streamflow series across 30 temporal scales ranging from 1 to 30 months, for both the original series and their anomalies. The analysis of the original rainfall series from $\tau = 1$ to $\tau = 12$ showed homogeneity for each temporal scale, with lower values at $\tau = 1$, $\tau = 6$, and $\tau = 12$, revealing that these time scales exhibited the greatest regularity and predictability of rainfall dynamics. This is a reflection of the rainfall variability in the region, with well-defined dry and wet seasons, and interannual seasonality, respectively. This effect was less pronounced for rainfall anomalies where seasonality is removed. For the intra-annual temporal scales of the station, the entropy values of the anomaly series were lower for $\tau = 6$, which can be attributed to the duration of rainfall seasons.

The streamflow analysis also showed a well-defined intra-variability for $\tau = 6$ for most stations. Comparing the MSE values for rainfall and streamflow, the streamflow showed

lower entropy values, indicating greater regularity or predictability in the time series data, and suggesting that the data is less complex or less variable over time. For comparison purposes, the Hurst exponent was calculated and showed long-term persistence for many stations. The analysis demonstrated spatial and temporal patterns of complexity. The results contribute to a better understanding of the behavior of rainfall and streamflow in the region and provide insights for water resource management and planning.

In addition, the trend of the hydrometeorological variables' complexity was analyzed. The rainfall data showed a decrease in rainfall complexity for half of the stations in the State. In contrast, the streamflow data had two stations in the southeastern region of the state that showed an increase in entropy, revealing an increase in the complexity of this time series.

Overall, the study's findings contribute to our understanding of rainfall and streamflow behaviors in this region, elucidating their interactions and evolution over time, which could have significant implications for water resource management, land use planning, and climate change adaptation. Future research should further explore the implications of these complexities and the mechanisms behind these observed patterns, as well as extend the multiscale entropy approach to other hydroclimatic variables for a more comprehensive understanding of the hydrometeorological dynamics.

6 PHASE-SPACE RECONSTRUCTION WITH MACHINE LEARNING METHODS: APPLICATION OF RAINFALL FORECASTING TO THE STATE OF CEARÁ, BRAZIL.

6.1 Introduction

Water resources management and planning issues, such as flow forecasting to support reservoir operation, flood mitigation, and spillway design studies, require accurate modeling and prediction of hydrological variables to be resolved. However, this modeling process remains a significant challenge due to the extremely complex and highly nonlinear characteristics of hydrological variables. Additionally, anthropogenic and climate change impacts further complicate the prediction process, particularly in the medium- and long-term (DWARAKISH & GANASRI, 2015; GU et al., 2015; ZHAO et al., 2017).

Rainfall dynamics have long been an area of interest for hydrologists. In the past, stochastic models were commonly applied to hydrological variables (SIVAKUMAR et al., 2001; SIVAKUMAR, 2009) due to limitations in data availability and computational power. Although linear approaches are still popular, advancements in computational power and data collection have enabled the use of nonlinear approaches, such as artificial neural networks (KARUNASINGHE et al., 2006), data-driven models (FENG et al., 2020), and deterministic chaos theory (SIVAKUMAR, 2016). In recent years, data-based techniques have become increasingly popular in water resources management, particularly due to a lack of physical information on the watershed. These methods have shown promising results in the modeling processes (MESHRAM et al., 2020).

Chaotic systems are nonlinear dynamics that are sensitive to small changes in initial conditions. Moreover, chaotic systems have a symmetric property that plays a crucial role in producing their chaotic behavior (RAMADEVI & BINGI, 2022). This symmetry can be observed through phase-space reconstruction, which reveals dynamic information hidden in chaotic time series. Consequently, chaotic data behaves differently than stochastic data and requires different approaches during the modeling process (BOARETTO et al., 2021). Given these characteristics, in recent decades, many researchers have attempted to predict the future evolution of chaotic systems using a wide range of models. Although traditional prediction methods have not produced satisfactory results, advanced techniques using machine learning-based approaches have been proposed recently (RAMADEVI & BINGI, 2022). One of the main challenges in understanding and predicting chaotic systems is the high dimensionality and

chaotic behavior of hydrological systems. Nevertheless, recent advances in computing power, algorithmic innovations, and data availability have led to the convergence of different approaches that can address this issue (VLACHAS et al., 2018).

The advancement of machine learning has allowed for the development of new algorithms and strategies for identifying, controlling, and analyzing complex systems, resulting in an increased use of machine learning across various fields (TANG et al., 2020). One of the major advantages of machine techniques is that they do not require prior knowledge of the relationship between input and output variables. In recent decades, there have been many attempts to predict the future evolution of chaotic systems, and studies that apply machine learning techniques in the field of time series analysis and prediction (RAMADEVI & BINGI, 2022). Popular models include support vector machines (SVMs), but the most commonly used architecture is artificial neural networks (ANNs). Research in complex systems and machine learning has a strong relationship, providing a foundation for cross-disciplinary interactions.

Among the studies applied to hydrological variables, a Radial Basis Function Neural Network (RBFNN) model was developed to estimate the nonlinear hydrological time series of the Mekong River in Thailand and Laos, the Chao Phraya River in Thailand, and sea-surface temperature anomaly data (JAYAWARDENA et al., 2006). Ding and Ding (2009) developed an adaptive fuzzy inference-based neural network model was developed to predict medium- and long-term hydrological residual time series, using data collected from the Guantai hydrological station in the Zhang River, China. Baydaroğlu and Koçak (2014) applied Support Vector Regression (SVR) with Radial Basis Function (RBF) and phase-space reconstruction to predict daily evaporation amounts. Ouyang et al. (2016) used ensemble empirical mode decomposition (EEMD) and the phase-space reconstruction method to design input vectors for the SVR forecasting model, which was applied to the monthly rainfall at a weather station in Changchun, China. Liang et al. (2019) applied a statistical and chaotic nonlinear dynamic model to forecast stream water temperature from solar radiation and air temperature data in the Lake Tahoe basin, California and Nevada, USA.

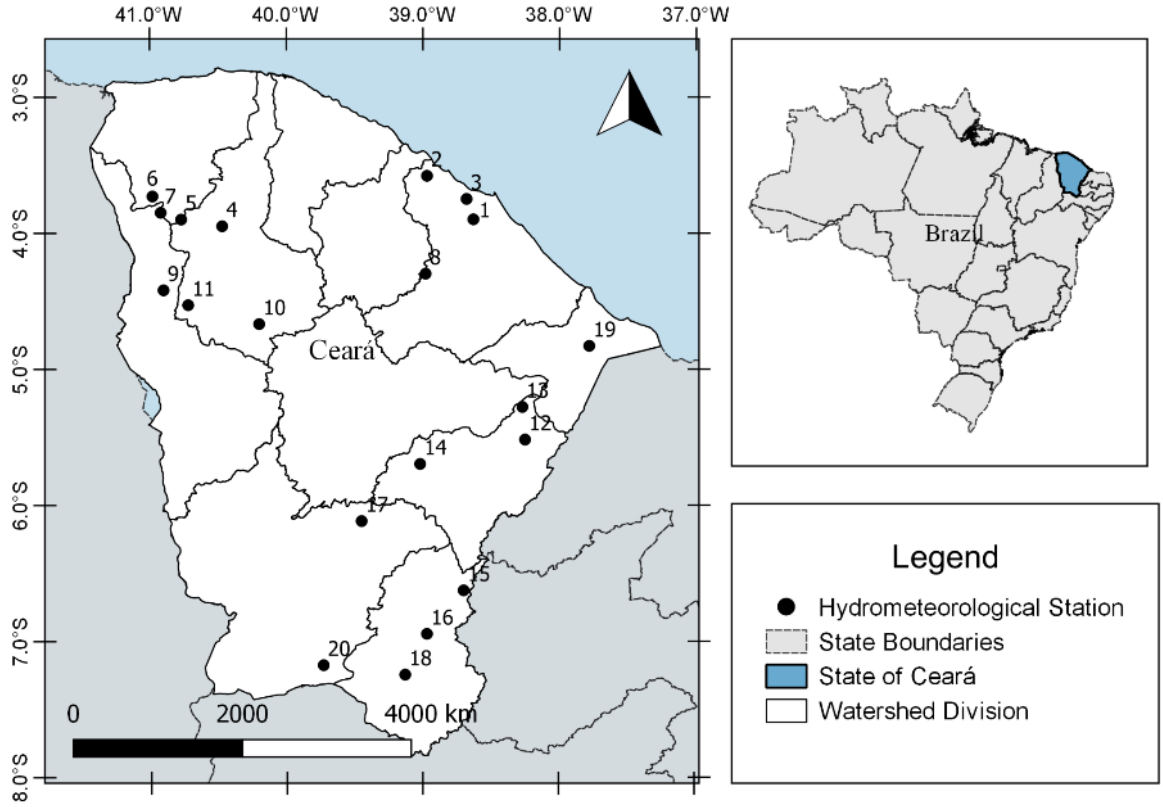
The selection of appropriate inputs is a crucial stage in rainfall forecasting, especially when using machine learning methods. The Phase-Space Reconstruction (PSR) approach can reveal unseen information in historical data. This study explores the integration of machine learning models with the PSR approach for rainfall pattern forecasting in Ceará, Brazil. To the authors' knowledge, there has been no documented research comparing machine learning models integrated with the PSR approach for rainfall pattern forecasting in Ceará or even in Brazil. Forecasting chaotic dynamics one- or a few-time steps ahead is often easy, as

demonstrated by the high performances obtained on many systems. However, the situation becomes more complex when considering mid- to long-term horizons, as small errors intrinsic to the chaotic nature of the data can expand. Even when the data is generated by a known chaotic system, small initialization or numerical errors can be amplified, leading to predictions that land on a random point on the system's attractor. Hence, this study aims to analyze the effect of these forecasts on the performance of the models. The main objective of this paper is to investigate the ability of multiple models such as Local Approximation Prediction (LAP), SVM, Decision Tree (DT), Random Forests (RF) and Long Short-Term Memory (LSTM) Artificial Neural Network for forecasting monthly rainfall data belong to 20 stations in the Ceará State over the period of 1962 to 2006.

6.2 Study area and hydrological data

Brazil is a country with high variability in rainfall due to its vast territory and diverse climates. The northeastern region, where Ceará is located, is known for its semi-arid climate with low rainfall and high evaporation rates. The rainy season in this region typically occurs from February to May, accounting for about 70% of the annual rainfall. The average annual rainfall is around 810 mm, with higher values in the coastal area and lower values in the central portion, where the semi-arid landscape is more pronounced (COSTA et al., 2021). The rainfall data used in this study is from the Brazilian National Water Agency (ANA). The monthly rainfall series ranges from January 1962 to December 2006. This study considered 20 rain gauges located in the State of Ceará shown in Figure 28.

Figure 28 - The location of the hydrometeorological stations in Ceará State, Brazil.



Source: Prepared by the author

The rainfall time series have vastly dissimilar statistical characteristics. The stations situated in the northern part of the state have a higher average rainfall, while those in the central region have a lower average. Furthermore, the stations in the northwest exhibit the largest range of values (Table 10).

Table 10 - Statistical summary of the monthly rainfall time series.

Station number	Station name	Mean (mm)	Variation coefficient
1	Maracanau	110.82	1.19
2	São Gonçalo do Amarante	80.60	1.41
3	Caucaia	89.66	1.32
4	Carire	64.39	1.42
5	Mucambo	111.18	1.25
6	Tianguá	64.83	1.37
7	Ubajara	78.56	1.45
8	Mulungu	65.32	1.31
9	Croatá	86.52	1.33
10	Catunda	66.17	1.31
11	Ipueiras	104.76	1.26

12	Alto Santo	77.19	1.35
13	São João do Jaguaribe	126.40	1.21
14	Solonópole	66.50	1.32
15	Umari	99.26	1.07
16	Aurora	83.85	1.29
17	Acopiara	47.92	1.52
18	Missão Velha	68.73	1.58
19	Jaguaruana	59.22	1.61
20	Santana do Cariri	76.71	1.31

Source: Prepared by the author

6.3 Methods

6.3.1 Phase Space Reconstruction (PSR)

Considering dynamic systems, systems can be classified as deterministic or stochastic. Deterministic systems follow a set of rules or equations that determine their behavior and can be predicted exactly if their current state is known, while stochastic systems are affected by randomness or uncertainty and cannot be predicted exactly. The phase space reconstruction (PSR) approach can be used to identify the chaotic, deterministic, or stochastic nature of dynamic systems (SIVAKUMAR et al., 2001).

PSR is a technique that extracts information about the dynamic behavior of a system from a time series. The technique maps the time series dataset onto a higher-dimensional phase space, allowing the underlying dynamic behavior of the system to be easily observed (LIANG et al., 2019). The Takens theorem (TAKENS, 1981) states that a one-dimensional time series of length k can be expressed in terms of phase space vectors as described in Equation 11.

$$\boldsymbol{Y}_i = \left(X_i, X_{i+\tau}, X_{i+2\tau}, \dots, X_{i+(m-1)\tau}\right), \quad i = 1, 2, \dots, K \qquad \text{(Eq. 13)}$$

where τ is the time delay, m is the embedding dimension, and $K = k - (m-1)$ (TAKENS, 1981). Even though phase space is a powerful method to gain qualitative information about how a system evolves, the calculation of the reconstruction parameters m and τ is crucial for a proper reconstruction (JIANG et al., 2020). In this study, a fixed delay time is used to make predictions for different embedding dimensions. The time delay is determined using mutual information (MI), which is a statistical measure of the dependence between two variables. The delay that minimizes the dependence between the chosen variables is selected (STROZZI et al., 2002). The MI has been commonly used for analyzing nonlinear time series and the time delay

is often chosen as the local minimum of the MI (FRASER & SWINNEY, 1986). Further details on the calculation of MI can be found in Sivakumar (2016).

6.3.2 Nonlinear Local Approximation Prediction

The Local Approximation Prediction (LAP) (Farmer and Sidorowich 1987) of a time series requires the reconstruction of the phase space. The PSR parameters enable the interpretation of the underlying dynamics in the form of an m-dimensional map, which can be expressed in Equation 12.

$$\boldsymbol{Y}_{j+T} = f_T(\boldsymbol{Y}_j) \qquad \text{(Eq. 14)}$$

where $\boldsymbol{Y}_j$ and $\boldsymbol{Y}_{j+T}$ are vectors of dimension m, describing the state of the system at times j (i.e., the current state) and $j + T$ (i.e., the future state), respectively. The problem then is to find an appropriate expression for f_T. LAP involves dividing the points on the attractor into an initial learning set, used to predict points in the remaining test set. The parameters are then fitted to each neighborhood separately, allowing the underlying system dynamics to be represented step by step in the phase space. To predict X_{j+T} based on $\boldsymbol{Y}_j$ (an m-dimensional vector) and past history, k nearest neighbors of $\boldsymbol{Y}_j$ are found on the basis of the minimum distance between points in the attractor. In general, however, m is varied to find out the optimum predictions (SIVAKUMAR, 2016). For a chaotic time series, the forecast the accuracy of the nonlinear forecast falls off with increasing the embedding dimension. For stochastic time series, by contrast, the forecast accuracy does not change with the embedding dimension (SUGIHARA & MAY, 1990).

6.3.3 Long Short-Term Memory (LSTM) Neural Network

Artificial neural networks (ANNs) are parallel-distributed information-processing systems that are inspired by the structure of the human brain, where knowledge is acquired through a learning process and finding optimum weights for the different connections between the individual neurons (KALIN et al., 2010; NOORI et al., 2020). ANNs are particularly useful for forecasting as they handle the nonlinearity and instability of hydrological time series effectively when the input vectors are designed using PSR method (PENG et al., 2017).

One ANN that has obtained outstanding performances in natural language

processing in recent years is the Long Short-Term Memory (LSTM) neural network (HOCHREITER & SCHMIDHUBER, 1997). Traditional Recurrent Neural Networks (RNNs) have a problem with the vanishing gradient, which occurs when the gradients of the error backpropagated through the network become very small, making it difficult for the network to learn from the data (CHO & KIM, 2022). LSTMs were developed to address this problem. The hidden layers in LSTMs contain recurrent neurons called LSTM cells, rather than traditional feed-forward neurons. Each LSTM cell or memory cell has two internal states, which are the hidden (h_t) and cell state (C_t) and three gates: the input (i_t), the output (o_t), and forget gate (f_t). This structure maintains its state and regulate the flow of information in the cell-state structure. This allows the LSTM to selectively choose what information to keep and what to discard, effectively avoiding the vanishing gradient problem. This makes LSTM suitable for tasks such as speech recognition, natural language processing, and time series prediction. The fundamentals of LSTM were well described in Hrnjica and Mehr (2020).

The architecture of the LSTM unit is illustrated in Figure 29. Equations 15 to 20 describes the LSTM equations.

$$i_t = \sigma(W_i x_i + U_i h_{i-1} + b_i) \quad \text{(Eq. 15)}$$

$$f_t = \sigma(W_f x_f + U_f h_{f-1} + b_f) \quad \text{(Eq. 16)}$$

$$o_t = \sigma(W_o x_o + U_o h_{o-1} + b_o) \quad \text{(Eq. 17)}$$

where W_i, W_f, and W_o indicate the weights linking the input, forget, and output gates with the input, respectively; U_i, U_f, and U_o represent the weights from the input, forget, and output gates to the hidden layer, respectively; b_i, b_f, and b_o denote the bias vectors for the input, forget, and output gate, respectively. The cell state is responsible for keeping track of the relevant information provided by past inputs, while the hidden state synthesizes the information provided by the current input, cell state, and previous hidden state. The gates regulate the flow of information in and out of the neuron using a sigmoid activation function. The input and output gates act as signal filters, distinguishing time-dependent and time-independent signals to prevent input and output weight conflicts. The input and forget gates control the extent to which a new input value and the current cell state, respectively, affect the new cell state.

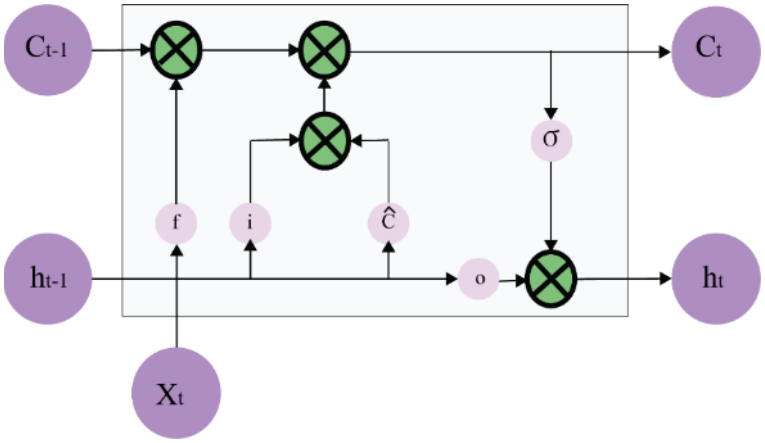
$$\tilde{C}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad \text{(Eq. 18)}$$

$$C_t = f_t \otimes C_{t-1} + i_{t-1} \otimes \tilde{C}_t \quad \text{(Eq. 19)}$$

$$h_t = o_t \otimes tanh(C_{t-1}) \quad \text{(Eq. 20)}$$

$\tilde{C}_t$ is the state of the cell at the previous time; C_t is the current state of the cell; h_{t-1} refers to output of the cell at the previous time point; h_t refers to output of the cell at the current time (LIU et al., 2020).

Figure 29 - LSTM-based neural network structure.



Source: Adapted from HOCHREITER & SCHMIDHUBER (1997).

6.3.4 Decision Tree (DT)

The decision trees (DTs) are a widely used and important method for supervised learning, which builds a tree-like model of decisions and their possible outcomes in a flowchart-like structure for easy interpretation (BREIMAN, 1984). DTs are constructed by repeatedly dividing the data into smaller sets based on the values of input variables, with each internal node in the tree representing an evaluation of an input variable, and each leaf node representing a prediction of the target variable (RAGETTLI et al., 2017). Data division is usually based on the input variable values, using criteria such as variance reduction or mean squared error to find the best split at each node. The predicted value of a new observation is obtained by traversing the tree from the root to a leaf node, making decisions based on the input variable values of the observation. To improve the accuracy of the predictions, the algorithm is run for multiple iterations on the tree. A common process in DT modeling is pruning the tree to a subtree to prevent overfitting and the misclassification risk of new samples (XU & LIANG, 2021). One of the major advantages of decision trees is their interpretability. However, one disadvantage of decision trees is their statistical instability even after pruning, meaning that small perturbations

or noise in the training data may result in a substantially different structure of the learned tree (Hastie et al., 2009). In this paper, the statistical package of R called 'Rpart' (THERNEAU & ATKINSON, 2015) is used.

6.3.5 Random Forest (RF)

A random forest (RF) is an ensemble learning method that combines multiple decision trees to create a more accurate and stable prediction model for continuous target variables. The model is known for being a robust non-parametric model capable of handling nonlinear, noisy, and multivariate correlated data (WANG et al., 2022). In this approach, each decision tree in the forest is trained on a random subset of the data and a random subset of the features. Each tree makes a prediction for a given input, and the final prediction of the random forest is the average of the predictions made by all the decision trees (YANG et al., 2016). The RF explores predictor importance through subsets of data and bootstrapped predictors for tree growth. Each node in the decision tree is split using the best predictor selected randomly at that node. For each bootstrap iteration, the RF identifies data not included (out-of-bag) using the tree grown in the bootstrap sample. After aggregating all out-of-bag (OOB) predictions, the overall error rate is estimated (BREIMAN, 2001). RF predicts variable importance by observing how the forecasting error increases when the out-of-bag data for that variable is changed, while others remain fixed (CATANI et al., 2013).

6.3.6 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are machine learning algorithms that have good generalization capability applying a structural risk minimization on a limited number of learning patterns (BAYDAROĞLU & KOÇAK, 2014). SVM is one of the machine learning methods utilized in hydrology and has proved to be an alternative to ANNs (MESHRAM et al., 2020). SVM is a supervised learning algorithm which seek to minimize model complexity and errors (NHU et al., 2020). The model was first introduced by Vapnik (1995). The goal of SVM is to find a hyperplane (a line or a plane in a higher-dimensional space) that separates the data points of different classes, or, in the case of regression, predicts the target variable. The hyperplane is chosen so that it maximizes the margin, which is the distance between the hyperplane and the closest data points of each class, known as support vectors. The data points closest to the hyperplane are called support vectors and have the greatest impact on the position

of the hyperplane (PHAM et al., 2016).

SVM can handle nonlinearly separable data by using kernel functions that map the data into a higher-dimensional space where a linear hyperplane can be used for separation (TEHRANY et al., 2014). An optimal separating hyper-plane is projected in the original space of n coordinates (x_i parameters in vector x) between the points of two distinct classes within a certain error limit (TEHRANY et al., 2014). Consider that x and y correspond to input variables and output variables, respectively. If $x_i \in R^n, y_i \in \{-1,1\}$ and $i = 1, \dots, n$, then the optimal separating hyper-plane is calculated using a classification decision function in Equation 21.

$$f(x) = \text{sgn}(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x_j) + b) \quad \text{(Eq. 21)}$$

where n is the number of input variables, α_i are Lagrange multipliers, $K(x_i, x_j)$ is the kernel function and b is the offset of the hyper-plane from the origin.

Ideally, the choice of kernel function should be made based on structure of the structure of the input data and its relation to the output (WANG et al 2022). In literature, there are different kernel functions such as linear, polynomial, sigmoid and radial basis (Gaussian) function (RBF). In the application of SVR, RBF is chosen as it is the best kernel function (Equation 22) according to past hydrometeorological studies (see ADNAN et al., 2020; BAESENS et al., 2000; LEONG et al., 2021) and it also represents the hydrometeorological processes, which have yielded good results.

$$K(x_i, x_j) = \exp(-\gamma x_i - x_j)^2 \quad \text{(Eq. 22)}$$

where γ controls the degree of nonlinearity of the SVM model. Small and large values of γ cause under- and over-fitting of the training data, respectively. For further details about SVMs, see Hamel (2009), and Steinwart and Christmann (2008), Tehrany et al. (2014).

6.3.7 Cross-validation in time series

Cross-validation is a statistical method used to evaluate the performance of machine learning models. Different parameters must be defined for each situation. Since the case studies addressed here refer to time series data, the approach of leave-one-out cross-validation

(LOOCV) is suitable for DT, RF and SVM. In this perspective, training is conducted in an iterative way, in which training, and validation sets are used. The process is carried out several times and in each iteration the training set contains more observations than in the previous iteration and the validation set has less. The corresponding training set consists only of observations that occurred prior to the observation that forms the test set. For the case studies addressed here, the trainControl function contained in caret package is used. The main parameters adopted are: (i) method equals time slice, (ii) initial window equals 90% of training set, (iii) 10% of training set to compose the validation set and (iv) fixed window equals true. In this process, the training window moves ahead with fixed horizon. The process is carried out until entire training set is used. The training performance is obtained by the average of performance measures of iterations. Further details can be obtained in Kuhn et al. (2020)

The LSTM model consists of two stacked LSTM layers and a single-unit fully connected layer. The LSTM layers have 50 units, the batch size is set to 6, and the number of epochs is 100. The network is optimized using Adamax, an adaptive moment estimation extension based on the infinity norm, and the Huber loss function. To prevent overfitting, 10% of the training data is used for validation. To capture longer dependencies, a stateful LSTM truncation layer is applied.

6.3.9 Chaos theory and machine learning

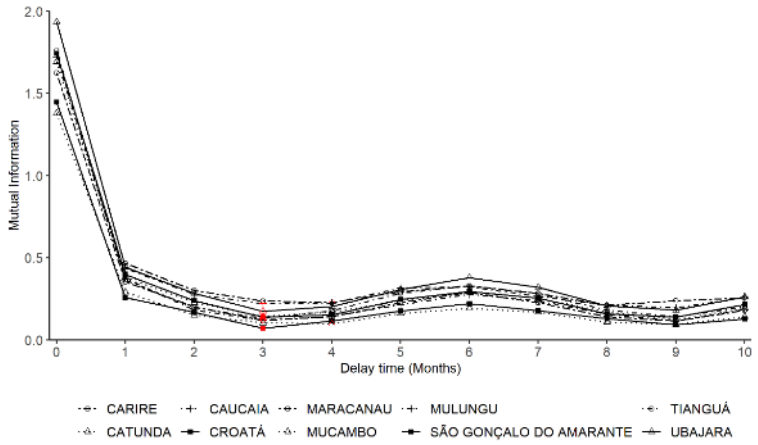
To predict the monthly rainfall, the MI is first used to identify the time delay and then a sequential procedure is used to estimate the embedding dimension m for PSR by trying out dimensions from 1 to 30. Each PSR is then fed to the proposed models in this study, and the fit of the models is evaluated using Root Mean Square Error (RMSE) and Nash-Sutcliffe Efficiency (NSE) metrics in the training phase. The reconstructed phase space with the dimension that results in a stabilization of the errors in most models is selected. The chosen dimension is then used to reconstruct the phase space and PSR is used as input to the forecast models. The training period is then divided into a training and validation phase to tune the parameters of the machine learning models. The forecasts are compared using numerical metrics such as RMSE, Mean Absolute Error (MAE), NSE and Correlation Coefficient (R), and a Taylor diagram is used to show the prediction skill of the models.

6.4 Results and discussion

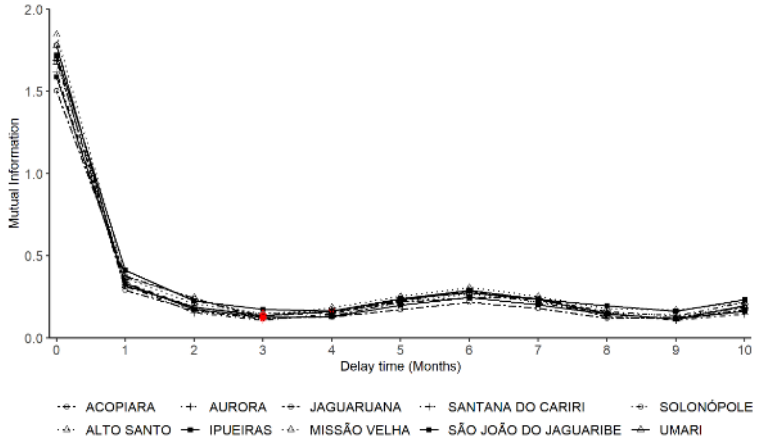
6.4.1 Phase-space reconstruction

The phase space was reconstructed with a delay time τ calculated using the mutual information method and embedding dimensions ranging from 1 to 30. Figure 30 illustrates the mutual information for various lag times for the stations analyzed in this study. The first minimum for most of the stations is at lag of 3, apart from Maracanaú, Catunda and Ipueiras stations which have a 4-month lag.

Figure 30 - Mutual information with delay time for monthly rainfall stations (a) 1 through 10 and (b) 11 through 20.



(a)



(b)

Source: Prepared by the author

All data was normalized to enable the machine learning models to reasonably

consider both extremely low and extremely high values in the feature phase-space reconstruction, regardless of their magnitude. The combination of delay time and varying embedding dimensions results in different prediction accuracy. The choice of dimensions for each station was based on how the model errors stabilized for most of the applied machine learning models according to the NSE and RMSE values. Table 11 describes the choose dimension for the training dataset based on the RMSE and NSE for each station. The results for the RMSE and NSE in the training dataset, based on the variation of embedding dimensions, are displayed in APPENDIX B.

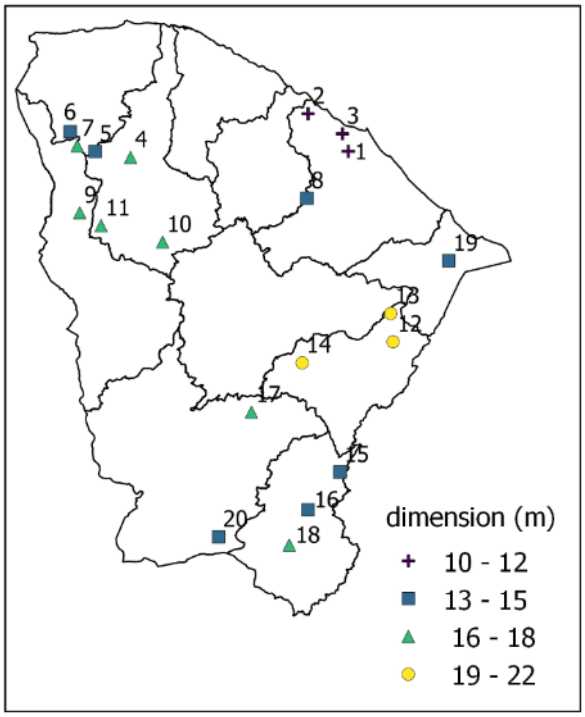
For most of the models, the selected embedding dimension ranged from 10 to 21 (Figure 31), suggesting that they need information that is older than 2-5 years in order to have a better fit of the models. This variability in time period is consistent with events such as El Niño, which are known to impact rainfall in the region (KAYANO & ANDREOLI, 2006). The RF and SVM methods showed limited improvement in prediction accuracy as the dimensions increased, as measured by the RMSE in APPENDIX B. However, for other methods, such as LAP, the RMSE varied significantly with changes in the embedding dimension. Spatial analysis of the chosen dimensions revealed that the coastal region of the state has lower embedding dimensions, while the central-eastern region has higher embedding dimensions, as illustrated in Figure 31.

Table 11 – Chosen embedding dimension for the training dataset according to the accuracy of the models.

Station	Chosen dimension	Station	Chosen dimension
MARACANAU	10	IPUEIRAS	17
SÃO GONÇALO DO AMARANTE	11	ALTO SANTO	21
CAUCAIA	13	SÃO JOÃO DO JAGUARIBE	21
CARIRE	17	SOLONÓPOLE	22
MUCAMBO	15	UMARI	15
TIANGUÁ	15	AURORA	15
UBAJARA	17	ACOPIARA	18
MULUNGU	14	MISSÃO VELHA	18
CROATÁ	17	JAGUARUANA	14
CATUNDA	17	SANTANA DO CARIRI	15

Source: Prepared by the author

Figure 31 - Spatial distribution of the chosen dimension for the rainfall stations according to the prediction accuracy.



Source: Prepared by the author

6.4.2 Comparisons of the results

In this study, the performance of various machine learning methods and a nonlinear prediction method for modeling rainfall time series is analyzed. To compare the model, the rainfall time series were divided into training (80%), validation (10%) and test (10%) phases. Table 12 presents the performance metrics for the analyzed stations during the validation period. The best scores on the validation sets are highlighted in bold. The results show that most models are capable of accurately modelling rainfall and providing information about the time series. The RF model performed the best during the validation period. The DT, SVM, and LSTM models had similar NSE values for most stations, while the LPA model had the worst performance across most stations.

Table 12 - Performance of machine learning models for monthly rainfall in the validation phase

Station	Chosen dimension and time delay	Model	MAE (mm)	RMSE (mm)	NSE	R
1 – Maracanau	$m = 10, \tau = 4$	DT	46.51	66.77	0.64	0.65
		LPA	103.05	135	-0.49	0.02
		LSTM	52.1	77.38	0.51	0.55
		RF	**22.67**	**35.08**	**0.9**	**0.92**
		SVM	43.95	64.55	0.66	0.66
2 – São Gonçalo do Amarante	$m = 11, \tau = 3$	DT	37.41	58.31	0.6	0.6
		LPA	151.52	164.09	-2.16	0.01
		LSTM	64.89	104.83	-0.29	0.48
		RF	**17.53**	**29.87**	**0.9**	**0.92**
		SVM	33.33	56.38	0.63	0.63
3 – Caucaia	$m = 13, \tau = 3$	DT	57.52	103.23	0.54	0.56
		LPA	137.11	179.3	-0.4	0
		LSTM	67.54	104.34	0.53	0.58
		RF	**31.56**	**63.6**	**0.82**	**0.9**
		SVM	57.29	110.02	0.47	0.53
4 – Carire	$m = 17, \tau = 3$	DT	31.47	53.53	0.68	0.68
		LPA	100.18	124.32	-0.73	0.04
		LSTM	44.23	77.62	0.32	0.39
		RF	**19.01**	**33.43**	**0.87**	**0.9**
		SVM	33.88	59.59	0.6	0.6
5 – Mucambo	$m = 15, \tau = 3$	DT	30.65	50.41	0.76	0.78
		LPA	106.29	131.87	-0.61	0.02
		LSTM	51.92	88.79	0.27	0.37
		RF	**16.68**	**30.86**	**0.91**	**0.94**
		SVM	34.88	62.54	0.64	0.65
6 – Tianguá	$m = 15, \tau = 3$	DT	53.76	89.81	0.55	0.57
		LPA	127.61	163.29	-0.49	0
		LSTM	58.32	104.16	0.4	0.47
		RF	**24.72**	**44.63**	**0.89**	**0.92**
		SVM	43.18	77.15	0.67	0.7
7 – Ubajara	$m = 17, \tau = 3$	DT	46.03	76.26	0.7	0.7
		LPA	146.12	177.34	-0.63	0
		LSTM	54.41	89.19	0.59	0.64
		RF	**18.37**	**31.19**	**0.95**	**0.96**
		SVM	38.33	67.43	0.76	0.77
8 – Mulungu	$m = 14, \tau = 3$	DT	62.55	95.21	0.34	0.38
		LPA	128.83	155.5	-0.75	0.12
		LSTM	78.41	118.91	-0.02	0.16
		RF	**30**	**50.11**	**0.82**	**0.91**
		SVM	31.42	62.07	0.72	0.77
9 – Croatá	$m = 17, \tau = 3$	DT	22.53	35.43	0.55	0.58
		LPA	71.67	77.97	-1.18	0.01

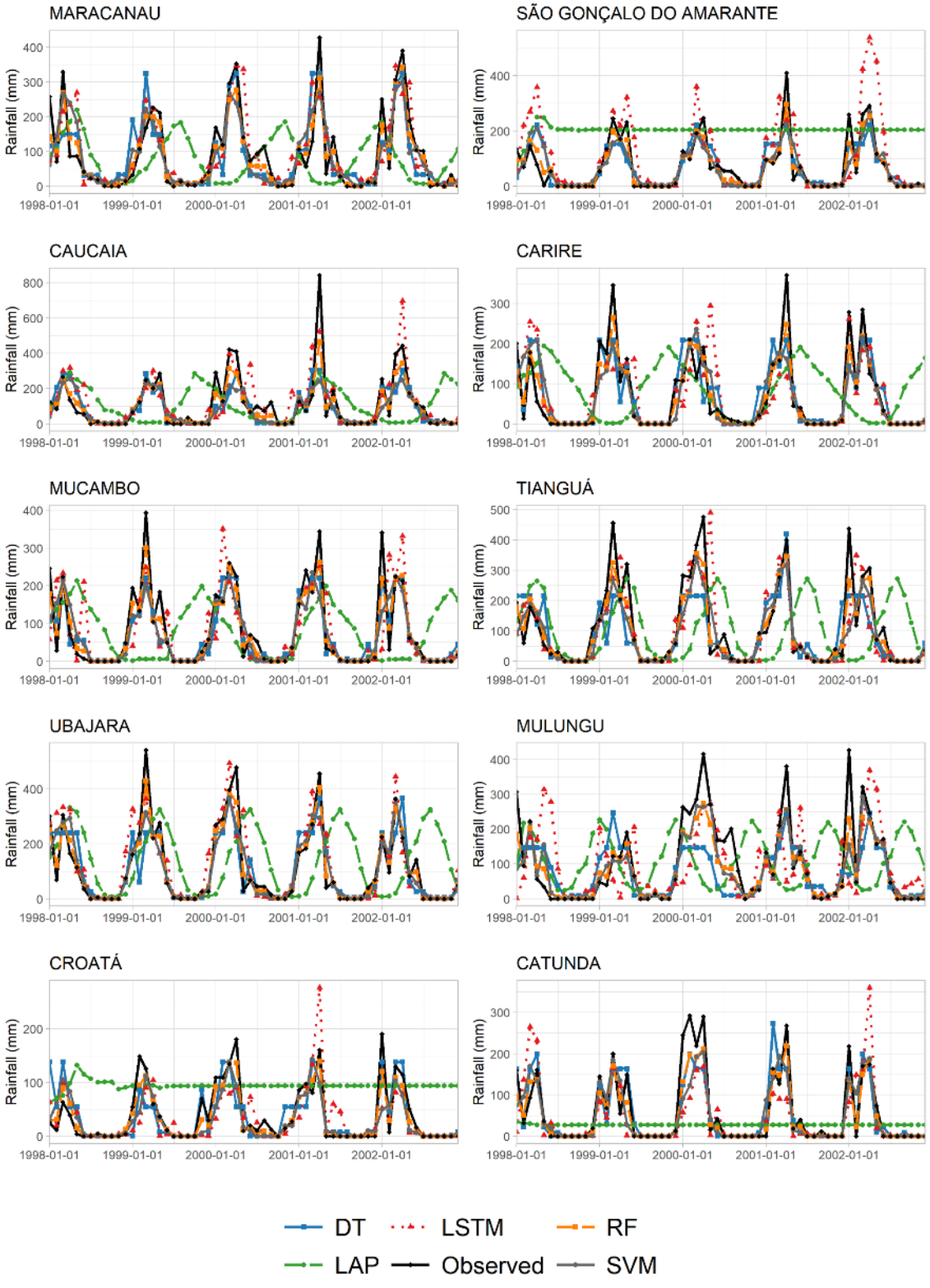
		LSTM	27.29	42.01	0.37	0.44
		RF	**10.29**	**17.26**	**0.89**	**0.92**
		SVM	19.83	34.3	0.58	0.6
10 – Catunda	$m = 17, \tau = 4$	DT	24.25	41.65	0.77	0.77
		LPA	61.89	92.49	-0.12	0.01
		LSTM	41.26	72.32	0.32	0.39
		RF	**16.58**	**30.47**	**0.88**	**0.92**
		SVM	27.26	49.47	0.68	0.69
11 – Ipueiras	$m = 17, \tau = 4$	DT	31.95	52.8	0.74	0.75
		LPA	74.06	109.04	-0.12	0.01
		LSTM	44.76	76.27	0.45	0.51
		RF	**20.59**	**35.48**	**0.88**	**0.92**
		SVM	39.18	66.78	0.58	0.61
12 – Alto Santo	$m = 21, \tau = 3$	DT	31.79	47.22	0.6	0.62
		LPA	69.69	92.35	-0.54	0
		LSTM	42.24	65.74	0.22	0.44
		RF	**15.24**	**24.47**	**0.89**	**0.92**
		SVM	31	49.54	0.56	0.57
13 – São João do Jaguaribe	$m = 21, \tau = 3$	DT	31.9	50.48	0.6	0.6
		LPA	54.75	75.34	0.1	0.18
		LSTM	45.21	78.89	0.02	0.27
		RF	**17.53**	**30.58**	**0.85**	**0.91**
		SVM	34.51	60.91	0.41	0.43
14 – Solonópole	$m = 22, \tau = 3$	DT	27.5	44.38	0.6	0.61
		LPA	36.6	49.57	0.5	0.52
		LSTM	37.17	69.32	0.02	0.43
		RF	**14.04**	**23.28**	**0.89**	**0.91**
		SVM	27.7	42.67	0.63	0.63
15 – Umari	$m = 15, \tau = 3$	DT	31.65	57.22	0.57	0.58
		LPA	102.02	111.42	-0.64	0.02
		LSTM	42.96	63.18	0.47	0.53
		RF	**16.19**	**30.58**	**0.88**	**0.93**
		SVM	33.58	60.78	0.51	0.52
16 – Aurora	$m = 15, \tau = 3$	DT	33.63	57.00	0.56	0.58
		LPA	96.68	122.39	-1.04	0
		LSTM	44.25	65.92	0.41	0.55
		RF	**16.51**	**26.80**	**0.9**	**0.92**
		SVM	30.13	49.11	0.67	0.68
17 – Acopiara	$m = 18, \tau = 3$	DT	34.37	63.32	0.34	0.37
		LPA	76.07	99.99	-0.65	0
		LSTM	35.9	60.29	0.40	0.48
		RF	**15.34**	**30.21**	**0.85**	**0.89**
		SVM	28.89	52.29	0.55	0.57
18 – Missão Velha	$m = 18, \tau = 3$	DT	37.4	58.27	0.61	0.62
		LPA	106.94	133.32	-1.04	0.02
		LSTM	49.26	72.02	0.41	0.48

		RF	**17.36**	**28.81**	**0.9**	**0.92**
		SVM	34.71	54.93	0.65	0.66
19 – Jaguaruana	$m = 14, \tau = 3$	DT	32.32	50.19	0.43	0.45
		LPA	55.9	86.63	-0.69	0
		LSTM	39.77	65.1	0.04	0.39
		RF	**13.16**	**23.41**	**0.88**	**0.9**
		SVM	27.64	48.47	0.47	0.49
20 – Santana do Cariri	$m = 15, \tau = 3$	DT	35.59	57.27	0.58	0.58
		LPA	89.17	118.76	-0.8	0
		LSTM	42.61	68.03	0.41	0.46
		RF	**14.91**	**27**	**0.91**	**0.93**
		SVM	31.03	52.55	0.65	0.65

Source: Prepared by the author

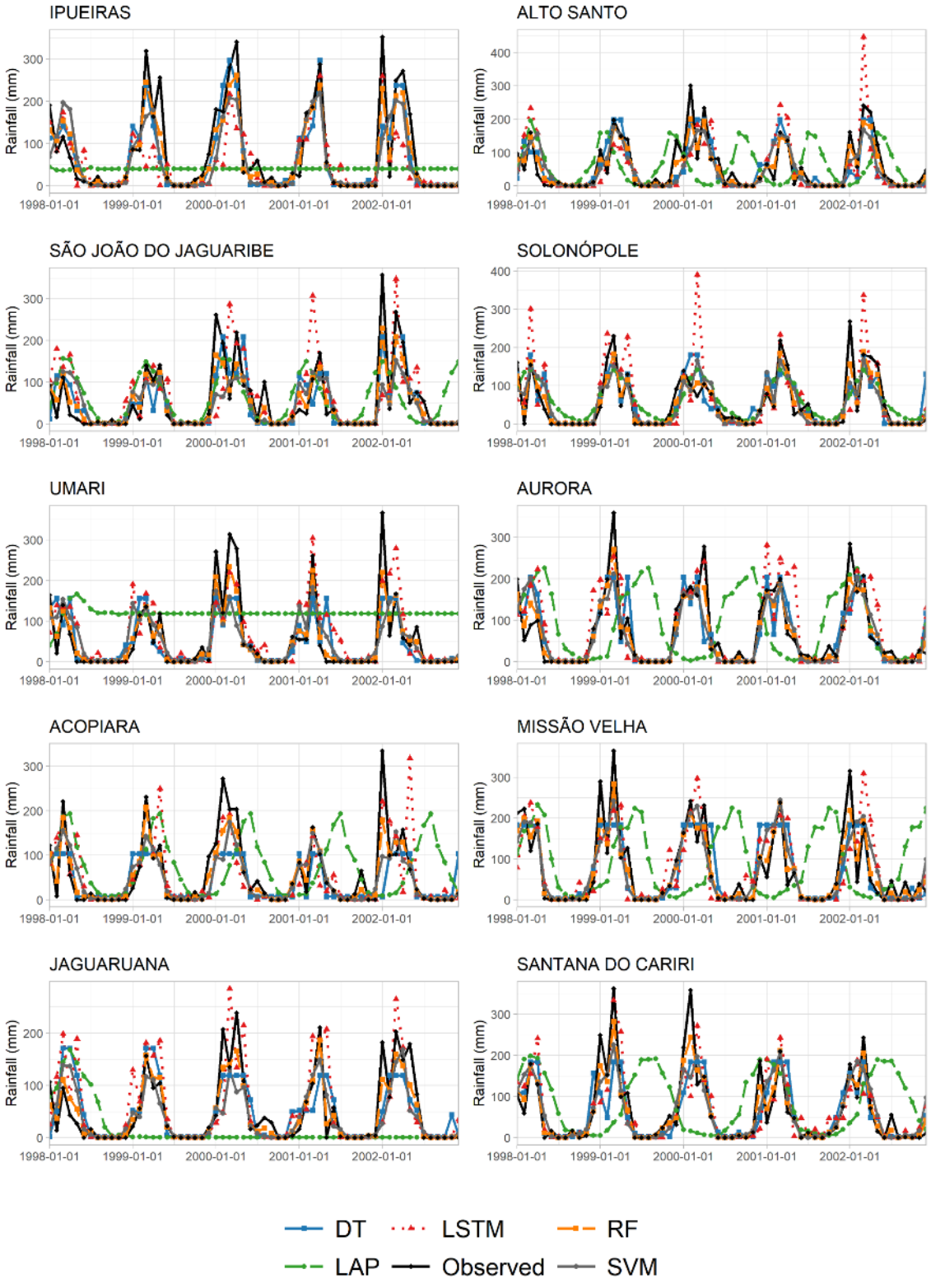
Figures 32 and 33 display the validation period of the analyzed datasets. It is observed that the LPA is able to track the monthly seasonality of the rainfall series for some stations, but not for others, such as the São Gonçalo do Amarante Station. The LSTM overestimates the high rainfall peaks for the year 2002 at stations such as São Gonçalo do Amarante, Caucaia, Catunda, Alto Santo, among others. Conversely, the SVM and DT models underestimate the high rainfall peaks.

Figure 32 - Comparison between measured and forecasted rainfall over the validation phase for stations 1 through 10.



Source: Prepared by the author

Figure 33 - Comparison between measured and forecasted rainfall over the validation phase for stations 11 through 20.



Source: Prepared by the author

The performance of rainfall time series modeling is analyzed using five machine learning models and a nonlinear prediction method. Table 13 reports the performance measures based on four metrics. The Ubajara Station achieved the highest accuracy in the test phase with an RMSE of 77.32mm and an NSE of 0.72 using SVM, followed by the Tianguá Station with an RMSE of 65.69mm and an NSE of 0.71 using RF. RF was the best-performing model in 9 stations, followed by SVM in 8 stations, DT in 2, and LSTM in 1. The LPA model performed poorly in all stations, as also observed in the validation period. For some stations the SVM and RF showed very similar values for the NSE and RMSE.

Table 13 - Performance of machine learning models for monthly rainfall in the testing phase

Station	Chosen dimension and time delay	Model	MAE (mm)	RMSE (mm)	NSE	R
1 – Maracanau	$m = 10, \tau = 4$	DT	69.97	101.92	0.35	0.37
		LPA	88.65	119.25	0.11	0.18
		LSTM	64.91	100.63	0.37	0.4
		RF	65.58	94.14	0.45	0.45
		SVM	**63.34**	**93.42**	**0.46**	**0.46**
2 – São Gonçalo do Amarante	$m = 11, \tau = 3$	DT	56.9	90.92	0.41	0.43
		LPA	92.89	122.77	-0.07	0.04
		LSTM	66.93	110.51	0.13	0.38
		RF	51.33	91.82	0.4	0.46
		SVM	**50.99**	**89.45**	**0.43**	**0.48**
3 – Caucaia	$m = 13, \tau = 3$	DT	70.46	113.97	0.46	0.47
		LPA	128.67	164.92	-0.12	0.05
		LSTM	87.3	131.93	0.28	0.33
		RF	69.19	110.92	0.49	0.53
		SVM	**69.55**	**116.25**	**0.44**	**0.58**
4 – Carire	$m = 17, \tau = 3$	DT	43.18	73.19	0.37	0.43
		LPA	93.78	120.74	-0.71	0
		LSTM	44.34	71.72	0.4	0.46
		RF	**35.36**	**64.59**	**0.51**	**0.53**
		SVM	37.23	68.74	0.45	0.49
5 – Mucambo	$m = 15, \tau = 3$	DT	38.18	59.77	0.59	0.62
		LPA	89.16	118.15	-0.6	0
		LSTM	43.17	83.76	0.2	0.46
		RF	**27.22**	**53.17**	**0.68**	**0.68**
		SVM	30.41	56.08	0.64	0.65
6 – Tianguá	$m = 15, \tau = 3$	DT	51.63	78.34	0.59	0.6
		LPA	95.15	133.97	-0.19	0.08
		LSTM	53.07	83.43	0.54	0.56
		RF	**38.13**	**65.69**	**0.71**	**0.74**
		SVM	40.3	68.82	0.69	0.7

7 – Ubajara	$m = 17, \tau = 3$	DT	60.32	92.8	0.6	0.62
		LPA	54.68	85.46	0.66	0.67
		LSTM	67.81	99.99	0.54	0.6
		RF	53.7	78.74	0.71	0.72
		SVM	**46.47**	**77.32**	**0.72**	**0.73**
8 – Mulungu	$m = 14, \tau = 3$	DT	59.48	85.11	0.39	0.41
		LPA	106.74	131.03	-0.44	0.01
		LSTM	56.49	89.24	0.33	0.38
		RF	**54.12**	**79.84**	**0.47**	**0.51**
		SVM	69.68	99.33	0.17	0.24
9 – Croatá	$m = 17, \tau = 3$	**DT**	**26.63**	**65.97**	**0.35**	**0.36**
		LPA	68.09	99.46	-0.47	0.07
		LSTM	36.24	82.58	-0.01	0.08
		RF	28.26	71.93	0.23	0.27
		SVM	27.02	71.35	0.24	0.29
10 – Catunda	$m = 17, \tau = 4$	**DT**	**39.86**	**67.49**	**0.47**	**0.49**
		LPA	62.39	98.18	-0.12	0.01
		LSTM	38.93	77.77	0.3	0.4
		RF	37.25	70.53	0.42	0.47
		SVM	36.74	67.91	0.46	0.48
11 – Ipueiras	$m = 17, \tau = 4$	DT	54.18	94.53	0.26	0.32
		LPA	78.01	118.34	-0.16	0.02
		LSTM	**35.58**	**71.37**	**0.58**	**0.62**
		RF	39.17	71.69	0.57	0.61
		SVM	38.97	71.14	0.58	0.62
12 – Alto Santo	$m = 21, \tau = 3$	DT	50.06	82.91	0.3	0.32
		LPA	80.22	107.15	-0.16	0.03
		LSTM	53.31	92.81	0.13	0.3
		RF	**39.74**	**73.2**	**0.46**	**0.52**
		SVM	37.81	72.63	0.46	0.5
13 – São João do Jaguaribe	$m = 21, \tau = 3$	DT	47.1	76.7	0.28	0.32
		LPA	67.72	93.51	-0.07	0.06
		LSTM	45.31	78.05	0.26	0.32
		RF	**38.83**	**67.58**	**0.44**	**0.46**
		SVM	38.07	69.44	0.41	0.43
14 – Solonópole	$m = 22, \tau = 3$	DT	51.7	90.61	0.07	0.17
		LPA	56.87	87.81	0.13	0.14
		LSTM	50.5	88.41	0.12	0.22
		RF	43.42	77.86	0.32	0.36
		SVM	**42.83**	**77.87**	**0.32**	**0.35**
15 – Umari	$m = 15, \tau = 3$	DT	39.38	61.42	0.37	0.4
		LPA	57.58	78.67	-0.04	0.04
		LSTM	51.15	84.38	-0.2	0.46
		RF	31.61	50.17	0.58	0.58
		SVM	**31.09**	**49.27**	**0.59**	**0.60**
16 -	$m = 15, \tau = 3$	DT	41.19	64.75	0.58	0.60

Aurora		LPA	98.14	129.27	-0.67	0.00
		LSTM	42.67	72.32	0.48	0.53
		RF	**35.32**	**59.32**	**0.65**	**0.68**
		SVM	37.51	60.48	0.64	0.66
17 - Acopiara	$m = 18, \tau = 3$	DT	46.58	79.87	0.09	0.15
		LPA	48.96	80.69	0.07	0.20
		LSTM	47.5	79.82	0.09	0.25
		RF	37.04	68.93	0.32	0.33
		SVM	**37.54**	**68.91**	**0.32**	**0.34**
18 – Missão Velha	$m = 18, \tau = 3$	DT	58.59	99.44	0.42	0.53
		LPA	110.39	151.7	-0.36	0.00
		LSTM	73.47	117.41	0.19	0.24
		RF	55.97	96.98	0.44	0.58
		SVM	**52.73**	**93.3**	**0.49**	**0.60**
19 - Jaguaruana	$m = 14, \tau = 3$	DT	49.61	86.32	0.28	0.29
		LPA	72.34	106.68	-0.11	0.02
		LSTM	46.6	87.95	0.25	0.33
		RF	**43.02**	**79.51**	**0.38**	**0.41**
		SVM	46.93	81.13	0.36	0.40
20 – Santana do Cariri	$m = 15, \tau = 3$	DT	46.03	92.96	0.41	0.42
		LPA	109.77	150.91	-0.57	0.02
		LSTM	45.48	84.62	0.51	0.51
		RF	**42.4**	**83.12**	**0.52**	**0.60**
		SVM	43.56	89	0.45	0.51

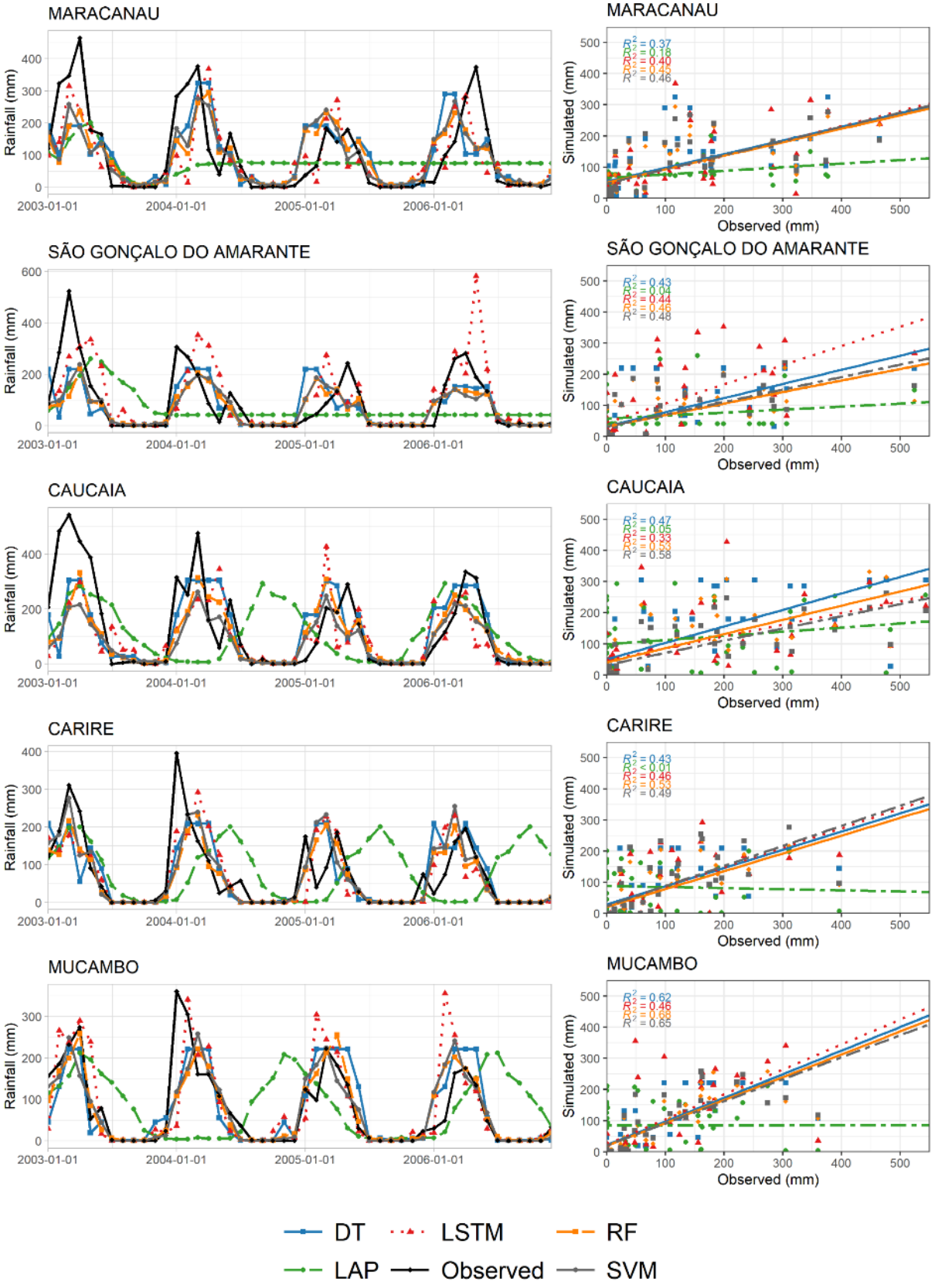
Source: Prepared by the author

The accuracy of the models was evaluated and assessed. A direct comparison of the models is shown in Figures 33 through 36. Based on these figures, the LPA performed poorly and was not able to accurately forecast rainfall for the stations. Most of the models generally followed the seasonal patterns of the precipitation series for most stations. However, some differences in performance in relation to model complexity and hydrological regimes were also evident. The LSTM tended to capture the high peaks of rainfall, but at the expense of some overestimation, particularly for the Umari, São Gonçalo do Amarante, and Alto Santo stations. The simpler model, RF, captures the seasonal behavior in rainfall but tends to underestimate the magnitude of the monthly peaks. Hagen et al. (2021) found similar results in their analysis of snowmelt-driven flood regimes. Comparing the peaks in these rainfall forecast regimes shows that increasing model complexity does not always lead to more accurate predictions and distinct peaks. As stated by Papacharalampous et al. (2019), it is often argued that the best methodology is determined by the data and the case study, as evidenced in our results. Furthermore, the small

decrease in the evaluation metrics for the validation and testing phase for the SVM, RF, and LSTM models could indicate that these models are more robust.

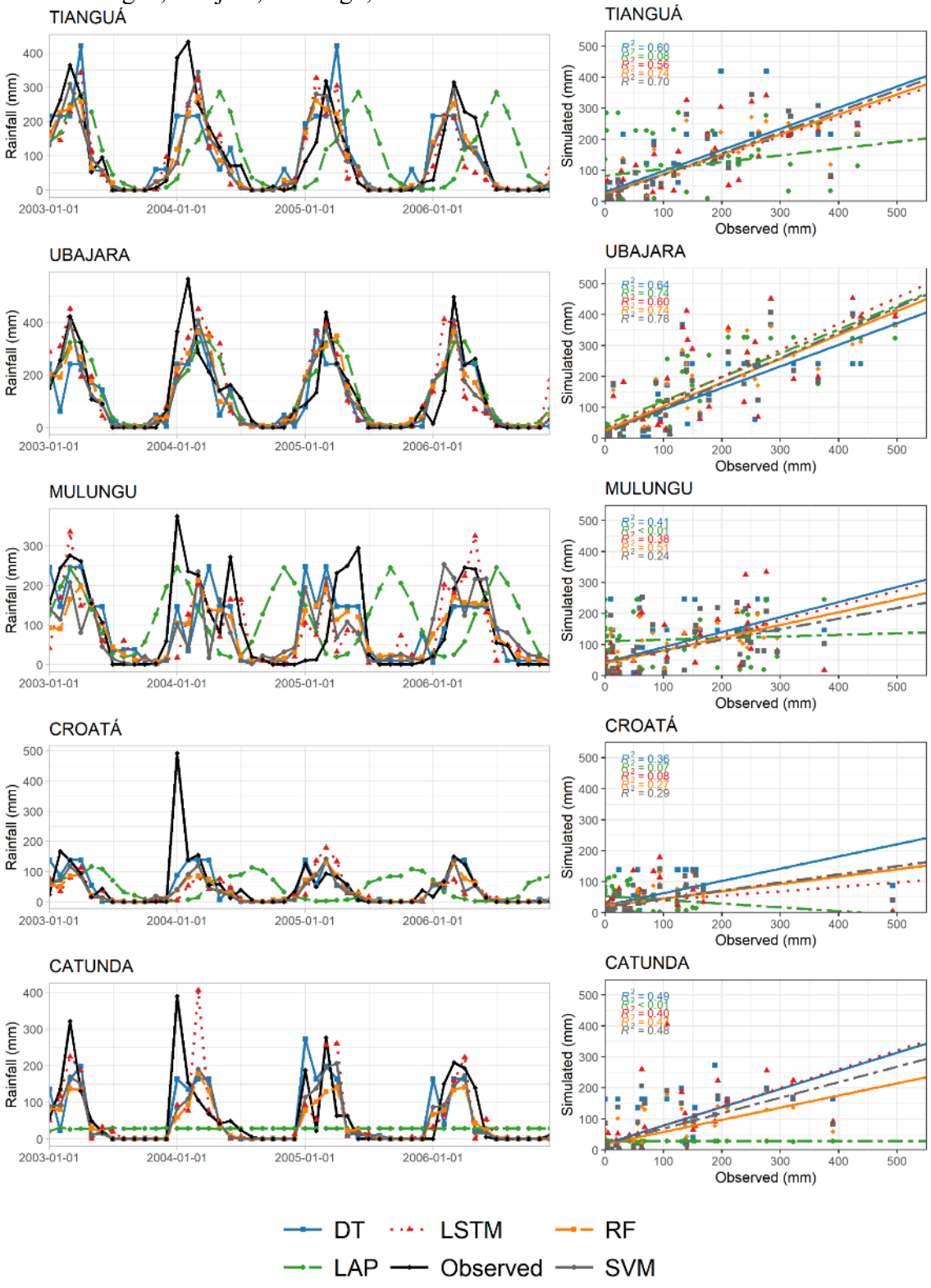
As stated by Lellep et al. (2020) and confirmed by our results, machine learning methods can provide a useful set of tools for predicting rainfall time series if the underlying dynamics, such as increasing complexity with prediction time and fragmentation of the state space, are taken into account. When these conditions are met, machine learning tools have great potential for various applications.

Figure 34 - Comparison between measured and forecasted rainfall and best fit lines by the DT, LAP, LSTM, RF and SVM for the best accurate input combinations over the testing phase for the Maracanu, São Gonçalo do Amarante, Caucaia, Carire and Mucambo station.



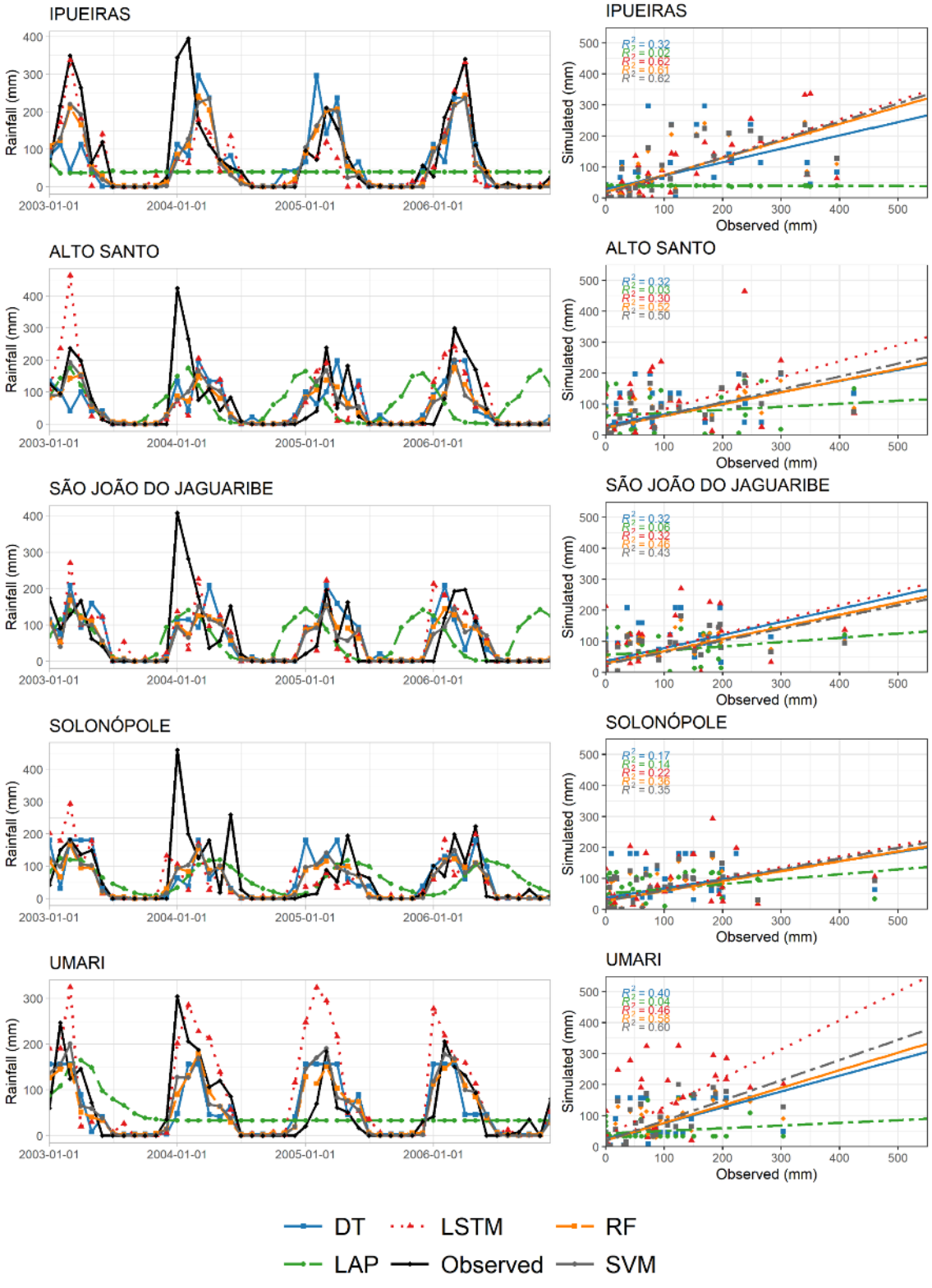
Source: Prepared by the author

Figure 35 – Comparison between measured and forecasted rainfall and best fit lines by the DT, LAP, LSTM, RF and SVM for the best accurate input combinations over the testing phase for the Tianguá, Ubajara, Mulungu, Croatá and Catunda stations.



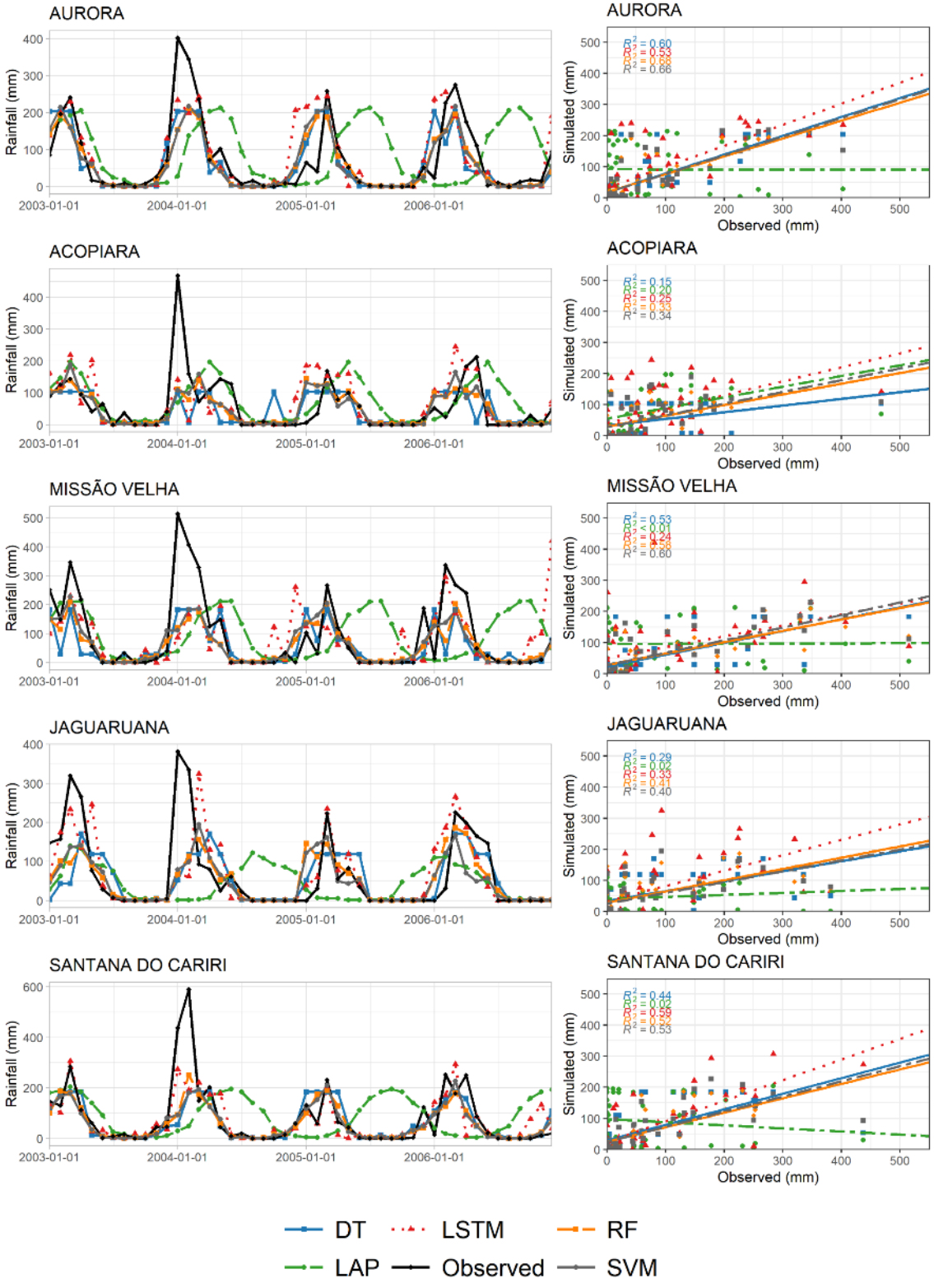
Source: Prepared by the author

Figure 36 – Comparison between measured and forecasted rainfall and best fit lines by the DT, LAP, LSTM, RF and SVM for the best accurate input combinations over the testing phase for the Ipueiras, Alto Santo, São João do Jaguaribe, Solonópole and Umari stations.



Source: Prepared by the author

Figure 37 - Comparison between measured and forecasted rainfall and best fit lines by the DT, LAP, LSTM, RF and SVM for the best accurate input combinations over the testing phase for the Aurora, Acopiara, Missão Velha, Jaguaruana and Santana do Cariri stations.

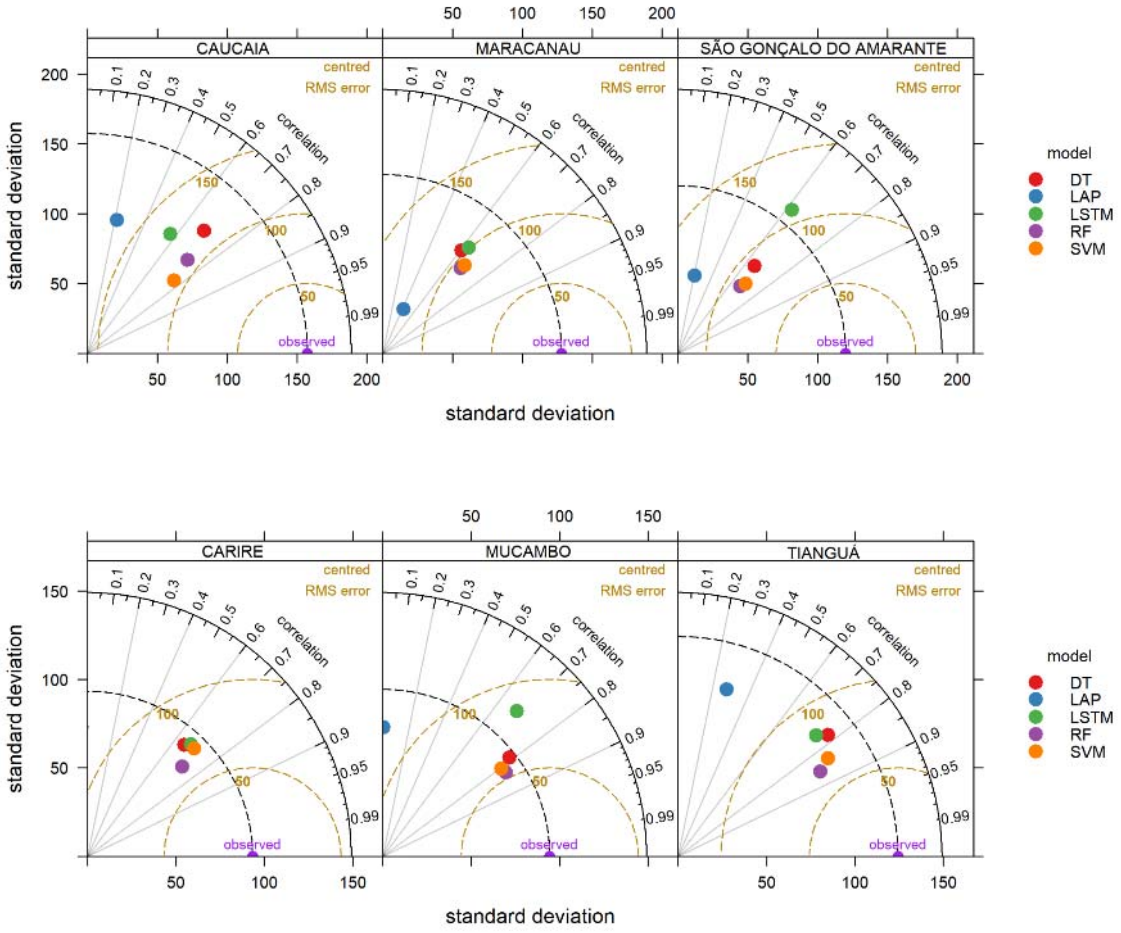


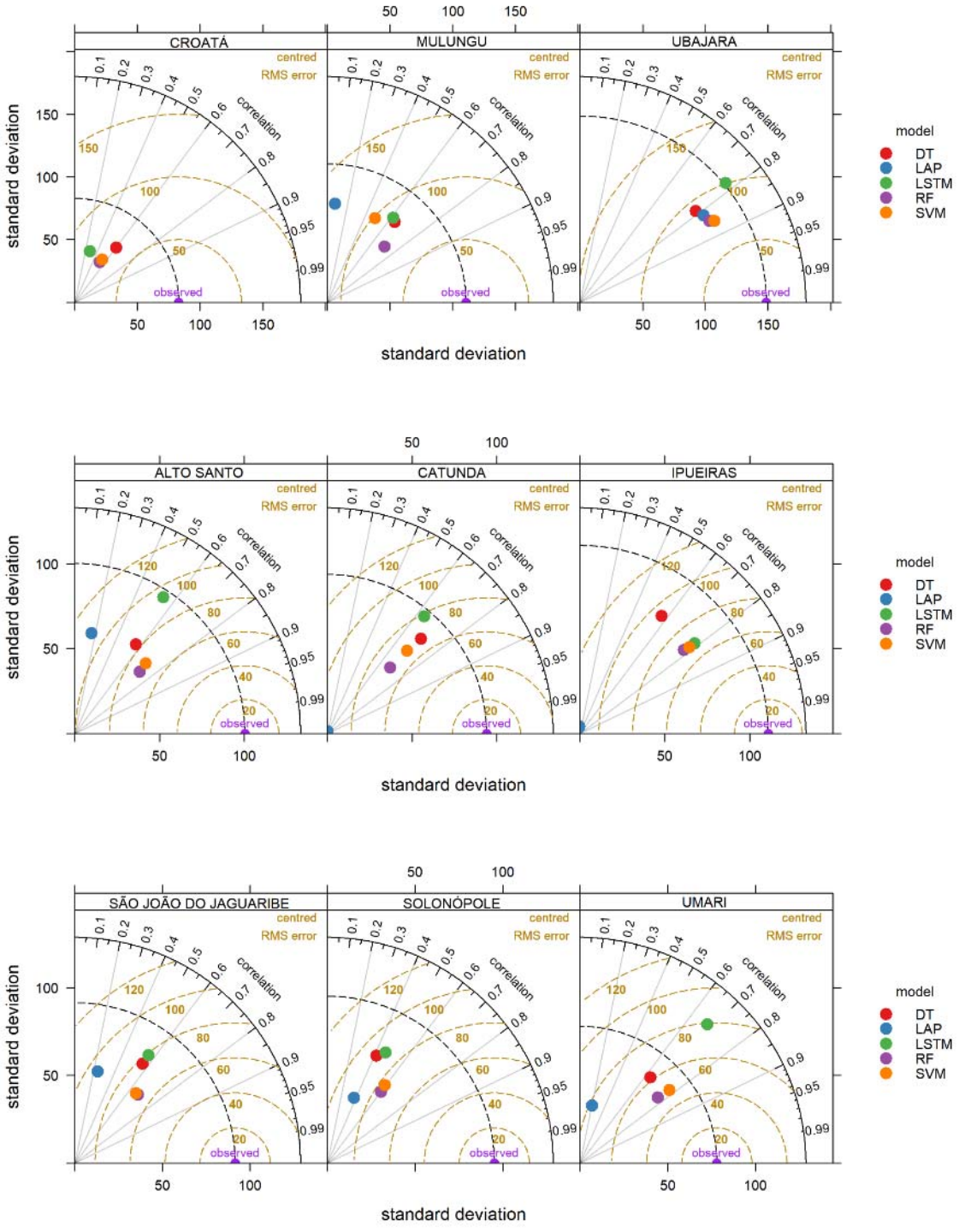
Source: Prepared by the author

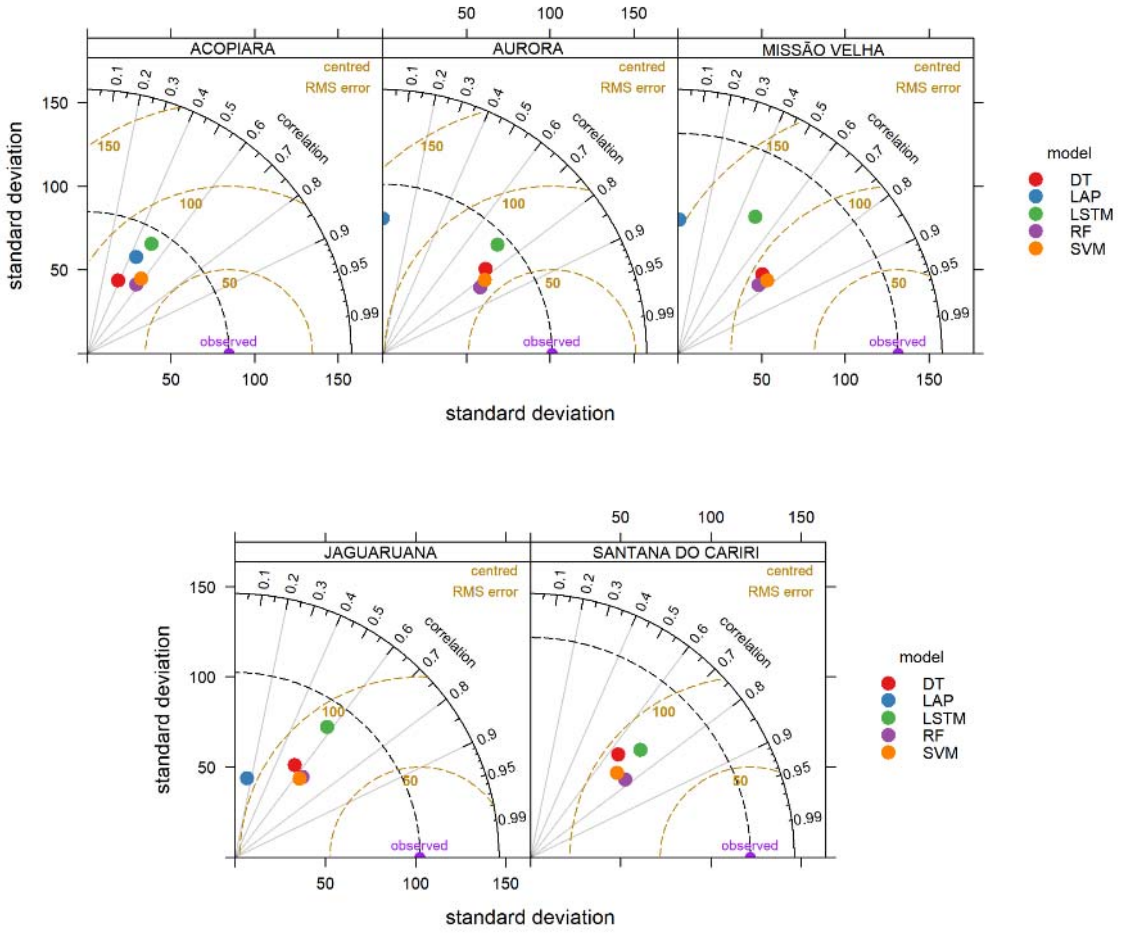
We also used the Taylor diagram to evaluate the models, as shown in Figure 38. The Taylor diagram is a graphical tool for comparing a simulation or set of cases to observations,

evaluated in terms of their correlation coefficient and standard deviation. The distance from the reference point (observations) represents the focused RMSE (TAYLOR, 2001). A perfect model that matches the observations is located at the reference point, with a correlation coefficient of 1 and similar error distribution compared to the observations (HEO et al., 2014). Among the stations, the SVM model stood out as it had similar variability to the observations and the highest correlation for most stations.

Figure 38 - Taylor diagram of the forecasted rainfall over the testing period for all stations.




CROATÁ
MULUNGU
UBAJARA
ALTO SANTO
CATUNDA
IPUEIRAS
SÃO JOÃO DO JAGUARIBE
SOLONÓPOLE
UMARI
centred
RMS error
correlation
observed
standard deviation
model
DT
LAP
LSTM
RF
SVM



Source: Prepared by the author

6.5 Conclusions

The purpose of this study is to evaluate the effectiveness of machine learning models on monthly rainfall time series in the Ceará State, in Northeastern Brazil. The study takes into account the principles of chaos theory and employs various machine learning techniques to analyze the rainfall dynamics. The monthly rainfall data from 20 hydrological stations in the region were used as input. The data was divided into training and testing windows, and the embedding dimension was chosen based on the accuracy of the models during the training phase. Upon analyzing rainfall data, we found distinct rainfall patterns at each location, requiring different models for optimal forecasting. Most stations showed embedding dimensions greater than 10, necessitating information from a period as long as five years for accurate prediction.

The SVM and RF models performed well in many of the stations, proving the viability of data-driven methodologies in capturing rainfall dynamics without extensive

physical information. However, it's important to note a key limitation in the capability of some models in our study to adequately capture extreme rainfall events.

The results underscore the potential of machine learning techniques in rainfall forecasting, aiding in resource management. Future research should focus on enhancing model performance for rainfall broadening the scope to include other climatic variables. The incorporation of physical data could also enrich the predictive power of these models, bridging data-driven and physics-based approaches.

7 A MULTI-MODEL FRAMEWORK FOR STREAMFLOW FORECASTING BASED ON STOCHASTIC MODELS: AN APPLICATION TO THE STATE OF CEARÁ, BRAZIL[2].

7.1 Introduction

Accurate and reliable long-term streamflow prediction at monthly, annual, inter-annual, or even decadal scales can be a valuable tool for optimal allocation and management of water systems (LIANG et al., 2018; SINGH, 2016; WANG et al., 2019). Particularly in systems where decadal-scale variability is prominent, and under stress due to surface and groundwater scarcity (SZOLGAYOVA et al., 2014). However, the forecast of hydrological variables is a major challenge due to its nonlinear, nonstationary, overly complex processes and multi-scale characteristics, which can be directly affected by climate change and land use (NAZIR et al. 2019; WANG et al., 2019; WEN et al., 2019). Recent attention has focused on long lead-time streamflow prediction to improve the operation of the hydroelectric system (ROLIM & SOUZA FILHO, 2020), drought forecasting (KHAN et al., 2020; KISI et al., 2019), and other practical water activities (DARIANE et al., 2018). Therefore, short- and long-term streamflow prediction (3-15 years) can be used to meet water demand over extended periods, being detrimental to water resource management (ERKYIHUN et al., 2017; MENG et al., 2019; MISHRA & SINGH, 2010; WEN et al., 2019).

Several approaches have been developed to simulate and predict streamflow time series at longer lead-time. These approaches can be categorized as data-driven and process-driven models (HE et al., 2014; ZHANG et al., 2018). Typically, physical models require complex model structures that need knowledge of physical processes, require extensive calculations, precise, and a great quantity of meteorological and hydrological data (LIANG et al., 2018; YANG et al., 2020). Physical models consider the physical mechanism of hydrological processes. Also, these models have been widely applied to hydrological series. However, there is still a great discussion around those models mainly centered around the parameterization process, uncertainty led by data limitations, and computational constraints on model analysis (CLARK et al., 2017). Instead, data-driven models have attracted considerable attention because of their simple formulation, the need for much less data, rapid simulations,

---

[2] Reproduced with permission from Springer Nature. This version of the article has been accepted for publication and is available online at: https://doi.org/10.1007/s41101-023-00184-1.

and versatility in modeling nonlinear processes compared to physical models (MENG et al., 2019; NAZIR et al., 2019; WEN et al., 2019). Furthermore, physical models need information and extensive data for a complete analysis, which can be an expensive computational process, whereas data-driven models do not require such physical constraints and can accurately represent nonlinear and nonstationary processes, e.g., the temporal variability in streamflow (SARAIVA et al., 2021).

One of the main types of data-driven models is the traditional stochastic model approach, which success rate has always encouraged its implementation over physics-based models (REMESAN & MATHEW, 2016). Stochastic models were extensively developed since the 1960s, beginning with the original generation of synthetic streamflow (THOMAS & FIERING, 1962). Studies mostly focused on the applications of parametric methods during the 1980s and early 1990s. Autoregressive models (AR) and autoregressive moving average (ARMA) assume the series to be cyclostationary or stationary, however, these assumptions are not likely to reproduce the features of long-term streamflow time series (SALAS & OBEYSEKERA, 1982). According to Erkyihun et al. (2017), short- and long-term forecasts (1–20 years) can be made through a fitted AR model. Nevertheless, the assumption of stationarity in AR models limits their capability to capture nonstationary processes. Later, models like auto-regressive with exogenous inputs (ARX) and auto-regressive moving average with exogenous inputs (ARMAX) also showed satisfactory time series prediction skills and fairly easy implementation. Studies also show that stochastic models can perform as well as machine learning (ML) methods (KOUTSOYIANNIS & GEORGAKAKOS, 2008; PAPACHARALAMPOUS et al., 2017; PAPACHARALAMPOUS et al., 2019). Further-more, as record lengths have increased, modelers became aware of decadal-scale variability. Hence, improved modeling and prediction methods have been developed, such as nonparametric time series simulation models, singular spectrum analysis (SSA) to analyze the temporal variability of temperature time series (DETTINGER et al., 1995), a semiparametric approach for forecasting streamflow at multiple gaging locations on climate precursors (SOUZA FILHO & LALL, 2003), and wavelet-based models as the wavelet autoregressive model (WARM) introduced by Kwon et al. (2007).

Many hybrid algorithms, which are data preprocessing techniques combined with forecast models, have been applied to incorporate nonstationary and nonlinear features to hydrological time series (CHOU & WANG, 2004; HUMPHREY et al., 2016; MENG et al., 2019; WANG et al., 2019). Many multi-resolution tools have been applied to decompose signals,

such as the Fourier transform window function, the wavelet trans-form (WT), singular spectrum analysis, and principal component analysis. WT has recently become more widely applied in hydrology because it overcomes the major limitations of Fourier analysis, such as the lack of time-frequency localization and infinite domain of sine and cosine waves (CHOU & WANG, 2004; ERKYIHUN et al., 2017; GUO et al., 2011; KASIVISWANATHAN et al., 2016; KWON et al., 2007; MENG et al., 2019; NOURANI et al., 2011, 2009; PENG et al., 2017). Although WT addresses nonlinearity and nonstationary successfully, drawbacks in the method's performance remain (e.g., proper choice of mother wavelet and optimal processing threshold for the frequencies to be modeled).

Alternative methods such as the empirical mode decomposition (EMD) (HUANG et al., 1998), ensemble EMD (EEMD) (WU & HUANG, 2009), and the complete ensemble EMD with adaptive noise (CEEMDAN) (TORRES et al., 2011) have been proposed to solve these issues related to multi-resolution data analysis. The methods do not require a pre-determine frequency band by the user, therefore, the process is self-adaptive. The set of EMD is a data-driven tool that decomposes the nonstationary and nonlinear data through an adaptive process into several oscillatory components called Intrinsic Mode Functions (IMF) and a residue. In recent decades, these methods became popular to analyze hydro-meteorological time series especially coupled with other ML methods such as support vector machines and extreme learning machines (ADARSH & REDDY, 2018; MENG et al., 2019; NAZIR et al., 2019; WEN et al., 2019).

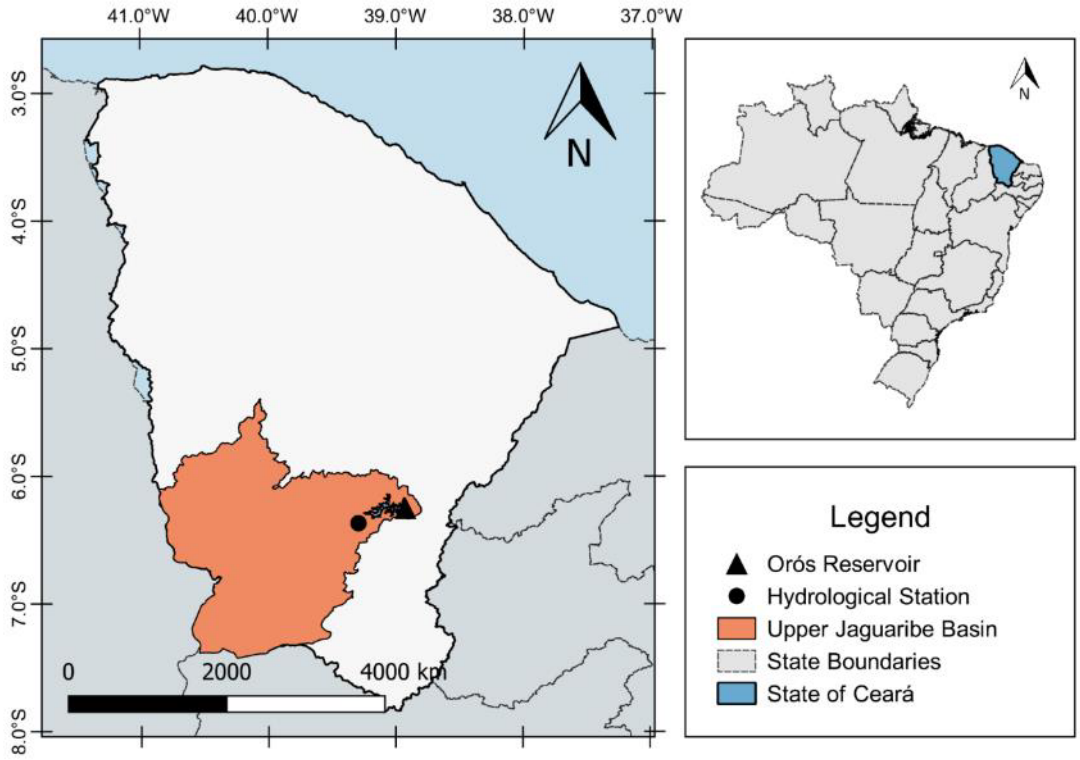
Reliable hydrological forecasting depends on accurate predictions in a longer lead time. Thus, the present study primary objective is to propose a multi-model framework including hybrid models to predict long-term average annual streamflow (3-15 years). Many recent studies have focused on applying preprocessing techniques with data-driven models to improve long-term forecasts. Nonetheless, despite several studies on these models' performance, no conclusive decision has been made about which model performs best in a longer lead time for streamflow prediction. It is often argued that it is the data and the case study that decide the 'best' methodology (PAPACHARALAMPOUS et al., 2019). Furthermore, many of the developed methods require a large amount of data. However, using annual streamflow makes it easier to model in data-scarce regions. We seek the simplification of the proposed multi-model using the LASSO (Least Absolute Shrinkage and Selection Operator) method to reduce the subsample of the multi-model and incorporate models that account for uncertainty and consider the gains that each model can bring in the prediction analysis. Also, we will evaluate these

hybrid approaches separately. The hybrid selected models were the wavelet autoregressive model (WARM), wavelet hidden Markov model (WHMM), complete ensemble empirical mode decomposition with adaptive noise autoregressive model (CEEMDAN-AR), and complete ensemble empirical mode decomposition with adaptive noise hidden Markov model (CEEMDAN-HMM). The models' forecast skills were measured for different forecast windows (3, 5, 10, and 15 years) to evaluate model performance and improvement for short-term and decadal-scale prediction for water resources planning and management applications.

7.2 Study area and hydrological data

The analyzed hydrological station is located in the Upper Jaguaribe basin. The Basin is located in the south of the State of Ceará, Brazil, and it is home to approximately half a million people. The Basin's area extends over more than 24,000 km². The basin's outlet is controlled by the Orós reservoir, which has a storage capacity of 1.94 billion m³ and was built in 1961. The annual streamflow data from Iguatu hydrological station was used in this study (Figure 39). The station has monthly streamflow data covering January 1937 through December 2016. The series has 21 months with missing data. The statistical metrics of the annual streamflow at the Iguatu station are shown in Table 14. The rainfall regime is mainly governed by the Intertropical Convergence Zone, with average annual rainfall in the region of around 780 mm. In contrast, the region presents high rates of annual evaporation (2500 mm). The high temporal variability in rainfall with the high rates of evaporation leads to ephemeral or intermittent rivers in the region (Gaiser et al. 2003; Malveira et al. 2011). The area is highly affected by the low water levels due to recurrent droughts (Lima Neto et al. 2011) Therefore, streamflow is an essential variable whose forecasts are of considerable implication to water allocation in this region.

Figure 39 - Location of the meteorological stations analyzed in this study.



Source: Prepared by the author

Table 14 - Statistical information of streamflow data observed at the hydrological station.

Max (m^3/s)	Min (m^3/s)	Median (m^3/s)	Mean (m^3/s)	Standard Deviation (m^3/s)	Variation Coefficient
172.20	0.00	11.58	21.73	30.76	1.41

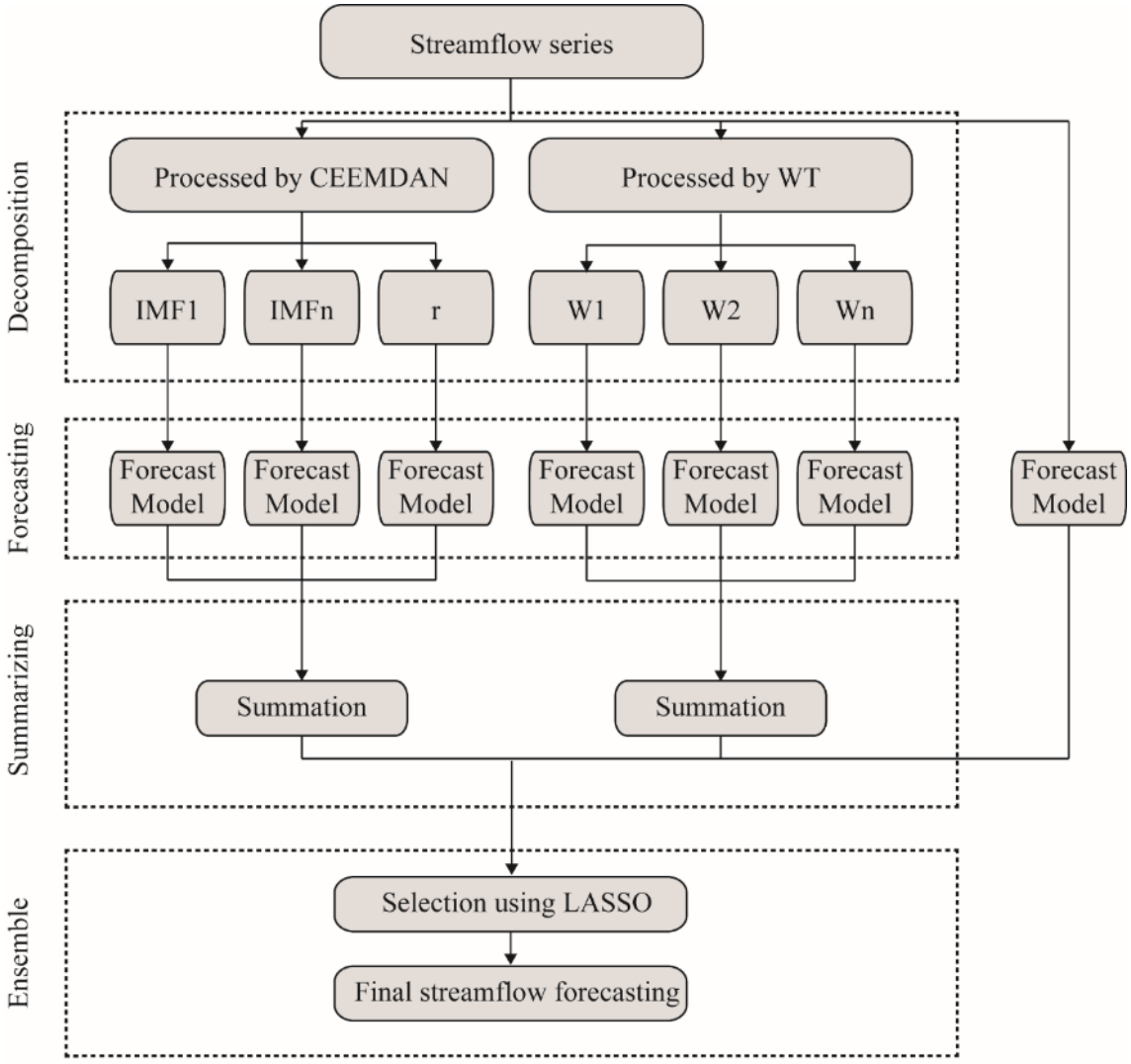
Source: Prepared by the author

7.3 Methods

The multi-model's schematics proposed to predict long-term average annual streamflow (3-15 years) and its comparative counterpart models are illustrated in Figure 40. First, different filtering techniques (WT and CEEMDAN) are applied to the streamflow time series to decompose the series into the pre-determinate frequencies and IMFs, respectively. The AR and HMM are used to forecast each WT decomposition and IMFs components and their respective residues datasets. In the prediction step, the procedure is repeated to generate an ensemble of streamflow forecasting time series according to the desired projection length. Next, the forecast series are summed. Finally, the multi-model was fitted using a regression model to

choose the models that would account for a more accurate prediction of long-term streamflow. The multi-model performance was compared with all the models analyzed to verify its performance in considering the uncertainties present in the streamflow time series. Their performance was evaluated using correlation coefficient (R), root mean squared error (RMSE), and mean absolute error (MAE). The methodology is detailed below.

Figure 40 - The overall processes of the multi-model methodologies applied in this study. The original series was decomposed using the CEEMDAN into n IMFs and a residue term (r) and using the WT into n daughter wavelets.



Source: Prepared by the author

7.3.1 Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

The CEEMDAN decomposition is an adaptation of the EMD. In the EMD method, the model separates time series into several IMFs, which represent the frequency modes and a residue. To decompose the time series, the model must satisfy the admissibility conditions, which are: (i) the mean value of the upper and lower bounds of the IMFs is zero at any point, and (ii) the number of extrema and zero crossings in the signal differ at most by one. A complete description of the EMD method can be found in Huang et al. (HUANG et al., 1998).

To overcome some limitations exhibited in the EMD process, Wu & Huang (2009) proposed an ensemble approach called EEMD, which decomposes a signal by summing different white noise to the original signal, and then the final modes are defined as averages of the IMFs obtained through EMD over an ensemble of trials. Limitations, such as mode mixing issues in the reconstructed series, may still prevail. Thus, a new method named CEEMDAN was proposed by Torres et al. (2011). In the CEEMDAN method, the decomposition process starts via the EMD method, where the series is separated into several IMFs and residues until a stopping criterion is attained. The method requires a controlled noise to be added at each stage of the decomposition with a singular residue for each IMF (ANTICO et al., 2014).

In the CEEMDAN, the level of added noise used is fixed to obtain all modes. Previous studies suggested that enough ensembles must be applied (REN et al., 2015). In this study, the number of ensembles is set as 100. The CEEMDAN process is enabled using the package hht (BOWMAN & LEES, 2013) in the software R.

7.4.2 Wavelet Transform

Wavelet transform is a widely applied method for the periodic phenomenon in nonstationary time series. The wavelet transform is a local transformation of frequency and time through multi-scale analysis using scaling and translation (MENG et al., 2019; TORRENCE & COMPO, 1998). The method divides the time series into a subset of continuous or discrete wavelets, and each sub-signal plays a different role and has a unique behavior. A more detailed revision can be found in Labat (2005) and Sang (2013).

WT has been used as a decomposition tool for analyzing streamflow time series in many studies (DANANDEH et al., 2013; KWON et al., 2007; MENG et al., 2019; PATHAK et

al., 2016; SUN et al., 2019). The Morlet wavelet was applied to the time series, and the package WaveletComp in R (ROSCH & SCHMIDBAUER 2016) was used in this study.

7.3.3 Hidden Markov Model

A hidden Markov model is a stochastic model used to represent dependencies between successive hidden states. Each state is associated with a probability distribution to the observed states. The model is comprised of two parts: an unobserved process that satisfies the Markov property; and a state-dependent process whose distribution depends only on the current state and not on previous states or observations.

Transitions among the hidden states are ruled by probabilities denoted as transition probabilities and are represented by the matrix **A**(t). The prior probabilities (i.e., initial conditions) π need to be defined to estimate the transition's parameters in **A**.

The HMM consists of a triplet of observed probability parameters (**A**,B,π), which are the state transition probability matrix, the parameter vector of the response models, and the initial state probabilities. The expectation-maximization algorithm (EM) is used to determine the optimal model parameters through an iterative process. The EM algorithm is used to establish the parameter of the model which maximizes the likelihood function in HMMs. To decode observation sequences into hidden state sequences, the Viterbi algorithm is applied. Further details can be found in Zucchini et al. (2016).

7.3.4 Multi-Model Framework

The CEEMDAN and WT were used to decompose the average annual streamflow series into pre-determinate frequencies and IMFs for the training and validation period (Table 15). Then, the respective decomposed series were modeled by HMM and AR models. The prediction with the HMM was based on the relation of the current states and the transition probability matrix. Based on each IMF or WT state at current and past times with the transition probability matrix, a component state is generated for time t+1, and a value from the corresponding state probability density function is sampled for time t+1. A random noise based on the residual of the fitted models was added to the prediction value of each decomposition component. This process is repeated to generate an ensemble of streamflow sequences for the

chosen lengths (ERKYIHUN et al., 2017). Next, the outputs of the subseries are summed up. Further, the CEEMDAN and WT are also coupled with an AR model.

The medians of the summed modeled ensembles series were coupled through the use of LASSO regression developed by (TIBSHIRANI, 1996). The LASSO method is used to restrict the number of variables by adding a penalty term on the absolute value of the regression coefficients, which results in some coefficients being set to zero. This process was performed using the glmnet package in R (Friedman et al 2010). The regression model associated with the penalty value that resulted in the lowest mean cross-validated error was selected.

The framework of the long-term prediction models is displayed in Figure 40. In addition, the prediction of the average annual streamflow for different time windows (3, 5, 10, and 15-years) was compared to examine the effects on the time-varying forecast window of each model. Table 15 summarizes the period of model training, validation and prediction window.

Table 15 - Time period of model training, validation, and prediction for the multi-model.

Forecast window	**Training period**	**Validation period**	**Prediction period**
3 years (D1)	1937 – 2003	2004 – 2013	2014 – 2016
5 years (D2)	1937 – 2001	2002 – 2011	2012 – 2016
10 years (D3)	1937 – 1996	1997 – 2006	2007 – 2016
15 years (D4)	1937 – 1991	1992 – 2001	2002 – 2016

Source: Prepared by the author

7.4 Results and discussion

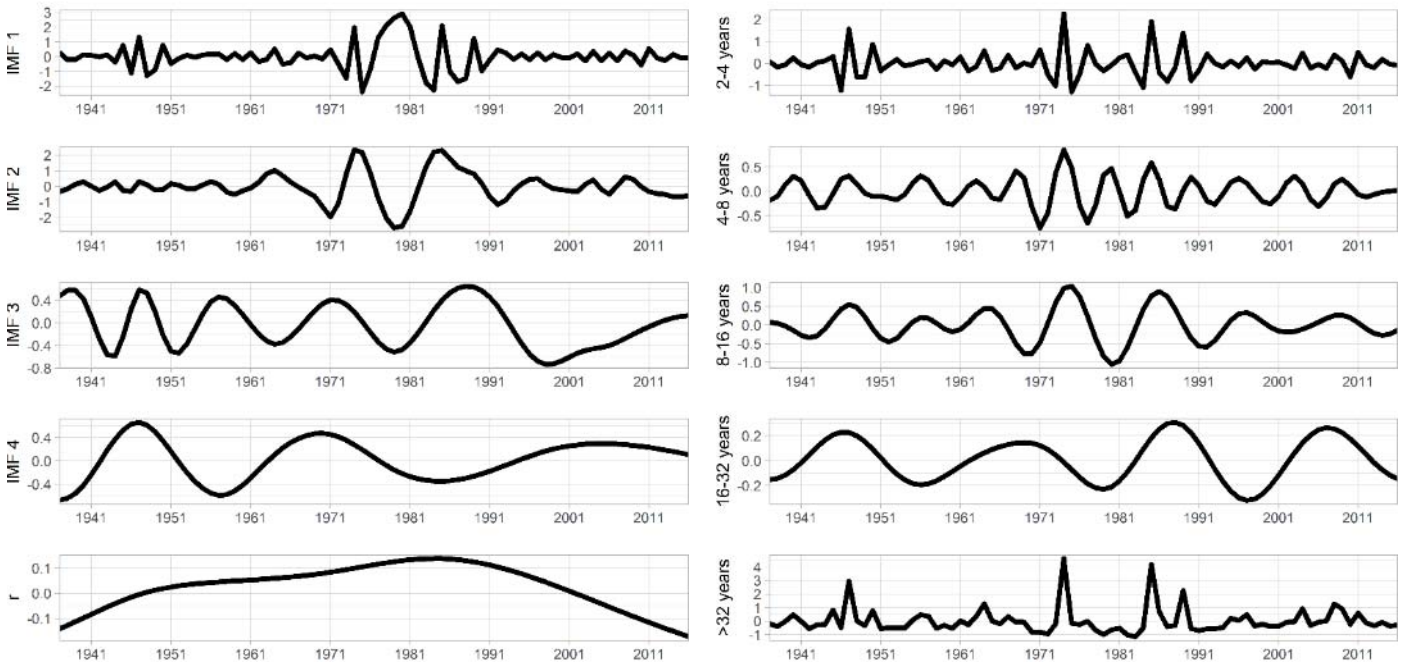
7.4.1 The original series decomposed

For the analysis of the original series, the annual streamflow from 1937 to 2016 was decomposed by CEEMDAN and WT. The decomposition according to the preprocessing methods used for the annual time series is exhibited in Figure 41. The streamflow series was decomposed into four IMFs and one residue component based on the CEEMDAN. The IMF1 (Figure 41b) represents the irregular part of the original series. Further, in the WT decomposition, a highly irregular pattern can also be seen in the 2-4 years frequency of WT in Figure 41c.

Figure 41b shows the residue component using the CEEMDAN method presents a downward inclination around 1991. Meanwhile, the last level subseries of WT shows a periodic behavior. This characteristic may be due to the lack of physical mechanism during the decomposition process by the WT method. Meng et al. (2019) found equivalent results when comparing the decomposition of streamflow time series applying a modified EMD, EMD, and WT. As streamflow is an important variable for water allocation in water-scarce regions, predicting it accurately is vital and provides essential information in the short and long-term planning process for water resources in the region.

Figure 41 - (a) Original annual streamflow time series, (b) decomposed results for the Orós reservoir using CEEMDAN and (c) decomposed results for the Orós reservoir using WT.

(a)



(b) (c)

Source: Prepared by the author

Furthermore, to account for some persistence present in the lower frequencies of the decomposition, a cross-wavelet and a coherence analysis were applied considering the known climatic indices to directly impact the variability of precipitation in the studied region. The Cross-Wavelet Transform (XWT) is used to calculate the coincident powers among two-time series. The Wavelet Transform Coherence (WTC) is applied to detect the frequency bands and time intervals wherein the two series are related (ROCHA et al., 2019). Additional information on the description of the XWT and WTC can be found in (TORRENCE & WEBSTER, 1999; TANG et al., 2014).

In the analysis of the XWT for the streamflow time series and Niño 3.4 anomaly index (Figure 42a), results show that they share areas in common in the power spectrum in the 4-8 years period between the years 1970-1990, starting as out of phase. Another common area is seen in the band of 8-16, from 1965 up to the end of the series, presenting an out-of-phase behavior.

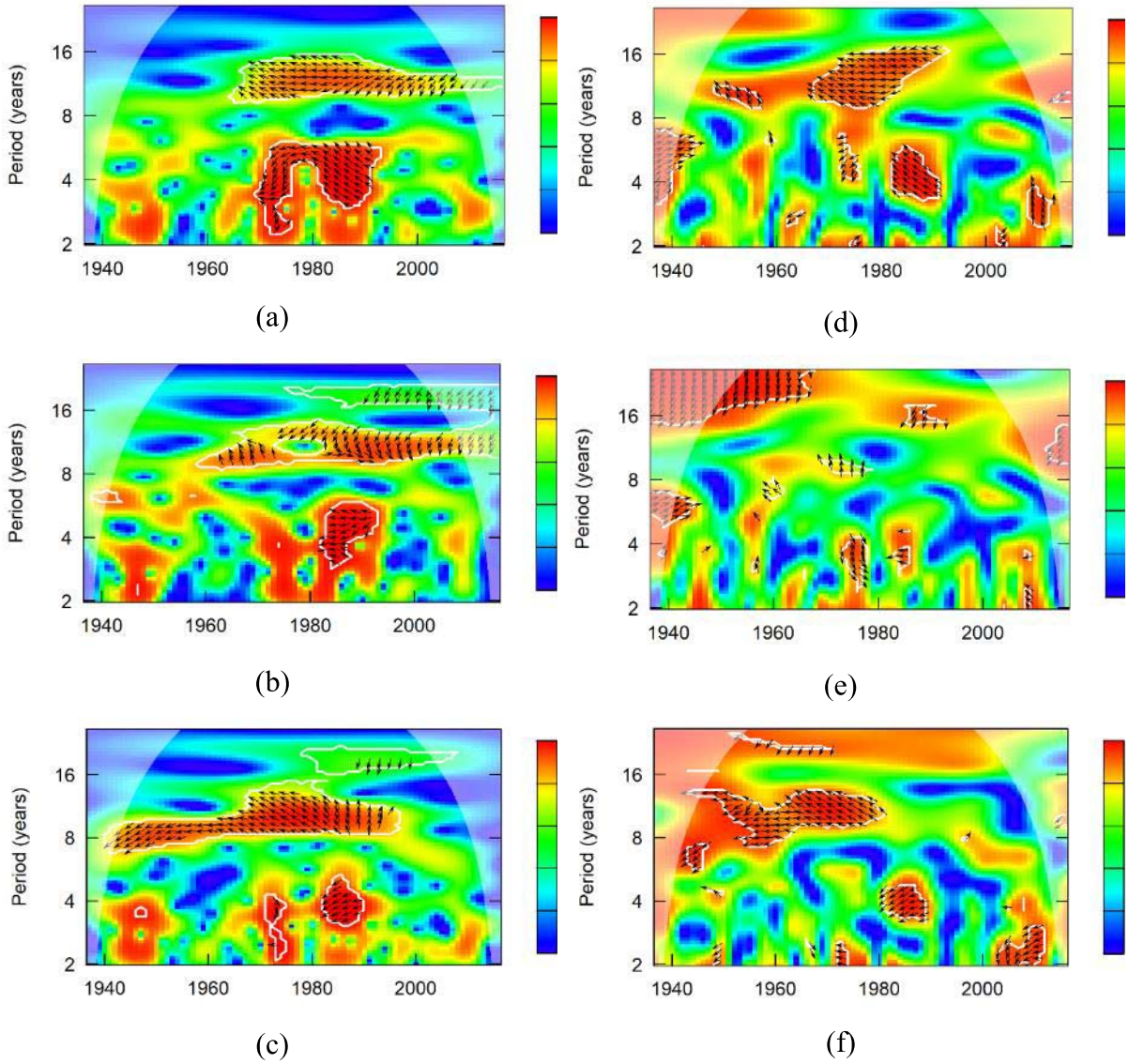
In XWT analysis for the streamflow and the Pacific Decadal Oscillation (PDO) (Figure 42b), both variables share areas in common on the power spectrum in the periods of 4-8 years between 1980 and 1990, starting out-of-phase with the streamflow leading. The 8-16 years period also shows a common area from the 1960s until the end of the series, starting out-of-phase. In the 80s and 90s, the indices are in phase and then return to out of phase. From the '90s onwards, there is a common area in the frequency above 16 years that shows out-of-phase behavior.

Unlike the other indices previously analyzed, for streamflow and Atlantic Multidecadal Oscillation (AMO) (Figure 42c), there are common areas on the power spectrum in the periods of 8-16 from the beginning of the series until the end of the '90s, starting out-of-phase and then staying in a positive phase. Similar to the PDO, the AMO analyses beginning in the '90s show a common area in the power spectrum for frequencies above 16-years that presents an out-of-phase behavior until the end of the '90s.

When comparing the decomposed time series, we observed that the Niño 3.4 anomaly and the PDO presented similar communal areas with the streamflow time series between 1970 and 1990 for the 2-8 years period. This can be observed through the higher peaks in the WT (2-4 years and 4-8 years frequency) and in the CEEMDAN (IMF1 and IMF2) decompositions in Figure 41b-c. Similar behavior can be seen in the 8-16 years period from

1970, while the decomposition (wavelet in the 8-16 years frequency and IMF4) composite shows oscillations with similar amplitudes. Also, the increase in the oscillation amplitude in the 16-32 years period coincides with the PDO and AMO's influence for a similar frequency (above 16-years). These results indicate that the streamflow of the study area is related to the analyzed indices, mainly in the variability between 8-16 years and less frequently (~ 16 to 32 years).

Figure 42 - Power spectrum XTC (a) streamflow and El Nino 3.4, (b) streamflow and PDO and (c) streamflow and AMO and Power spectrum WTC (d) streamflow and El Nino 3.4, (e) streamflow and PDO and (f) streamflow and AMO.



Source: Prepared by the author

The WTC between the streamflow and the Niño 3.4 anomaly index (Figure 42d) indicates significant areas with high values (> 0.8) over almost the entire series, at the beginning

of the 1940s for the 4-8 years period with influence in phase between the analyzed series. From the mid-1940s, the index influences throughout the analyzed period are noted, but out of phase.

The WTC between streamflow and PDO (Figure 42e) and between streamflow and AMO (Figure 42f) indicates areas with high values (> 0.8) over almost the entire series. For the PDO, there is an out-of-phase behavior at the beginning of the series until 1960 in the above 16-years period. As for the AMO, there is also an out-of-phase influence between 1940-1980. Furthermore, both indices' effects throughout the analyzed series are observed in the results. Rocha et al. (2019) investigated the influence of the PDO and AMO on the Standard Precipitation Index using WTC and XWT for the same area and detected a correlation between the 4-8 years and 16-32 years period.

7.4.2 Method for parameter selection

Hydrological time series are likely to be highly nonlinear and nonstationary. Thus, the CEEMDAN and WT were applied to obtain multiple subseries, which amount to different frequencies and parts of the streamflow time series. Prior to decomposition, the streamflow series was standardized for faster convergence of the parameters. Subseries of the streamflow for the different training and validation periods were selected as the input variables for the AR model and HMM. In the AR model, each component of the decomposition process was fitted for the training period, and the model order was chosen by the partial autocorrelation function plot analysis and the model's performance in the validation period. The orders of the AR models for the WARM and CEEMDAN-AR are presented in Table 16. In the HMM, the model was trained to vary the number of states from one to six. The states selected for the model were based on the lowest Bayesian information criterion value and the model's performance in the validation period. The number of states for the models WHMM and CEEMDAN-HMM is presented in Table 16. For the models without decomposition, the AR order and the number of states for the HMM are shown in Table 17.

Table 16 - The optimal order of AR model and number of hidden states for the HMM in the hybrid models for 3-, 5-, 10- and 15-year forecasts.

WT Decomposition	**WARM**				**WHMM**			
	D1	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
W1	4	4	4	1	4	4	3	3
W2	3	2	9	2	4	4	4	4
W3	2	2	12	13	4	2	3	2

W4	3	3	2	6	4	3	2	2
W5	4	4	1	5	4	4	3	4
CEEMDAN	**CEEMDAN-AR**				**CEEMDAN-HMM**			
Decomposition	**D1**	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
IMF1	9	9	9	9	4	4	3	3
IMF2	1	12	3	3	3	3	4	4
IMF3	9	1	9	1	4	3	4	3
IMF4	2	2	1	2	3	3	2	3
R	1	1	1	1	2	2	4	3

Source: Prepared by the author

Table 17 - The optimal order of AR model and number of hidden states for the HMM in the models without pre-processing for 3-, 5-, 10- and 15-year forecasts.

Forecast Window	AR	HMM
D1	9	4
D2	10	4
D3	8	4
D4	11	4

Source: Prepared by the author

The correlation coefficient R was used as a parameter for the order of AR and HMM models and the metrics were calculated both for the training and validation periods (Tables 18 and 19). The WARM was the model that performed better, while the WHMM and CEEMDAN-HMM had similar performances. The AR and HMM models had poor performance compared to the hybrid models in all analyzed periods (Table 19). IMFs 1 and 2 showed a lower correlation coefficient compared to the other components, this is due to the irregular behavior that these series have, as can be seen in Figure 41. In the multi-model framework, the fitted models were coupled using the lasso regression. In the lasso regression, a penalty for non-zero coefficients is added by penalizing the sum of their absolute values. As a result, some coefficients are exactly zeroed. The lasso regression was adjusted by performing a k-fold cross-validation and it showed a correlation coefficient of 0.98, 0.96, 0.95, and 0.78 for the analyzed periods. The regression coefficients for each prediction window can be found in Table 20. The highest regression coefficient belongs to the WARM, which performed better in all analyzed periods.

Table 18 - Evaluation of hybrid models in modeling the subseries of streamflow for the training and validation periods.

WT Decomposition	**R2 - WARM**	
	Training	**Validation**

	D1	D2	D3	D4	D1	D2	D3	D4
W1	0.87	0.87	0.87	0.24	0.36	0.28	0.27	0.23
W2	0.95	0.93	0.99	0.93	0.91	0.98	0.85	0.65
W3	0.99	0.99	1.00	0.99	0.96	0.96	0.98	0.85
W4	1.00	1.00	0.98	1.00	0.93	0.93	0.85	0.91
W5	1.00	1.00	0.98	1.00	0.99	0.98	0.00	0.99
WT Decomposition	**R2 - WHMM**							
	Training				**Validation**			
	D1	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
W1	0.87	0.66	0.55	0.58	0.15	0.35	0.04	0.17
W2	0.54	0.52	0.39	0.60	0.42	0.02	0.26	0.19
W3	0.67	0.28	0.72	0.42	0.20	0.11	0.05	0.43
W4	0.79	0.78	0.62	0.61	0.20	0.77	0.82	0.49
W5	0.85	0.84	0.87	0.89	0.75	0.81	0.18	0.39
CEEMDAN Decomposition	**CEEMDAN-AR**							
	Training				**Validation**			
	D1	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
IMF1	0.12	0.12	0.11	0.14	0.29	0.17	0.12	0.23
IMF2	0.66	0.96	0.93	0.93	0.12	0.25	0.53	0.93
IMF3	0.98	0.82	0.98	0.80	0.85	0.99	0.86	0.95
IMF4	0.98	0.99	0.92	0.98	0.98	0.13	0.98	0.98
R	0.99	0.99	0.98	0.99	0.99	1.00	0.99	0.91
CEEMDAN Decomposition	**CEEMDAN-HMM**							
	Training				**Validation**			
	D1	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
IMF1	0.74	0.75	0.75	0.75	0.59	0.55	0.14	0.16
IMF2	0.56	0.41	0.49	0.68	0.25	0.26	0.08	0.74
IMF3	0.75	0.80	0.90	0.79	0.61	0.26	0.95	0.82
IMF4	0.72	0.76	0.56	0.68	0.39	0.56	0.80	0.84
R	0.66	0.60	0.87	0.60	0.78	0.81	0.84	0.67

Source: Prepared by the author

Table 19 - Evaluation of AR and HMM models in modeling the subseries of streamflow for the training and validation periods.

Period	**AR**		**HMM**	
	Training	**Validation**	**Training**	**Validation**
D1	0.06	0.07	0.61	0.14
D2	0.08	0.13	0.51	0.08
D3	0.10	0.22	0.75	0.22
D4	0.25	0.06	0.76	0.01

Source: Prepared by the author

Table 20 - Regression coefficients for the LASSO regression for each forecast window for 3-, 5-, 10- and 15-year forecasts.

Regression coefficients	**D1**	**D2**	**D3**	**D4**

Intercept	0.09	0.09	0.11	0.15
AR	-0.16	-	-	0.02
WARM	0.89	0.88	0.88	0.71
CEEMDAN-AR	-0.04	-	-	-
HMM	0.19	0.18	0.16	0.26
WHMM	-	-	-	-
CEEMDAN-HMM	-	-0.02	-	0.08

Source: Prepared by the author

7.4.3 Comparative analysis

The forecast algorithm applied the AR and HMM using as input variables the subseries of the decomposition methods. For the models without the decomposition, the forecast was performed using only the AR and the HMM. The medians of the models were employed to form the multi-model. A comprehensive comparative analysis based on different metrics is needed to reveal the effect of the applied methods on streamflow prediction accuracy (Table 21). The R, RMSE, and MAE, were used as metrics to evaluate the performances of the median forecast values of the HMM, AR, WARM, WHMM, CEEMDAN-HMM, CEEMDAN-AR, and the multi-model. The RMSE is an ideal error metric to evaluate the global fitness of series that have high streamflow values, while the MAE gives a measure of overall errors (WANG et al., 2019; WEN et al., 2019).

Table 21 - The prediction performance indicators of WARM, WHMM, CEEMDAN-AR, CEEMDAN-HMM, AR, HMM, and Multi-model at Iguatu station.

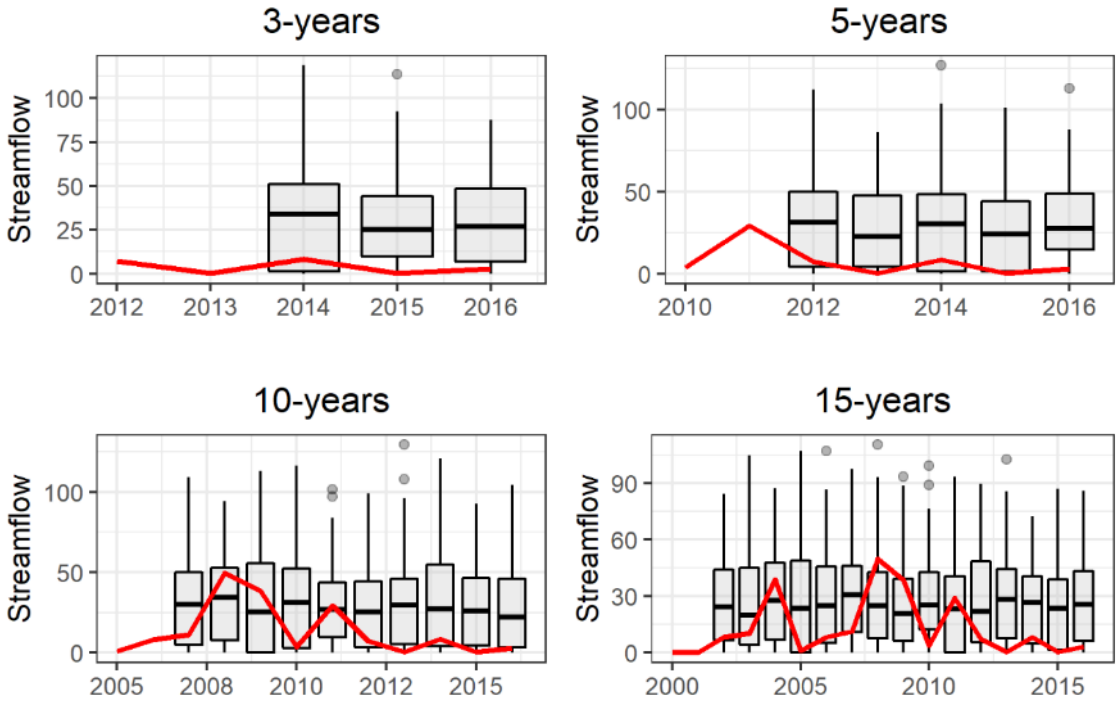
Model	Indicator	3-years	5-years	10-years	15-years
AR	MAE (m^3/s)	26.78	26.99	20.00	19.31
	RMSE (m^3/s)	26.84	27.09	22.06	20.48
	R2	0.96	0.54	0.30	0.01
WARM	MAE (m^3/s)	22.04	20.81	15.03	17.79
	RMSE (m^3/s)	22.49	21.53	16.42	18.67
	R2	0.24	0.01	0.08	0.00
CEEMDAN-AR	MAE (m^3/s)	16.2	23.17	14.27	16.39
	RMSE (m^3/s)	17.2	23.53	17.76	18.11

	R2	1.0	0.38	0.08	0.06
HMM	MAE (m^3/s)	14.91	28.24	19.92	19.65
	RMSE (m^3/s)	16.86	34.23	21.53	21.65
	R2	0.36	0.41	0.00	0.01
WHMM	MAE (m^3/s)	18.54	26.66	18.50	17.70
	RMSE (m^3/s)	20.87	27.39	20.76	19.72
	R2	0.89	0.54	0.13	0.00
CEEMDANHMM	MAE (m^3/s)	12.92	29.65	17.62	19.53
	RMSE (m^3/s)	14.65	30.80	19.24	22.45
	R2	0.85	0.92	0.35	0.04
MULTI-MODEL FORECAST	MAE (m^3/s)	17.85	22.27	15.99	16.80
	RMSE (m^3/s)	18.74	22.90	17.41	18.10
	R2	0.18	0.11	0.01	0.00

Source: Prepared by the author

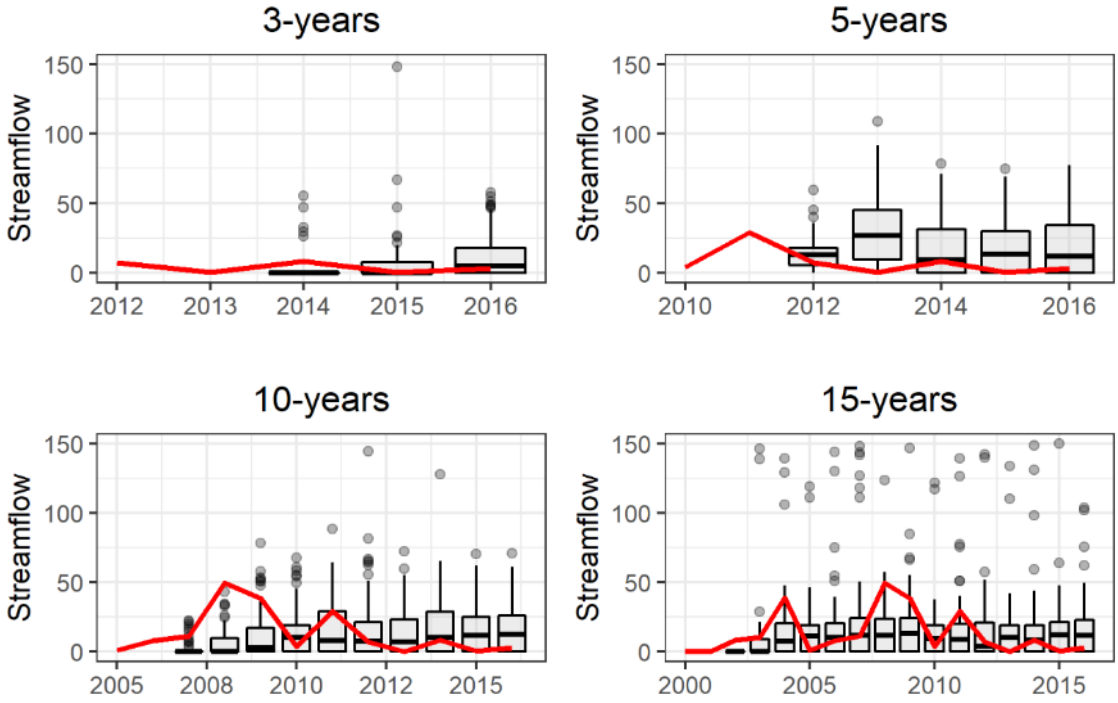
The comparison between the AR and the HMM models indicates that the forecasting accuracy of the AR model is greater than the HMM model, particularly for short-term forecasts, however, the HMM presents a lower MAE and RMSE for the 3-year forecast. This is an indication that the HMM can capture essential features present in streamflow series in the short term; however, at longer timescales, the Markov transition does not perform well, which according to Erkyihun et al. (2017), is a common characteristic of Markov chains. For the analyzed gauge station at the 3-year forecasting, the HMM shows an improvement over the AR model with a 44% decrease in RMSE and a 37% decrease in MAE. This difference between the AR and HMM in performance in terms of MAE and RSME can be attributed to nonlinear and nonstationary features present in the streamflow time series, which the AR model has limited ability to accurately represent due to linearity and normality assumptions. For the other forecasting windows, the AR shows a decrease of an average of 5% in RMSE and MAE. The forecast ensembles and the historic flows for the AR model are shown in Figure 43. In the Figure, during short-term forecasting, the median of the ensembles (horizontal line in the boxes) tracks the variability of the historic flows (red line) well, but not well during the long-term forecast. The AR model shows a larger uncertainty in the projections, particularly for longer forecast windows. Comparing the AR and HMM (Figure 44), the HMM shows a lower uncertainty.

Figure 43 - Forecasting ensembles (boxplot) and the historic flows (red) for the AR model.



Source: Prepared by the author

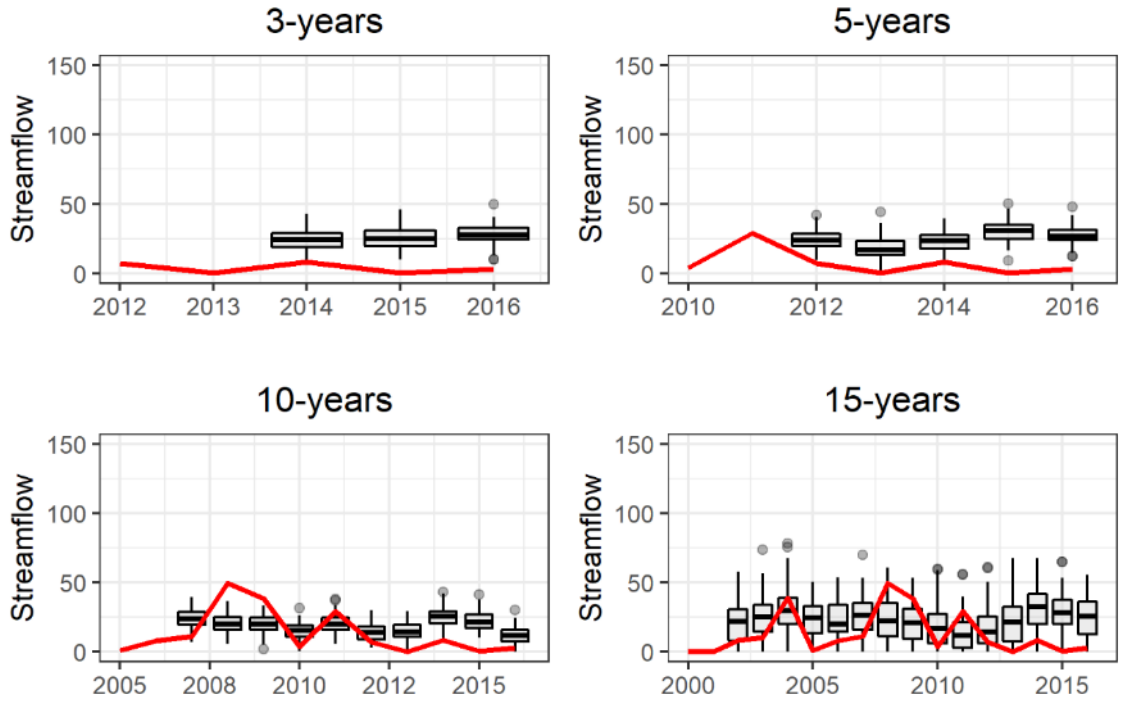
Figure 44 - Forecasting ensembles (boxplot) and the historic flows (red) for the HMM model.



Source: Prepared by the author

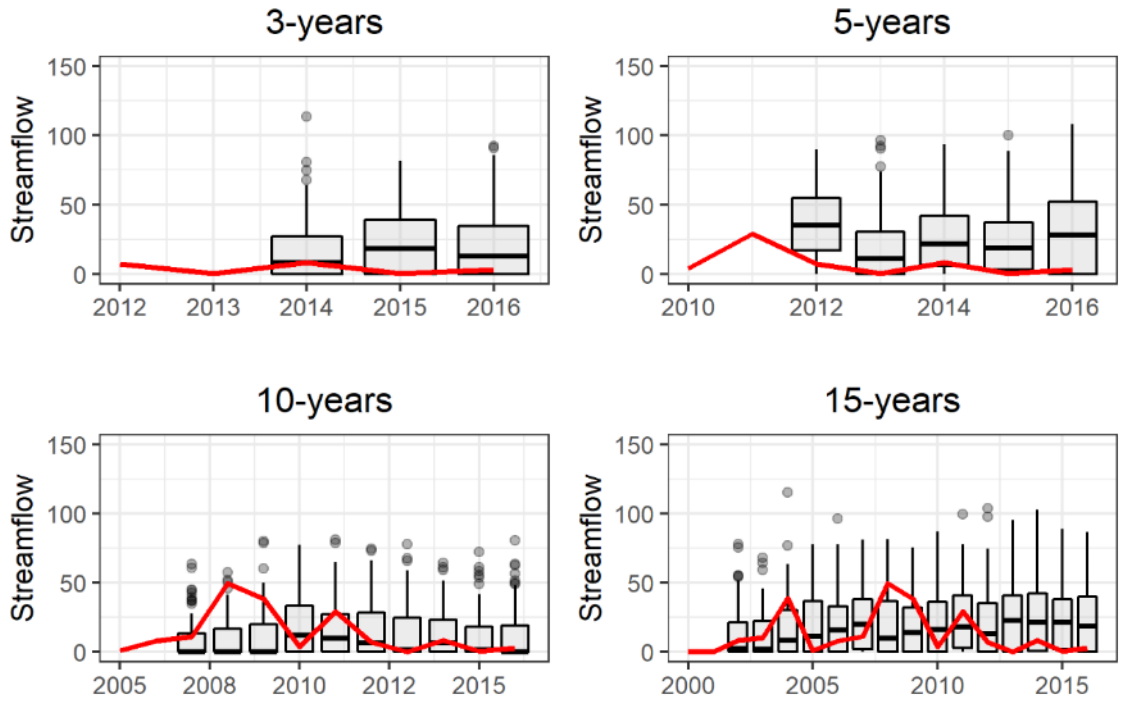
The hybrid models that applied the AR model (WARM and CEEMDAN-AR) showed an improvement regarding the AR model for the all projections, presenting an average decrease of 18% and 24% in MAE, and a decrease of 17% and 20% in RMSE, respectively. The CEEMDAN-AR (Figure 46) shows great uncertainty compared to the WARM (Figure 45). For the 15-year window, the CEEMDAN-AR can track the streamflow's variability quite well for the first years, however, the accuracy reduces after some time. In this study, the aim was to predict the state of the system and anticipate its future state. It was observed that most boxplots show little variability when compared to the observed values, indicating no transitions between states of low and high flows. Among the hybrid models that applied HMM (WHMM and CEEMDAN-HMM), the forecast accuracies showed an improvement regarding the HMM for the forecasted windows with a decrease in the MAE and RSME, and an increase in the R. Thus, the arrangement of these preprocessing approaches with the HMM model can improve the annual streamflow forecast accuracy, particularly for long-term forecasting. The finding of our study agrees with the other studies (WANG et al., 2019; WEN et al., 2019), which stated that hybrid model can enhance the forecasting accuracy of hydrological time series. Both models (WHMM and CEEMDAN-HMM) showed larger uncertainty for all the forecasting windows (Figures 47 and 48).

Figure 45 - Forecasting ensembles (boxplot) and the historic flows (red) for the WARM model.



Source: Prepared by the author

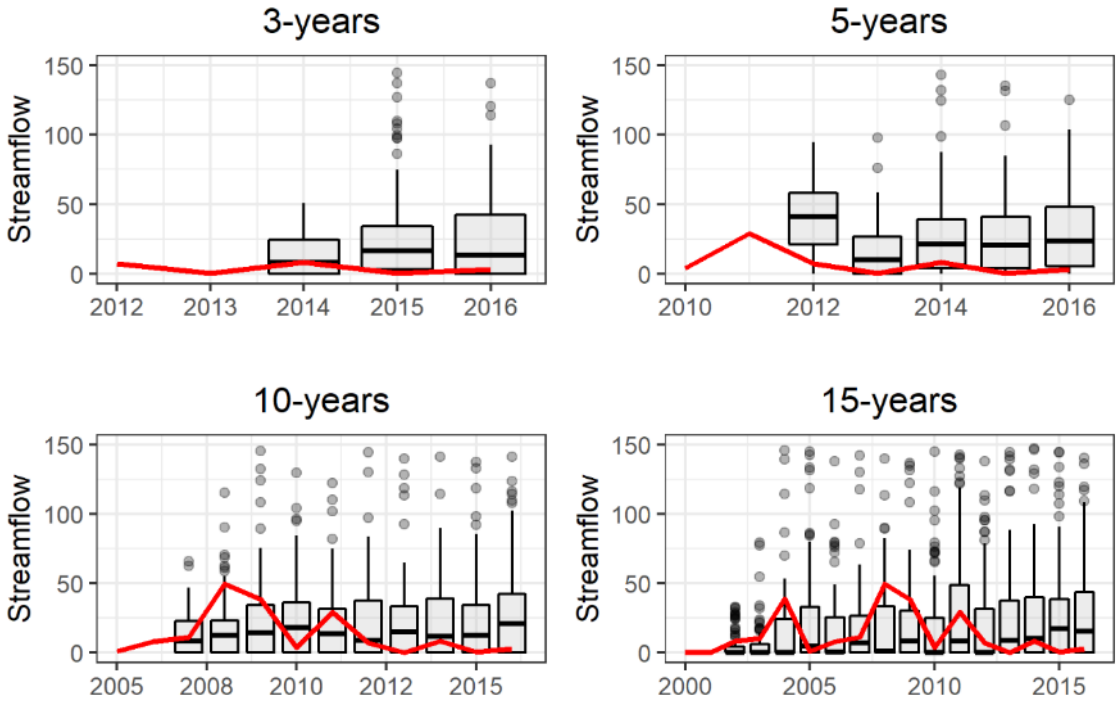
Figure 46 - Forecasting ensembles (boxplot) and the historic flows (red) for the CEEMDAN-AR model.



Source: Prepared by the author

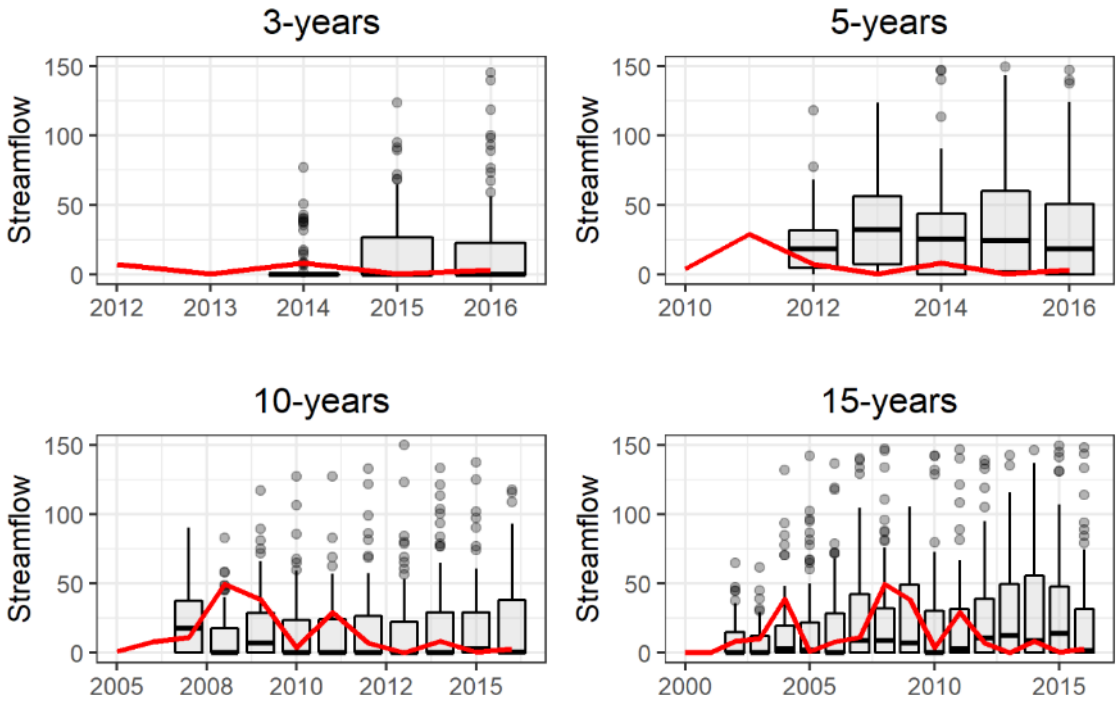
Figure 47 - Forecasting ensembles (boxplot) and the historic flows (red) for the WHMM

model.



Source: Prepared by the author

Figure 48 - Forecasting ensembles (boxplot) and the historic flows (red) for the CEEMDAN-HMM model.

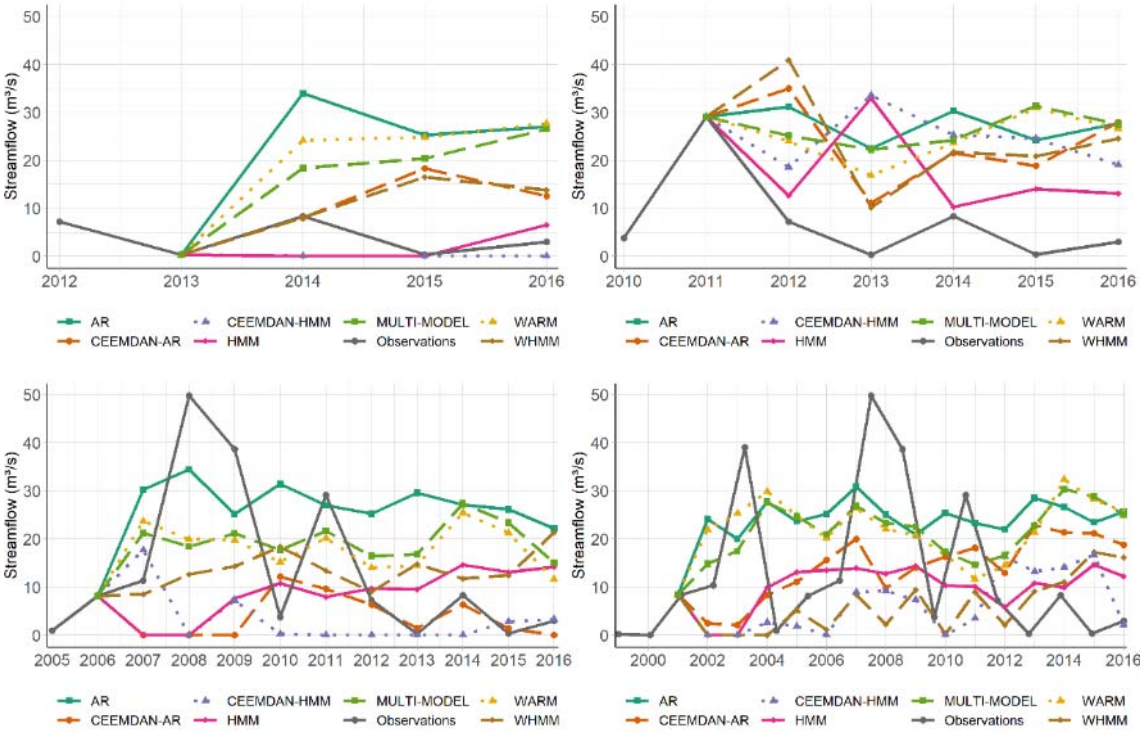


Source: Prepared by the author

Figure 49 shows the streamflow prediction at the four prediction windows for the AR, HMM, WARM, WHMM, CEEMDAN-AR, CEEMDAN-HMM, and multi-model. Looking at the LASSO regression coefficients, WARM contributes more to the multi-model, so it is observed that the behavior of the multi-model is similar to that of the WARM. The models that incorporate the decomposition into IMFs can predict better low flows. Although many studies apply hybrid models with WT in hydrological and water resources forecasting, Quilty and Adamowski (2018) stated that during wavelet decomposition sources of errors may be introduced, which are related to the boundary conditions. These sources of errors arise during wavelet decomposition when one seeks to calculate wavelet or scaling coefficients at a particular time and scale that cannot be calculated correctly (i.e., without introducing error). The boundary condition issues are associated with three main issues: the use of 'future data,' inappropriately selecting decomposition levels and wavelet filters, and not carefully partitioning calibration and validation data (QUILTY & ADAMOWSKI, 2018). The CEEMDAN can adaptively decompose a series of frequency components without a predetermined basis function, this makes it an improved method for analyzing nonlinear and nonstationary time series. Although some decomposition models have some drawbacks, the hybrid models can outperform their counterpart models, without decomposition, particularly for long-term projections and for the average annual time series such as the studied case, where the series has a large variability from year-to-year. The multi-model performance outperformed all the other models in terms of MAE and RMSE, particularly for long-period forecasts. Therefore, the multi-model can achieve better prediction accuracy than the other models applied in this study, demonstrating that the decomposition method coupled with a prediction algorithm and a regression method can improve the prediction accuracy of hydrological series and capture the dynamic and highly variable characteristics present in hydrological processes.

Figure 49 - Forecasted and observed average annual streamflow during the 3-, 5-, 10-, and 15-year period by WARM, WHMM, CEEMDAN-AR, CEEMDAN-HMM, AR, HMM, and

Multi-model at Iguatu station.



Source: Prepared by the author

Some models analyzed showed a remarkable amount of persistence from initial conditions, where the model can predict long-term well. This persistence can be associated with decadal and multidecadal time frames. Several studies relate low rainfall and extended drought in the Northeast region of Brazil with the El Niño phenomenon (MARENGO, 2008), and decadal variability of rainfall is associated with climate indices such as the PDO and the AMO (KAYANO & ANDREOLI, 2007; KNIGHT et al., 2006). Also, long-term projections indicate that there will be a trend for an increase in the frequency of consecutive dry days in the Northeast region (MARENGO, 2008). In addition, Rolim and Souza Filho (2020) showed that the low-frequency component of a streamflow time series of the Northeast region conditions the variability of a given year of the streamflow series. These features of hydrologic variability of the region where the station gauge is located are likely to influence the forecast accuracy. Furthermore, as shown in the XWT and WTC results, there is a strong influence of high and low-frequency climate indices in different frequencies and periods along the analyzed time series, which may influence the prediction of the analyzed dataset. The last years of the analyzed time series show an even more challenging feature due to the long-drought period (2012-2018) that affected the region influenced by a serial combination and association of the climatic phenomenon (PONTES FILHO et al., 2020).

7.5 Conclusion

In this study, a multi-model framework to predict short- and long-term average annual streamflow and a comparative analysis of six prediction models were applied to an annual streamflow time series. The models selected to form the multi-model were the WARM, WHMM, CEEMDAN-AR, CEEMDAN-HMM, AR, and HMM. The hybrid models' that applied the WARM methods had a greater contribution to the multi-model, showing that the hybrid method overcomes the limitation of stand-alone models.

For short-term (3-years) forecasting, the hybrid models that applied the HMM showed better accuracy over the AR model, while for long-term forecasting, the model that applied the WT decomposition performed better. Further, the arrangement of these preprocessing approaches can improve the annual streamflow forecast accuracy, particularly for long-term forecasting. Further, the multi-model presented the lowest RSME and MAE on average for all the forecasted windows compared to the other models. Thus, the multi-model can be helpful for reservoir management and provide new methods for hydrological forecasting based on data-driven modeling. Although the proposed models obtained satisfactory prediction performance, this study focuses on univariate streamflow forecasting without considering exogenous variables that affect streamflow, such as rainfall and climatic indices. However, we presented results that show that climatic indices such as El Nino, PDO, and AMO present great influence over the analyzed streamflow. In future research, other variables will be considered in the modeling of streamflow.

8 FINAL CONCLUSIONS AND REMARKS

This study aimed to develop a methodology for detecting deterministic chaos and analyzing complexity in rainfall and streamflow time series using nonlinear, chaos theory, and information theory methods. Additionally, a methodology was applied to predict those time series using machine learning based on the diagnosis of the presence of deterministic chaos, and a multi-model based on stochastic models was proposed.

In the article on chaos detection using nonlinear and chaos theory methods, the correlation dimension (CD) was employed to identify chaos. The results showed that more than 70% of the rainfall and 80% of streamflow exhibited saturation a point on the monthly timescale, indicating the presence of chaos. However, as the timescales increased, the detection of chaotic series decreased. Another method, the largest Lyapunov exponent (LLE), was used to detect the presence of deterministic chaos and assess predictability. While rainfall data showed positive LLE values, indicating the presence of deterministic chaos, three streamflow stations displayed negative exponents, suggesting non-chaotic behavior. Overall, streamflow exhibited higher LLE values, indicating stronger chaos compared to rainfall. Since predictability is inversely proportional to the LLE value, precipitation stations with deterministic chaos are predictable for longer periods compared to streamflow.

The presence of chaos in rainfall and streamflow can be detected even when considering different timescales. This study highlights the limitations of long-term predictions for streamflow due to its chaotic nature, which has significant implications for water resource management and the development of integrated plans. These insights shed light on the evolution of hydrological variables and their dynamics in both time and space, offering valuable information for water resource management applications.

The second article used information theory methods, specifically multiscale entropy (MSE), to understand complexity in the time series. The analysis of rainfall and streamflow data uncovered interesting patterns and insights about their complexity and predictability. When comparing the MSE values between rainfall and streamflow, it was observed that streamflow had lower entropy values for smaller timescales, indicating greater regularity or predictability in the time series data, suggesting reduced complexity or variability over time. However, it's important to note that this predictability is only for short-term durations, as indicated by the findings of the first article. Notably, this results from MSE takes into consideration complexity over multiple timescales. Furthermore, in the analysis of the Hurst exponent, most streamflow time series displayed higher persistence compared to the rainfall analysis. This higher

persistence can have an impact on predictability. Additionally, the study revealed a decrease in complexity for half of the rainfall stations in the State. However, two streamflow stations in the southeastern region showed increased entropy, suggesting heightened complexity in those specific time series.

Understanding the dynamics of hydrology is crucial in comprehending the behavior of hydrometeorological variables. The complexity of these series varies across different locations and over time. Specifically, in Ceará, the northwestern part of the state exhibits higher complexity in both rainfall and streamflow. This knowledge is relevant for various processes and plays a vital role in managing water resources. The methodology employed in this thesis can be extended to analyze other hydrometeorological systems, thereby highlighting the presence of deterministic chaos in natural environments. We argue that chaos theory could provide a better understanding of hydrological systems compared to the commonly used deterministic/stochastic paradigm.

In the third article, based on the results of the first article, 20 rainfall time series that exhibited the presence of deterministic chaos were selected as input data for machine learning models. The study incorporates the principles of chaos theory and employs various machine learning techniques to analyze the rainfall dynamics. The data was divided into training and testing windows, and the embedding dimension was chosen based on the accuracy of the models during the training phase. It was found that most stations required embedding dimensions greater than 10 and relied on information from at least three years prior to the current time for accurate forecasting. This order of magnitude in the embedding dimension, determined by forecasting accuracy, is lower than the embedding dimension found using the CD method in the first article. This finding aligns with numerous studies suggesting that the optimal dimension for forecasting may not necessarily be the best dimension to represent the chaotic dynamics of the time series. Furthermore, the study found that each location had unique rainfall patterns and required different models for accurate forecasting. During the training phase, the RF model outperformed the other models, and the same was observed in the testing phase. However, the SVM also performed well in many stations and had similar accuracy to the RF model.

The stations that had the had the highest NSE in testing phase, namely, Tianguá e Ubajara, had a predictability horizon of approximately 48 and 49 months, respectively, which corresponded to the duration of the testing period in the third article. On the other hand, the Mulungu station exhibited an even higher predictability horizon of approximately 60 months. This indicates that the rainfall data at the Mulungu station could be accurately forecasted for a longer period compared to other stations. However, it is important to note that despite having a

higher predictability horizon, the station had a low NSE value. This occurrence could be attributed to a possibility that the low NSE value indicates a discrepancy between the observed and predicted rainfall values at the Mulungu station. This discrepancy may be influenced by factors that are not adequately captured by the forecasting model used in the study, such as local terrain features, microclimates, or other complex interactions that affect rainfall.

Considering the results of articles two and three, it was observed that most stations exhibiting a Hurst exponent close to 0.5 had the lowest values of NSE. This can be justified by their random behavior, making the time series hard to predict. On the other hand, stations like Ipueira and Ubajara, which had a Hurst exponent higher than 0.5, also demonstrated high performance in terms of NSE with the LSTM model. This can be seen as an indication of long-term persistence in the series, which the model was able to capture.

This prediction study demonstrates that these data-driven methods can provide satisfactory results for rainfall dynamics without requiring much physical information. However, including well-known climatic indices to influence the rainfall pattern could improve the forecasting accuracy, but it may significantly increase the problem dimension and impact the algorithm's runtime, requiring further studies to assess its feasibility. Another limitation of this work was the lack of an uncertainty analysis regarding the rainfall forecast. We believe that including this step could enhance the results and further investigate the potential of the forecasting.

The final article utilized a multi-model framework to forecast the average annual streamflow, both in the short-term and long-term, for a specific station. The framework consisted of six prediction models: WARM, WHMM, CEEMDAN-AR, CEEMDAN-HMM, AR, and HMM. Among the hybrid models, the WARM had the highest contribution to the multi-model. This finding suggests that the hybrid method overcomes the limitations of individual models. The arrangement of these preprocessing approaches could improve the accuracy of annual streamflow forecasting, particularly for long-term predictions. The multi-model outperformed the other models in terms of average RSME and MAE for all the forecasted windows. However, similar to the previous article, this study focused on univariate streamflow forecasting and excluded exogenous variables that may impact streamflow, such as rainfall and climatic indices. Nevertheless, the study revealed that climatic indices like El Nino, PDO, and AMO have a significant impact on the analyzed streamflow.

Furthermore, the station analyzed in the fourth article was selected to apply a more stochastic approach due to its relevance in terms of water supply and the absence of deterministic chaos, in the seasonal and annual timescale. However, at the monthly timescale,

this station had a predictability horizon of only up to 4 months. This limited predictability horizon could be one of the reasons why the stochastic forecast models performed poorly.

To summarize, the main objective of this work was achieved through a combination of articles, each presenting a different methodological approach for detecting chaos and analyzing complexity in rainfall and streamflow series, while also proposing methods for the forecast of these variables. For future studies, the following recommendations can be made:

- Additional studies are needed to determine the most effective complementary methods for simultaneously assessing the reliability of results and making direct inferences regarding these methods' outcomes.
- Future studies aiming to explore complexity should consider employing multiple methodologies to capture a wide range of cases and develop complementary approaches. This could involve examining the complexity of rainy days or exploring complexity across different decades.
- Further investigation is necessary to address the uncertainty associated with machine learning-based rainfall forecasting.
- More research is needed to incorporate climate indices into the structure between stations in the rainfall and streamflow forecasting framework.

REFERENCES

ABARBANEL, H.D.I.; LALL, U. Nonlinear dynamics of the Great Salt Lake: system identification and prediction. **Climate Dynamics**, Germany, v. 12, n. 4, p. 287-297, 1996. DOI: https://doi.org/10.1007/BF00219502. Disponível em: https://link.springer.com/article/10.1007/BF00219502. Acesso em: 16 jun. 2023.

ADARSH, S.; REDDY, M. Janga. Multiscale characterization and prediction of monsoon rainfall in India using Hilbert–Huang transform and time-dependent intrinsic correlation analysis. **Meteorology and Atmospheric Physics**, Austria, v. 130, n. 6, p. 667-688, 2018. Disponível em: https://doi.org/10.1007/s00703-017-0545-6. Disponível em: https:// https://link.springer.com/article/10.1007/s00703-017-0545-6. Acesso em: 16 jun. 2023.

ADARSH, S.; Nourani, V.; Archana, D.S.; Dharan, D.S. Multifractal description of daily rainfall fields over India. **Journal of Hydrology**, Amsterdam, v. 586, n. 3, p. 124913, 2020. DOI: https://doi.org/10.1016/j.jhydrol.2020.124913. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169420303735?via%3Dihub. Acesso em: 16 jun. 2023.

ADNAN, R.M.; Liang, Z.; Heddam, S.; Zounemat-Kermani, M.; Kisi, O.; Li, B. Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. **Journal of Hydrology**, Amsterdam, v. 586, n. 7, p. 124371, 2020. DOI: https://doi.org/10.1016/j.jhydrol.2019.124371. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169419311060. Acesso em: 16 jun. 2023.

ALFARO, M., FUERTES, G., VARGAS, M. SEPÚLVEDA, J., VELOSO-POBLETE, M. Forecast of chaotic series in a horizon superior to the inverse of the maximum Lyapunov exponent. **Complexity**, London, v. 2018, 2018. DOI: https://doi.org/10.1155/2018/1452683. Disponível em: https://www.hindawi.com/journals/complexity/2018/1452683/. Acesso em: 17 jun. 2023.

ANDREOLI, R.V.; KAYANO, M.T. ENSO-related rainfall anomalies in South America and associated circulation features during warm and cold Pacific decadal oscillation

regimes. **International Journal of Climatology**, United Kingdom, v. 25, n. 15, p. 2017-2030, 2005. DOI: https://doi.org/10.1002/joc.1222. Disponível em: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.1222. Acesso em: 17 jun. 2023.

ANTICO, A.; SCHLOTTHAUER, G.; TORRES, M.E. Analysis of hydroclimatic variability and trends using a novel empirical mode decomposition: application to the Paraná River Basin. **Journal of Geophysical Research: Atmospheres**, Hoboken, v. 119, n. 3, p. 1218-1233, 2014. DOI: https://doi.org/10.1002/2013JD020420. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2013JD020420. Acesso em: 17 jun. 2023.

BABOVIC, V.; KEIJZER, M. Forecasting of river discharges in the presence of chaos and noise. *In***:** MARSALEK, J., WATT, W.E., ZEMAN, E., SIEKER, F. (ed) **Flood Issues in Contemporary Water Management. NATO Science Series**, Dordrecht, 2020, p. 405-419. DOI: https://doi.org/10.1007/978-94-011-4140-6_42. Disponível em: https://link.springer.com/chapter/10.1007/978-94-011-4140-6_42. Acesso em 18 jun. 2023.

BAESENS, B.; VIAENE, S.; VAN GESTEL, T.; SUYKENS, J.A.; DEDENE, G.; DE MOOR, B.; VANTHIENEN, J. An empirical assessment of kernel type performance for least squares support vector machine classifiers. *In*: KES'2000. FOURTH INTERNATIONAL CONFERENCE ON KNOWLEDGE-BASED INTELLIGENT ENGINEERING SYSTEMS AND ALLIED TECHNOLOGIES, 1., 2000, Brighton. **Proceedings [...]**. Brighton, 2020, p. 313-316. DOI: https://doi.org/10.1109/KES.2000.885819. Disponível em: https://ieeexplore.ieee.org/document/885819. Acesso em: 18 jun. 2023.

BARRETO, I.D.; STOSIC, T.; FILHO, M.C.; DELRIEUX, C., SINGH, V.P.; STOSIC, B. Complexity analyses of Sao Francisco river streamflow: influence of dams and reservoirs. **Journal of Hydrologic Engineering**, United States, v. 25, n. 10, p. 05020036, 2020. DOI: https://doi.org/10.1061/(ASCE)HE.1943-5584.0001996. Disponível em: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29HE.1943-5584.0001996. Acesso em: 22 jun. 2023.

BAYDAROĞLU, Ö.; KOÇAK, K. SVR-based prediction of evaporation combined with chaotic approach. **Journal of Hydrology**, Amsterdam, v. 508, n. 1, p. 356-363, 2014. DOI: https://doi.org/10.1016/j.jhydrol.2013.11.008. Disponível em:

https://www.sciencedirect.com/science/article/abs/pii/S002216941300824X. Acesso em: 18 jun. 2023

BENETTIN, G.; FROESCHLE, C.; SCHEIDECKER, J. P. Kolmogorov entropy of a dynamical system with an increasing number of degrees of freedom. **Physical Review A**, College Park, v. 19, n. 6, p. 2454, 1979. DOI: https://doi.org/10.1103/PhysRevA.19.2454. Disponível em: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.19.2454. Acesso em: 19 jun. 2023.

BOARETTO, B.R.R.; Budzinski, R.C.; Rossi, K.L.; Prado, T.L.; Lopes, S.R.; Masoller, C. Discriminating chaotic and stochastic time series using permutation entropy and artificial neural networks. **Sci. Rep.**, United Kingdom, v. 11, n. 1, p. 15789, 2021. DOI: https://doi.org/10.1038/s41598-021-95231-z. Disponível em: https://www.nature.com/articles/s41598-021-95231-z. Acesso em: 19 jun. 2023.

BOWMAN, D.C.; LEES, J.M. The Hilbert–Huang transform: A high resolution spectral method for nonlinear and nonstationary time series. **Seismological Research Letters**, McLean, VA, v. 84, n. 6, p. 1074-1080, 2013. DOI: https://doi.org/10.1785/0220130025. Disponível em: https://pubs.geoscienceworld.org/ssa/srl/article-abstract/84/6/1074/349062/The-Hilbert-Huang-Transform-A-High-Resolution. Acesso em: 19 jun. 2023.

BRAGA A.C.; ALVES, L.G.A.; COSTA L.S.; RIBEIRO A.A.; DE JESUS M.M.A.; TATEISHI A.A.; RIBEIRO H.V. Characterization of river flow fluctuations via horizontal visibility graphs. **Physica A: Statistical Mechanics and its Applications**, Amsterdam, v. 444, n. 2, p. 1003-1011, 2016. Disponível em: https://doi.org/10.1016/j.physa.2015.10.102

BREIMAN, L. Random forests. **Machine learning**, v. 45, p. 5-32, 2001.

BREIMAN, L. et al. Classification and regression trees. Boca Raton, Florida, 1984.

CATANI, F.; LAGOMARSINO, D.; SEGONI, S.; TOFANI, V. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. **Natural Hazards and Earth System Sciences**, Germany, v. 13, n. 11, p. 2815-2831, 2013. DOI: https://doi.org/10.5194/nhess-13-2815-2013. Disponível em: https://nhess.copernicus.org/articles/13/2815/2013/. Acesso em: 20 jun. 2023.

CHO, K.; KIM, Y. Improving streamflow prediction in the WRF-Hydro model with LSTM networks. **Journal of Hydrology**, Amsterdam, v. 605, n. 2, p. 127297, 2022. DOI: https://doi.org/10.1016/j.jhydrol.2021.127297. Disponível em: https://www.sciencedirect.com/science/article/pii/S0022169421013470. Acesso em 20 jun. 2023.

CHOU, C.M.; WANG, R.Y. Application of wavelet-based multi-model Kalman filters to real-time flood forecasting. **Hydrological processes**, United Kingdom, v. 18, n. 5, p. 987-1008, 2004. DOI: https://doi.org/10.1002/hyp.1451. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.1451. Acesso em: 20 jun. 2023.

CHOU, C.M. Complexity analysis of rainfall and runoff time series based on sample entropy in different temporal scales. **Stochastic Environmental Research and Risk Assessment**, New York, v. 28, n. 6, p. 1401-1408, 2014. DOI: https://doi.org/10.1007 /s00477-014-0859-6. Disponível em: https://link.springer.com/article/10.1007/s00477-014-0859-6. Acesso em: 20 jun. 2023.

CHOU, C.M. Applying multiscale entropy to the complexity analysis of rainfall-runoff relationships. **Entropy**, Switzerland, v. 14, n. 5, p. 945-957, 2012. DOI: https://doi.org/10.3390/e14050945. Disponível em: https://www.mdpi.com/1099-4300/14/5/945. Acesso em: 20 jun. 2023.

CLARK, M.P.; BIERKENS, M.F.P.; SAMANIEGO, L.; WOODS, R.A.; UIJLENHOET R.; BENNETT K.E.; PAUWELS, V.R.N.; CAI, X.; WOOD, A.W.; PETERS-LIDARD, C.D. The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism. **Hydrology and Earth System Sciences**, Germany, v. 21, n. 7, p. 3427-3440, 2017. DOI: https://doi.org/10.5194/hess-21-3427-2017. Disponível em: https://hess.copernicus.org/articles/21/3427/2017/. Acesso em: 21 jun. 2023.

COSTA, A.C.; ESTÁCIO, A.B.S.; SOUZA FILHO, F.A.; LIMA NETO, I.E. Monthly and seasonal streamflow forecasting of large dryland catchments in Brazil. **Journal of Arid Land**, China, v. 13, n. 3, p. 205–223. DOI: https://doi.org/10.1007/s00000-021-0097-x. Disponível em: https://link.springer.com/article/10.1007/s40333-021-0097-y. Acesso em: 21 jun. 2023.

COSTA, M.; GOLDBERGER, A.L.; PENG, C.K. Multiscale entropy analysis of complex physiologic time series. **Physical review letters**, United States, v. 89, n. 6, p. 068102, 2002. DOI: https://doi.org/10.1103/PhysRevLett.89.068102. Disponível em: https://pubmed.ncbi.nlm.nih.gov/12190613/. Acesso em: 22 jun. 2023.

DA SILVA, V.D.P.R.; BELO FILHO, A.F.; ALMEIDA R.S.R.; DE HOLANDA, R.M.; CAMPOS J.H.B.C. Shannon information entropy for assessing space–time variability of rainfall and streamflow in semiarid region. **Science of The Total Environment**, Netherlands, v. 544, n. 2, p. 330-338, 2016. DOI: https://doi.org/10.1016/j.scitotenv.2015.11.082. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0048969715310639. Acesso em: 22 jun. 2023.

DARIANE, A.B.; AZIMI, S. Streamflow forecasting by combining neural networks and fuzzy models using advanced methods of input variable selection. **Journal of Hydroinformatics**, United Kingdom, v. 20, n. 2, p. 520-532, 2018. DOI: https://doi.org/10.2166/hydro.2017.076. Disponível em: https://iwaponline.com/jh/article/20/2/520/37910/Streamflow-forecasting-by-combining-neural. Acesso em: 22 jun. 2023.

DETTINGER, M.D.; GHIL, M; KEPPENNE, C.L. Interannual and interdecadal variability in United States surface-air temperatures, 1910-87. **Climatic Change**, Netherlands, v. 31, n. 1, p. 35-66, 1995. DOI: https://doi.org/10.1007/BF01092980. Disponível em: https://link.springer.com/article/10.1007/BF01092980. Acesso em: 22 jun. 2023

DHANYA, C.T.; KUMAR, D. Nagesh. Nonlinear ensemble prediction of chaotic daily rainfall. **Advances in Water resources**, United Kingdom, v. 33, n. 3, p. 327-347, 2010. DOI: https://doi.org/10.1016/j.advwatres.2010.01.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0309170810000138. Acesso em: 22 jun. 2023.

DING, G-B; DING, J-F. An improved hydrological forecast method based on chaos and trend term. *In*: 2009 CHINESE CONTROL AND DECISION CONFERENCE, 1., 2009, China. **Proceedings [...]**. China, 2009. p. 4929-4932. DOI: 10.1109/CCDC.2009.5194910. Disponível em: https://ieeexplore.ieee.org/document/5194910. Acesso em: 22 jun. 2023.

DWARAKISH, G.S.; GANASRI, B.P. Impact of land use change on hydrological systems: A review of current modeling approaches. **Cogent Geoscience**, United Kingdom, v. 1, n. 1, p.

1115691, 2015. DOI: https://doi.org/10.1080/23312041.2015.1115691. Disponível em: https://www.tandfonline.com/doi/full/10.1080/23312041.2015.1115691. Acesso em: 23 jun. 2023.

ECHI, I. M.; TIKYAA, E. V.; ISIKWUE, B. C. Dynamics of daily rainfall and temperature in Makurdi. **International journal of science and Research**, v. 4, n. 7, p. 493-499, 2015. Disponível em: https://www.researchgate.net/profile/Emmanuel-Tikyaa/publication/291339515_Dynamics_of_Daily_Rainfall_and_Temperature_in_Makurdi/links/56a0c4d108aee4d26ad800f9/Dynamics-of-Daily-Rainfall-and-Temperature-in-Makurdi.pdf. Acesso em 24 jun. 2023.

ECKMANN, J.P.; KAMPHORST, S.O.; RUELLE, D. Recurrence plots of dynamical systems. **World Scientific Series on Nonlinear Science Series A**, Bristol, v. 16, n. 9, p. 441-446, 1995. Disponível em: https://iopscience.iop.org/article/10.1209/0295-5075/4/9/004/meta. Acesso em: 24 jun. 2023

ELSHORBAGY, A.; SIMONOVIC, S. P.; PANU, U. S. Noise reduction in chaotic hydrologic time series: facts and doubts. **Journal of Hydrology**, Amsterdam, v. 256, n. 3-4, p. 147-165, 2002. DOI: https://doi.org/10.1016/S0022-1694(01)00534-0. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169401005340. Acesso em: 24 jun. 2023.

ERKYIHUN, S.T.; ZAGONA, E.; RAJAGOPALAN, B. Wavelet and hidden Markov-based stochastic simulation methods comparison on Colorado River streamflow. **Journal of Hydrologic Engineering**, United States, v. 22, n. 9, p. 04017033, 2017. DOI: https://doi.org/10.1061/(ASCE)HE.1943-5584.0001538. Disponível em: https://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0001538. Acesso em: 24 jun. 2023.

FALAYI, E.O.; ADEPITAN, J.O.; ADEWOLE, A.T.; ROY-LAYINDE, T.O. Analysis of rainfall data of some West African countries using wavelet transform and nonlinear time series techniques. **Journal of Spatial Science**, United Kingdom, p. 1-12, 2022. DOI: https://doi.org/10.1080/14498596.2021.2008539. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/14498596.2021.2008539. Acesso em: 24 jun. 2023.

FAN, J.; YUE, W.; WU, L.; ZHANG, F.; CAI, H.; WANG, X.; LU, X.; XIANG, Y. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. **Agricultural and forest meteorology**, Netherlands, v. 263, n. 12, p. 225-241, 2018. Disponível em: https://www.sciencedirect.com/science/article/pii/S0168192318302855?casa_token=28EH-83bFOoAAAAA:YRpraKEws82JegXSGCRL3bP5wlJGxmH3HYS9yb0EmsYHQEqrZ-v3a0jxZ_072Fl_K2nyGbVAdec. Acesso em 25 jun. 2023.

FANG, K.; SIVAKUMAR, B.; WOLDEMESKEL, F.M. Complex networks, community structure, and catchment classification in a large-scale river basin. **Journal of hydrology**, Amsterdam,v. 545, p. 478-493, 2017. DOI: https://doi.org/10.1016/j.jhydrol.2016.11.056. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169416307715. Acesso em: 25 jun. 2023.

FARMER, J.D.; SIDOROWICH, J.J. Predicting chaotic time series. **Physical review letters**, United States, v. 59, n. 8, p. 845, 1987. DOI em: https://doi.org/10.1103/PhysRevLett.59.845. Disponível em: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.59.845. Acesso em: 25 jun. 2023.

FENG, Z.K.; NIU, W.J.; LIU, S.; LUO, B.; MIAO, S.M.; LIU, K. Multiple hydropower reservoirs operation optimization by adaptive mutation sine cosine algorithm based on neighborhood search and simplex search strategies. **Journal of hydrology**, Amsterdam, v. 590, n. 11, p. 125223, 2020. DOI: https://doi.org/10.1016/j.jhydrol.2020.125223. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169420306831. Acesso em 25 jun. 2023.

FRASER, Andrew M.; SWINNEY, Harry L. Independent coordinates for strange attractors from mutual information. **Physical review A**, College Park, v. 33, n. 2, p. 1134, 1986. DOI: https://doi.org/10.1103/PhysRevA.33.1134. Disponível em: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.33.1134. Acesso em 26 jun. 2023.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of statistical software**, Los Angeles, v. 33, n. 1, p. 1, 2010. Disponível em: https://www.jstatsoft.org/article/view/v033i01. Acesso em: 26 jun.

2023.

FUWAPE, I.A.; OGUNJO, S.T.; OLUYAMO, S.S.; RABIU, A.B. Spatial variation of deterministic chaos in mean daily temperature and rainfall over Nigeria. **Theoretical and Applied Climatology**, Austria, v. 130, n. 1, p. 119-132, 2017. DOI: https://doi.org/10.1007/s00704-016-1867-x. Disponível em: https://link.springer.com/article/10.1007/s00704-016-1867-x. Acesso em: 26 jun. 2023.

GAISER, T.; KROL, M.; FRISCHKORN, H.; DE ARAÚJO, J.C. (ed.) **Global change and regional impacts:** Water availability and vulnerability of ecosystems and society in the semiarid Northeast of Brazil. Berlin, Heidelberg, Dordrecht, and New York City, 2003.

GARREAUD, R.D.; VUILLE, M.; COMPAGNUCCI, R.; MARENGO, J. Present-day South American climate, **Palaeogeography, Palaeoclimatology, Palaeoecology**, Netherlands, v. 281, n .3-4, p. 180–195, 2009. DOI: https://doi.org/10.1016/j.palaeo.2007.10.032. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0031018208005002. Acesso em: 26 jun. 2023.

GHORBANI, M.A.; KHATIBI, R.; MEHR, A.D.; ASADI, H. Chaos-based multigene genetic programming: a new hybrid strategy for river flow forecasting. **Journal of hydrology**, Amsterdam, v. 562, p. 455-467, 2018. DOI: https://doi.org/10.1016/j.jhydrol.2018.04.054. Disponível em: https://www.sciencedirect.com/science/article/pii/S002216941830307X?casa_token=6QJhiULjBSgAAAAA:Hth9umGthUPLgz4wnVCU5IO1lMRZ4wfxmx6DUiv9XlGRCBYYDheS2btsRLjgiOjjtiqggWy-hwc. Acesso em: 27 jun. 2023.

GIRI, F.; DEVERCELLI, M. Chaos arising from the hydrological behaviour of a floodplain river during the last century. **River Research and Applications**, United Kingdom, v. 39, n. 2, p. 241-254, 2023. DOI: https://doi.org/10.1002/rra.4080. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/rra.4080. Acesso em: 28 jun. 2023.

GRASSBERGER, P.; PROCACCIA, I. Characterization of strange attractors. **Physical review letters**, United States, v. 50, n. 5, p. 346, 1983a. DOI: https://doi.org/10.1103/PhysRevLett.50.346. Disponível em: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.50.346. Acesso em: 28 jun. 2023.

GRASSBERGER, P.; PROCACCIA, I. Estimation of the Kolmogorov entropy from a chaotic signal. **Physical review A**, College Park, v. 28, n. 4, p. 2591, 1983b. DOI: https://doi.org/10.1103/PhysRevA.28.2591. Disponível em: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.28.2591. Acesso em: 28 jun. 2023.

GU, H.; YU, Z.; WANG, G.; WANG, J.; JU, Q.; YANG, C.; FAN, C. Impact of climate change on hydrological extremes in the Yangtze River Basin, China. **Stochastic environmental research and risk assessment**, New York, v. 29, n. 3, p. 693-707, 2015. DOI: https://doi.org/10.1007/s00477-014-0957-5. Disponível em: https://link.springer.com/article/10.1007/s00477-014-0957-5. Acesso em: 28 jun. 2023.

GUNTU, R.K.; RATHINASAMY, M.; AGARWAL, A., SIVAKUMAR, B. Spatiotemporal variability of Indian rainfall using multiscale entropy. **Journal of Hydrology**, Amsterdam, v. 587, n. 8, p. 124916, 2020. DOI: https://doi.org/10.1016/j.jhydrol.2020.124916. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169420303760. Acesso em 28 jun. 2023.

GUO, J.; ZHOU, J.; QIN, H.; ZOU, Q.; LI, Q. Monthly streamflow forecasting based on improved support vector machine model. **Expert Systems with Applications**, United Kingdom, v. 38, n. 10, p. 13073-13081, 2011. DOI: https://doi.org/10.1016/j.eswa.2011.04.114. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0957417411006415. Acesso em: 29 jun. 2023.

HAGEN, J. S.; LEBLOIS, E.; LAWRENCE, D.; SOLOMATINE, D., SORTEBERG, A. Identifying major drivers of daily streamflow from large-scale atmospheric circulation with machine learning. **Journal of Hydrology**, Amsterdam, v. 596, n. 5, p. 126086, 2021. DOI: https://doi.org/10.1016/j.jhydrol.2021.126086. Disponível em: https://www.sciencedirect.com/science/article/pii/S0022169421001335. Acesso em: 1 jul. 2023.

HAMEL, L.H. **Knowledge discovery with support vector machines**. Hoboken, New Jersey: John Wiley & Sons, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and**

prediction. New York: Springer, 2009.

HE, Z.; WEN, X.; LIU, H.; DU, J. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. **Journal of Hydrology**, Amsterdam, v. 509, n. 2, p. 379-386, 2014. DOI: https://doi.org/10.1016/j.jhydrol.2013.11.054. https://www.sciencedirect.com/science/article/abs/pii/S0022169413008780. Acesso: 1 jul. 2023.

HEO, K.Y.; HA, K.J.; YUN, K.S.; LEE, S.S.; KIM, H.J.; WANG, B. Methods for uncertainty assessment of climate models and model predictions over East Asia. **International journal of climatology**, United Kingdom, v. 34, n. 2, p. 377-390, 2014. DOI: https://doi.org/10.1002/joc.3692. Disponível em: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3692. Acesso em 3 jul. 2023.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, Cambridge, Massachusetts, v. 9, n. 8, p. 1735-1780, 1997. DOI: 10.1162/neco.1997.9.8.1735. Disponível em: https://direct.mit.edu/neco/article-abstract/9/8/1735/6109/Long-Short-Term-Memory?redirectedFrom=fulltext. Acesso em 2 jul. 2023.

HONG, M.; WANG, D.; WANG, Y.; ZENG, X.; GE, S.; YAN, H.; SINGH, V.P. Mid-and long-term runoff predictions by an improved phase-space reconstruction model. **Environmental research**, Amsterdam, v. 148, n. 7, p. 560-573, 2016. DOI: https://doi.org/10.1016/j.envres.2015.11.024. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0013935115301523. Acesso em: 2 jul. 2023.

HRNJICA, B.; MEHR, A.D. Energy Demand Forecasting Using Deep Learning. *In*: AL-TURJMAN, F. (ed) **Smart Cities Performability, Cognition, & Security**. Switzerland, 2019. p 71-104. DOI: https://doi.org/10.1007/978-3-030-14718-1_4. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-14718-1_4. Acesso em: 2 jul. 2023.

HU, Z.; ZHANG, C.; LUO, G.; TENG, Z.; JIA, C. Characterizing cross-scale chaotic behaviors of the runoff time series in an inland river of Central Asia. **Quaternary International**, United Kingdom, v. 311, n. 10, p. 132-139, 2013. DOI: https://doi.org/10.1016/j.quaint.2013.07.031. Disponível em:

https://www.sciencedirect.com/science/article/abs/pii/S1040618213004308. Acesso em 3 jul. 2023.

HU, J.; LIU, Y.; SANG, Y.-F. Precipitation complexity and its spatial difference in the Taihu Lake Basin, China. **Entropy**, Switzerland, v. 21, n. 1, p. 48, 2019. DOI: https://doi.org/10.3390/e21010048. Disponível em: https://www.mdpi.com/1099-4300/21/1/48. Acesso em: 2 jul. 2023.

HUANG, N.E.; SHEN, Z.; LONG, S.R.; WU, M.C.; SHIH, H.H.; ZHENG, Q.; YEN, N.-C.; TUNG, C.C.; LIU, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. **Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences**, United Kingdom, v. 454, n. 1971, p. 903-995, 1998. DOI: https://doi.org/10.1098/rspa.1998.0193. Disponível em: https://royalsocietypublishing.org/doi/10.1098/rspa.1998.0193. Acesso em: 4 jul. 2023.

HUANG, S.C.; CHUANG, P.J.; WU, C.F.; LAI, H.J. Chaos-based support vector regressions for exchange rate forecasting. **Expert Systems with Applications**, United Kingdom, v. 37, n. 12, p. 8590-8598, 2010. DOI: https://doi.org/10.1016/j.eswa.2010.06.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0957417410005051. Acesso em: 4 jul. 2023.

HUANG, F.; XIA, Z.; ZHANG, N.; ZHANG, Y.; LI, J. Flow-complexity analysis of the upper reaches of the Yangtze River, China. **Journal of Hydrologic Engineering**, United States, v. 16, n. 11, p. 914-919, 2011. DOI: https://doi.org/10.1061/(ASCE)HE.1943-5584.0000392. Disponível em: https://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000392. Acesso em: 4 jul. 2023.

HUMPHREY, G.B.; GIBBS, M.S.; DANDY, G.C.; MAIER, H.R. A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network. **Journal of Hydrology**, Amsterdam, v. 540, n. 9, p. 623-640, 2016. DOI: https://doi.org/10.1016/j.jhydrol.2016.06.026. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169416303778. Acesso em: 4 jul. 2023.

IPCC, Climate Change. Synthesis Report. Contribution of working groups I. II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change, Switzerland, 2014,

v. 151. Disponível em: https://www.ipcc.ch/report/ar5/syr/. Acesso em: 4 jul. 2023.

JAYAWARDENA, A. W.; LAI, Feizhou. Analysis and prediction of chaos in rainfall and stream flow time series. **Journal of hydrology**, Amsterdam, v. 153, n. 1-4, p. 23-52, 1994. DOI: https://doi.org/10.1016/0022-1694(94)90185-6. Disponível: https://www.sciencedirect.com/science/article/abs/pii/0022169494901856. Acesso em: 4 jul. 2023.

JAYAWARDENA, A. W.; XU, P. C.; TSANG, F. L.; LI, W. K. Determining the structure of a radial basis function network for prediction of nonlinear hydrological time series. **Hydrological sciences journal**, United Kingdom, v. 51, n. 1, p. 21-44, 2006. DOI: https://doi.org/10.1623/hysj.51.1.21. Disponível em: https://www.tandfonline.com/doi/abs/10.1623/hysj.51.1.21. Acesso em: 4 jul. 2023.

JEMAI, S.; ELLOUZE, M.; ABIDA, H. Variability of precipitation in arid climates using the wavelet approach: case study of watershed of Gabes in South-East Tunisia. **Atmosphere**, Switzerland, v. 8, n. 9, p. 178, 2017. DOI: https://doi.org/10.3390/atmos8090178. Disponível em: https://www.mdpi.com/2073-4433/8/9/178. Acesso em: 4 jul. 2023.

JHONG, B.-C.; HUANG, J.; TUNG, C.-P. Spatial assessment of climate risk for investigating climate adaptation strategies by evaluating spatial-temporal variability of extreme precipitation. **Water Resources Management**, Netherlands, v. 33, n. 10, p. 3377-3400, 2019. DOI: https://doi.org/10.1007/s11269-019-02306-8. Disponível em: https://link.springer.com/article/10.1007/s11269-019-02306-8. Acesso em: 4 jul. 2023.

JIANG, J.; TANG, S.; LIU, R.; SIVAKUMAR, B.; WU, X.; PANG, T. A hybrid wavelet-Lyapunov exponent model for river water quality forecast. **Journal of Hydroinformatics**, United Kingdom, v. 23, n. 4, p. 864-878, 2021. DOI: https://doi.org/10.2166/hydro.2021.023. Disponível em: https://iwaponline.com/jh/article/23/4/864/82342/A-hybrid-wavelet-Lyapunov-exponent-model-for-river. Acesso em: 4 jul. 2023.

JIANG, Y.; BAO, X.; HAO, S.; ZHAO, H.; LI, X.; WU, X. Monthly streamflow forecasting using ELM-IPSO based on phase space reconstruction. **Water Resources Management**, Netherlands, v. 34, n. 11, p. 3515-3531, 2020. DOI: https://doi.org/10.1007/s11269-020-02631-3. Disponível em: https://link.springer.com/article/10.1007/s11269-020-02631-3. Acesso em: 4 jul. 2023.

KALIN, L.; ISIK, S.; SCHOONOVER, J.E.; LOCKABY, B.G. Predicting water quality in unmonitored watersheds using artificial neural networks. **Journal of environmental quality**, United States, v. 39, n. 4, p. 1429-1440, 2010. DOI: https://doi.org/10.2134/jeq2009.0441. Disponível em: https://acsess.onlinelibrary.wiley.com/doi/abs/10.2134/jeq2009.0441. Acesso em: 4 jul. 2023.

KANTZ, Holger. A robust method to estimate the maximal Lyapunov exponent of a time series. **Physics letters A**, Netherlands, v. 185, n. 1, p. 77-87, 1994. DOI: https://doi.org/10.1016/0375-9601(94)90991-1. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/0375960194909911. Acesso em: 4 jul. 2023.

KARMAKAR, S.; GOSWAMI, S.; CHATTOPADHYAY, S. Exploring the pre-and summer-monsoon surface air temperature over eastern India using Shannon entropy and temporal Hurst exponents through rescaled range analysis. **Atmospheric Research**, Netherlands, v. 217, n. 3, p. 57-62, 2019. DOI: https://doi.org/10.1016/j.atmosres.2018.10.007. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169809518309700. Acesso em: 5 jul. 2023.

KARUNASINGHE, D.S.K.; LIONG, S.-Y. Chaotic time series prediction with a global model: Artificial neural network. **Journal of Hydrology**, Amsterdam, v. 323, n. 1-4, p. 92-105, 2006. DOI: https://doi.org/10.1016/j.jhydrol.2005.07.048. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169405004142. Acesso em: 5 jul. 2023.

KASIVISWANATHAN K.S.; HE, J.; SUDHEER, K.P.; TAY, J.H. Potential application of wavelet neural network ensemble to forecast streamflow for flood management. **Journal of hydrology**, Amsterdam, v. 536, n. 5, p. 161-173, 2016. DOI: https://doi.org/10.1016/j.jhydrol.2016.02.044. Disponível em: https://www.sciencedirect.com/science/article/pii/S0022169416300786?casa_token=Frkhg8NZPU8AAAAA:pqfkk4VLurWZXDJnKq8-FapNThnAEXLKBJN2_arVNqH6frIWBztyncEY1bYquhg6MpQdyReAwFw. Acesso em: 5 jul. 2023.

KAYANO, M.T.; ANDREOLI, R.V. Relations of South American summer rainfall interannual

variations with the Pacific Decadal Oscillation. **International Journal of Climatology**, United Kingdom, v. 27, n. 4, p. 531-540, 2007. DOI: https://doi.org/10.1002/joc.1417. Disponível em: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.1417. Acesso em: 4 jul. 2023.

KAYANO, M.T.; ANDREOLI, R.V. Relationships between rainfall anomalies over northeastern Brazil and the El Niño–Southern Oscillation. **Journal of Geophysical Research: Atmospheres**, Hoboken, v. 111, n. D13, 2006. DOI: https://doi:10.1029/2005JD006142. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2005JD006142. Acesso em: 4 jul. 2023.

KAYANO, M.T.; ANDREOLI, R.V.; DE SOUZA, R.A.F. Pacific and Atlantic multidecadal variability relations to the El Niño events and their effects on the South American rainfall. **International Journal of Climatology**, United Kingdom, v. 40, n. 4, p. 2183-2200, 2020. DOI: https://doi.org/10.1002/joc.6326. Disponível: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.6326. Acesso em: 4 jul. 2023.

KĘDRA, M. Deterministic chaotic dynamics of raba river flow (polish carpathian mountains). **Journal of Hydrology**, Amsterdam, v. 509, n. 2, p. 474-503, 2014. DOI: https://doi.org/10.1016/j.jhydrol.2013.11.055. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169413008792. Acesso em: 4 jul. 2023.

KENNEL, M.B.; BROWN, R.; ABARBANEL, H.D.I. Determining embedding dimension for phase-space reconstruction using a geometrical construction. **Physical review A**, College Park, v. 45, n. 6, p. 3403, 1992. DOI: https://doi.org/10.1103/PhysRevA.45.3403. Disponível em: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.45.3403. Acesso em: 4 jul. 2023.

KHAN, M.M.H.; MUHAMMAD, N.S.; EL-SHAFIE, A. Wavelet based hybrid ANN-ARIMA models for meteorological drought forecasting. **Journal of Hydrology**, Amsterdam, v. 590, n. 11, p. 125380, 2020. DOI: https://doi.org/10.1016/j.jhydrol.2020.125380. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169420308404. Acesso em: 4 jul. 2023.

KHATIBI, R.; SIVAKUMAR, B.; GHORBANI, M.A.; KISI, O.; KOÇAK, K.; ZADEH,

D.F. Investigating chaos in river stage and discharge time series. **Journal of Hydrology**, Amsterdam, v. 414, n. 1, p. 108-117, 2012. DOI: https://doi.org/10.1016/j.jhydrol.2011.10.026. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169411007505. Acesso em: 5 jul. 2023.

KISI, O.; GORGIJ, A.D.; ZOUNEMAT-KERMANI, M.; MAHDAVI-MEYMAND, A.; KIM, S. Drought forecasting using novel heuristic methods in a semi-arid environment. **Journal of Hydrology**, Amsterdam, v. 578, n. 11, p. 124053, 2019. DOI: https://doi.org/10.1016/j.jhydrol.2019.124053. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169419307802. Acesso em: 5 jul. 2023.

KNIGHT, J.R.; FOLLAND, C.K.; SCAIFE, A.A. Climate impacts of the Atlantic multidecadal oscillation. **Geophysical Research Letters**, United States, v. 33, n. 17, 2006. DOI: https://doi.org/10.1029/2006GL026242. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2006GL026242. Acesso em: 5 jul. 2023.

KOUTSOYIANNIS, Demetris. The Hurst phenomenon and fractional Gaussian noise made easy. **Hydrological Sciences Journal**, United Kingdom, v. 47, n. 4, p. 573-595, 2002. DOI: https://doi.org/10.1080/02626660209492961. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/02626660209492961. Acesso em: 5 jul. 2023.

KOUTSOYIANNIS, D.; YAO, H.; GEORGAKAKOS, A. Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods/Prévision du débit du Nil à moyen terme: une comparaison de méthodes stochastiques et déterministes. **Hydrological Sciences Journal**, United Kingdom, v. 53, n. 1, p. 142-164, 2008. DOI: https://doi.org/10.1623/hysj.53.1.142. Disponível em: https://www.tandfonline.com/doi/abs/10.1623/hysj.53.1.142. Acesso em: 5 jul. 2023.

KUHN, Max et al. Package ‘caret’. **The R Journal**, United States, v. 223, n. 7, 2020. Disponível em: https://cran.r-project.org/web/packages/caret/caret.pdf. Acesso em: 5 jul. 2023.

KWON, Hyun-Han; LALL, Upmanu; KHALIL, Abedalrazq F. Stochastic simulation model for nonstationary time series using an autoregressive wavelet decomposition: Applications to

rainfall and temperature. **Water Resources Research**, United States, v. 43, n. 5, 2007. DOI: https://doi.org/10.1029/2006WR005258. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2006WR005258. Acesso em 5 jul. 2023.

LABAT, David. Recent advances in wavelet analyses: Part 1. A review of concepts. **Journal of Hydrology**, Amsterdam, v. 314, n. 1-4, p. 275-288, 2005. DOI: https://doi.org/10.1016/j.jhydrol.2005.04.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169405001769. Acesso em: 5 jul. 2023.

LABAT, David; SIVAKUMAR, B.; MANGIN, A. Evidence for deterministic chaos in long-term high-resolution karstic streamflow time series. **Stochastic Environmental Research and Risk Assessment**, New York, v. 30, n. 8, p. 2189-2196, 2016. DOI: https://doi.org/10.1007/s00477-015-1175-5. Disponível em: https://link.springer.com/article/10.1007/s00477-015-1175-5. Acesso em: 6 jul. 2023.

LELLEP, M.; PREXL, J.; LINKMANN, M.; ECKHARDT, B. Using machine learning to predict extreme events in the Hénon map. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, New York, v. 30, n. 1, p. 013113, 2020. DOI: https://doi.org/10.1063/1.5121844. Disponível em: https://pubs.aip.org/aip/cha/article-abstract/30/1/013113/1027422/Using-machine-learning-to-predict-extreme-events?redirectedFrom=fulltext. Acesso em: 6 jul. 2023.

LEONG, W.C.; BAHADORI, A.; ZHANG, J.; AHMAD, Z. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). **International Journal of River Basin Management**, United Kingdom, v. 19, n. 2, p. 149-156, 2021. DOI: https://doi.org/10.1080/15715124.2019.1628030. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/15715124.2019.1628030#:~:text=The%20current%20calculations%20of%20water,be%20predicted%20immediately%20using%20directly. Acesso em: 6 jul. 2023.

LI, Y.; SONG, Y.; LI, C. Selection of parameters for phase space reconstruction of chaotic time series. *In:* 2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 1., 2010, China. **Proceedings[..]**. China, 2010. p. 30-33.

DOI: 10.1109/BICTA.2010.5645296. Disponível em: https://ieeexplore.ieee.org/document/5645296. Acesso em: 6 jul. 2023.

LI, Z.; ZHANG, Y.-K. Multi-scale entropy analysis of Mississippi River flow. **Stochastic Environmental Research and Risk Assessment**, New York, v. 22, n. 4, p. 507-512, 2008. DOI: https://doi.org/10.1007/s00477-007-0161-y. Disponível em: https://link.springer.com/article/10.1007/s00477-007-0161-y. Acesso em: 6 jul. 2023

LIANG, Z.; LI, Y.; HU, Y.; LI, B.; WANG, J. A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework. **Theoretical and applied climatology**, Austria, v. 133, n. 1-2, p. 137-149, 2018. DOI: https://doi.org/10.1007/s00704-017-2186-6. Disponível em: https://link.springer.com/article/10.1007/s00704-017-2186-6. Acesso em: 6 jul. 2023.

LIANG, Z.; XIAO, Z.; WANG, J.; SUN, L.; LI, B.; HU, Y.; WU, Y. An improved chaos similarity model for hydrological forecasting. **Journal of Hydrology**, Amsterdam, v. 577, n. 10, p. 123953, 2019. DOI: https://doi.org/10.1016/j.jhydrol.2019.123953. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169419306730. Acesso em: 6 jul. 2023.

LIMA NETO, I.E.; WIEGAND, M.C.; CARLOS DE ARAÚJO, J. Redistribution des sédiments due à un réseau dense de réservoirs dans un grand bassin versant semi-aride du Brésil. **Hydrological sciences journal**, United Kingdom, v. 56, n. 2, p. 319-333, 2011. DOI: https://doi.org/10.1080/02626667.2011.553616. Acesso em: 5 jul. 2023.

LIU, M. et al. The applicability of LSTM-KNN model for real-time flood forecasting in different climate zones in China. **Water**, Switzerland, v. 12, n. 2, p. 440, 2020. DOI: https://doi.org/10.3390/w12020440. Disponível em: https://www.mdpi.com/2073-4441/12/2/440. Acesso em: 6 jul. 2023.

MA, H.-G.; HAN, C.-Z. Selection of embedding dimension and delay time in phase space reconstruction. **Frontiers of Electrical and Electronic Engineering in China**, Beijing, v. 1, n. 4, p. 111-114, 2006. DOI: https://doi.org/10.1007/s11460-005-0023-7. Disponível em: https://link.springer.com/article/10.1007/s11460-005-0023-7. Acesso em: 6 jul. 2023.

MAJUMDER, Swarnali; KANJILAL, Partha Pratim. Application of singular spectrum

analysis for investigating chaos in sea surface temperature. **Pure and Applied Geophysics**, Switzerland, v. 176, n. 8, p. 3769-3786, 2019. DOI: https://doi.org/10.1007/s00024-019-02140-4. Disponível em: https://link.springer.com/article/10.1007/s00024-019-02140-4. Acesso em: 5 jul. 2023.

MALVEIRA, V.T.C.; ARAÚJO, J.C.; GÜNTNER, A. Hydrological impact of a high-density reservoir network in semiarid northeastern Brazil. **Journal of Hydrologic Engineering**, United States, v. 17, n. 1, p. 109-117, 2012. DOI: https://doi.org/10.1061/(ASCE)HE.1943-5584.0000404. Disponível em: https://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000404. Acesso em: 5 jul. 2023.

MANN, H.B. Nonparametric tests against trend. **Econometrica: Journal of the econometric society**, p. 245-259, 1945.

MARENGO, J.A. Água e mudanças climáticas. **Estudos avançados**, São Paulo, v. 22, n. 63, p. 83-96, 2008. DOI: https://doi.org/10.1590/S0103-40142008000200006. Disponível em: https://www.scielo.br/j/ea/a/fXZzdm68cnzzt6Khr8zYx3L/abstract/?lang=pt. Acesso em 5 jul. 2023.

MARENGO, J.A.; TORRES, R.R.; ALVES, L.M. Drought in Northeast Brazil—past, present, and future. **Theoretical and Applied Climatology**, Austria, v. 129, n. 3-4, p. 1189-1200, 2017. DOI: https://doi.org/10.1007/s00704-016-1840-8. Disponível em: https://link.springer.com/article/10.1007/s00704-016-1840-8. Acesso em 5 jul. 2023.

MARWAN, N. How to avoid potential pitfalls in recurrence plot based data analysis. **International Journal of Bifurcation and Chaos**, Singapore, v. 21, n. 04, p. 1003-1017, 2011. DOI: https://doi.org/10.1142/S0218127411029008. Disponível em: https://www.worldscientific.com/doi/10.1142/S0218127411029008. Acesso em: 5 jul. 2023.

MARWAN. N.; ROMANO, M.C.; THIEL, M.; KURTHS, J. Recurrence plots for the analysis of complex systems. **Physics reports**, Netherlands, v. 438, n. 5-6, p. 237-329, 2007. DOI: https://doi.org/10.1016/j.physrep.2006.11.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0370157306004066. Acesso em: 5 jul. 2023.

MARWAN, N.; WESSEL, N.; MEYERFELDT, U.; SCHIRDEWAN, A.; KURTHS, J.

Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. **Physical review E**, College Park, v. 66, n. 2, p. 026702, 2002. DOI: https://doi.org/10.1103/PhysRevE.66.026702. Disponível em: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.66.026702. Acesso em 5 jul. 2023.

MEHR, A.D.; KAHYA, E.; OLYAIE, E. Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. **Journal of Hydrology**, Amsterdam, v. 505, n. 11, p. 240-249, 2013. DOI: https://doi.org/10.1016/j.jhydrol.2013.10.003. Disponível: https://www.sciencedirect.com/science/article/abs/pii/S0022169413007105. Acesso em: 5 jul. 2023.

MENG, E.; HUANG, S.; HUANG, Q.; FANG, W.; WU, L.; WANG, L. A robust method for non-stationary streamflow prediction based on improved EMD-SVM model. **Journal of hydrology**, Amsterdam, v. 568, n. 1, p. 462-478, 2019. DOI: https://doi.org/10.1016/j.jhydrol.2018.11.015. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169418308709. Acesso em: 5 jul. 2023.

MESHRAM, S.G.; SINGH, V.P.; KISI, O.; KARIMI, V.; MESHRAM, C. Application of artificial neural networks, support vector machine and multiple model-ANN to sediment yield prediction. **Water Resources Management**, Netherlands, v. 34, n. 15, p. 4561-4575, 2020. DOI: https://doi.org/10.1007/s11269-020-02672-8. Disponível em: https://link.springer.com/article/10.1007/s11269-020-02672-8. Acesso em: 5 jul. 2023.

MIHAILOVIĆ, D.T.; NIKOLIĆ-ĐORIĆ, E.; MALINOVIĆ-MILIĆEVIĆ, S.; SINGH, V.P.; MIHAILOVIĆ, A.; STOŠIĆ, T.; STOŠIĆ, B.; DREŠKOVIĆ, N. The choice of an appropriate information dissimilarity measure for hierarchical clustering of river streamflow time series, based on calculated Lyapunov exponent and Kolmogorov measures. **Entropy**, Switzerland, v. 21, n. 2, p. 215, 2019. DOI: https://doi.org/10.3390/e21020215. Disponível em: https://www.mdpi.com/1099-4300/21/2/215. Acesso em: 5 jul. 2023.

MISHRA, A.K.; SINGH, V.P. A review of drought concepts. **Journal of hydrology**, Amsterdam, v. 391, n. 1-2, p. 202-216, 2010. DOI: https://doi.org/10.1016/j.jhydrol.2010.07.012. Disponível:

https://www.sciencedirect.com/science/article/abs/pii/S0022169410004257. Acesso em: 5 jul. 2023.

MOCENNI, C.; FACCHINI, A.; VICINO, A. Comparison of recurrence quantification methods for the analysis of temporal and spatial chaos. **Mathematical and Computer Modelling**, Netherlands, v. 53, n. 7-8, p. 1535-1545, 2011. DOI: https://doi.org/10.1016/j.mcm.2010.04.008. Disponível em: https://www.sciencedirect.com/science/article/pii/S0895717710001986. Acesso em: 6 jul. 2023

NAZIR, H.M.; HUSSAIN, I.; FAISAL, M.; SHOUKRY, A.M.; GANI, S.; AHMAD, I. Development of multidecomposition hybrid model for hydrological time series analysis. **Complexity**, Egypt, v. 2019, n. 1, p. 1-14, 2019. DOI: https://doi.org/10.1155/2019/2782715. Disponível em: https://www.hindawi.com/journals/complexity/2019/2782715/. Acesso em 5 jul. 2023.

NEELIN, J.D.; BATTISTI, D.S.; HIRST, A.C.; JIN, F.F.; WAKATA, Y.; YAMAGATA, T.; ZEBIAK, S.E. ENSO theory. **Journal of Geophysical Research: Oceans**, United States, v. 103, n. C7, p. 14261-14290, 1998. DOI: https://doi.org/10.1029/97JC03424. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/97JC03424. Acesso em: 6 jul. 2023.

NHU, V.H.; KHOSRAVI, K; COOPER, J.R.; KARIMI, M.; KISI, O.; PHAM, B.T.; LYU, Z. Monthly suspended sediment load prediction using artificial intelligence: testing of a new random subspace method. **Hydrological Sciences Journal**, United Kingdom, v. 65, n. 12, p. 2116-2127, 2020. DOI: https://doi.org/10.1080/02626667.2020.1754419. Disponível em: https://www.tandfonline.com/doi/full/10.1080/02626667.2020.1754419. Acesso em: 6 jul. 2023.

NOORI, N.; KALIN, L.; ISIK, S. Water quality prediction using SWAT-ANN coupled approach. **Journal of Hydrology**, Amsterdam, v. 590, n. 11, p. 125220, 2020. DOI: https://doi.org/10.1016/j.jhydrol.2020.125220. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169420306806. Acesso em 6 jul. 2023.

NOURANI, V.; KISI, Ö.; KOMASI, M. Two hybrid artificial intelligence approaches for modeling rainfall–runoff process. **Journal of Hydrology**, Amsterdam, v. 402, n. 1-2, p. 41-

59, 2011. DOI: https://doi.org/10.1016/j.jhydrol.2011.03.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169411001612. Acesso em: 6 jul. 2023.

NOURANI, V.; KOMASI, M.; MANO, A. A multivariate ANN-wavelet approach for rainfall–runoff modeling. **Water resources management**, Netherlands, v. 23, n. 4, p. 2877-2894, 2009. DOI: https://doi.org/10.1007/s11269-009-9414-5. Disponível em: https://link.springer.com/article/10.1007/s11269-009-9414-5. Acesso em: 6 jul. 2023.

OGUNJO, S.T.; FUWAPE, I.A. Nonlinear characterization and interaction in teleconnection patterns. **Advances in Space Research**, United Kingdom, v. 65, n. 12, p. 2723-2732, 2020. DOI: https://doi.org/10.1016/j.asr.2020.03.023. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0273117720301782. Acesso: 6 jul. 2023.

OGUNJO, S., OLUSOLA, A., FUWAPE, I., & DUROWOJU, O. Temporal variation in deterministic chaos: the influence of Kainji dam on downstream stations along lower Niger River. **Arabian Journal of Geosciences**, Germany, v. 15, n. 3, p. 237, 2022. DOI: https://doi.org/10.1007/s12517-021-09297-0. Disponível em: https://link.springer.com/article/10.1007/s12517-021-09297-0. Acesso em: 7 jul. 2023.

OMBADI, M.; NGUYEN, P.; SOROOSHIAN, S.; HSU, K.L. Complexity of hydrologic basins: A chaotic dynamics perspective. **Journal of Hydrology**, Amsterdam, v. 597, n. 6, p. 126222, 2021. DOI: https://doi.org/10.1016/j.jhydrol.2021.126222. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169421002699. Acesso em 5 jul. 2023.

OUYANG, Q.; LU, W.; XIN, X.; ZHANG, Y.; CHENG, W.; YU, T. Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction. **Water resources management**, Netherlands, v. 30, n. 7, p. 2311-2325, 2016. DOI: https://doi.org/10.1007/s11269-016-1288-8. Disponível em: https://link.springer.com/article/10.1007/s11269-016-1288-8. Acesso em: 5 jul. 2023.

PAPACHARALAMPOUS, G.; TYRALIS, H.; KOUTSOYIANNIS, D. Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. **Stochastic environmental research and risk assessment**, New York, v. 33, n. 2,

p. 481-514, 2019. DOI: https://doi.org/10.1007/s00477-018-1638-6. Disponível em: https://link.springer.com/article/10.1007/s00477-018-1638-6. Acesso em: 6 jul. 2023.

PAPACHARALAMPOUS, G.A.; TYRALIS, H.; KOUTSOYIANNIS D. Forecasting of geophysical processes using stochastic and machine learning algorithms. **European Water**, Greece, v. 59, p. 161-168, 2017. Disponível em: https://www.researchgate.net/profile/Georgia-Papacharalampous/publication/318457064_Forecasting_of_geophysical_processes_using_stochastic_and_machine_learning_algorithms/links/5a3a2eae0f7e9baa5018b0aa/Forecasting-of-geophysical-processes-using-stochastic-and-machine-learning-algorithms.pdf. Acesso em: 6 jul. 2023.

PATHAK, P.; KALRA, A.; AHMAD, S.; BERNARDEZ, M. Wavelet-aided analysis to estimate seasonal variability and dominant periodicities in temperature, precipitation, and streamflow in the Midwestern United States. **Water resources management**, Netherlands, v. 30, n. 13, p. 4649-4665, 2016. DOI: https://doi.org/10.1007/s11269-016-1445-0. Disponível em: https://link.springer.com/article/10.1007/s11269-016-1445-0. Acesso em: 6 jul. 2023.

PENG, T.; ZHOU, J.; ZHANG, C.; FU, W. Streamflow forecasting using empirical wavelet transform and artificial neural networks. **Water**, Switzerland, v. 9, n. 6, p. 406, 2017. DOI: https://doi.org/10.3390/w9060406. Disponível em: https://www.mdpi.com/2073-4441/9/6/406. Acesso em: 6 jul. 2023.

PHAM, V.-T.; VAIDYANATHAN, S.; KAPITANIAK, T. Complexity, dynamics, control, and applications of nonlinear systems with multistability. **Complexity**, Egypt, v. 2020, p. 1-7, 2020. DOI: https://doi.org/10.1155/2020/8510930. Disponível em: https://www.hindawi.com/journals/complexity/2020/8510930/. Acesso em 6 jul. 2023.

PHAM, B.T.; BUI, D.; PRAKASH, I.; DHOLAKIA, M. Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. **J. Geomatics**, India, v. 10, n. 1, p. 71-79, 2016. Disponível em: https://www.researchgate.net/profile/Indra-Prakash-4/publication/308723439_A_Comparison_Study_of_predictive_ability_of_Support_Vector_Machine_and_Naive_Bayes_Tree_Methods_in_Landslide_Susceptibility_Assessment_at_an_Area_between_Tehri_Garhwal_and_Pauri_Garhwal_Uttarakhand_St/links/61a601d60cfb7a4f

aa74500a/A-Comparison-Study-of-predictive-ability-of-Support-Vector-Machine-and-Naive-Bayes-Tree-Methods-in-Landslide-Susceptibility-Assessment-at-an-Area-between-Tehri-Garhwal-and-Pauri-Garhwal-Uttarakhand-S.pdf. Acesso em 6 jul. 2023.

PINCUS, Steven M. Approximate entropy as a measure of system complexity. **Proceedings of the National Academy of Sciences**, United States, v. 88, n. 6, p. 2297-2301, 1991. Disponível em: https://www.pnas.org/doi/abs/10.1073/pnas.88.6.2297. Acesso em: 6 jul. 2023.

PONTES FILHO, J.D.; SOUZA FILHO, F.A.; MARTINS, E.S.P.R.; STUDART, T.M.C. Copula-based multivariate frequency analysis of the 2012–2018 drought in Northeast Brazil. **Water**, Switzerland, v. 12, n. 3, p. 834, 2020. DOI: https://doi.org/10.3390/w12030834. Disponível em: https://www.mdpi.com/2073-4441/12/3/834. Acesso em 6 jul. 2023.

QUILTY, John; ADAMOWSKI, Jan. Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. **Journal of hydrology**, Amsterdam, v. 563, n. 8, p. 336-353, 2018. DOI: https://doi.org/10.1016/j.jhydrol.2018.05.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169418303317. Acesso em: 7 jul. 2023.

RAGETTLI, S.; ZHOU, J.; WANG, H.; LIU, C.; GUO, L. Modeling flash floods in ungauged mountain catchments of China: A decision tree learning approach for parameter regionalization. **Journal of Hydrology**, Amsterdam, v. 555, n. 12, p. 330-346, 2017. DOI: https://doi.org/10.1016/j.jhydrol.2017.10.031. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169417307060. Acesso em: 6 jul. 2023.

RAMADEVI, B.; BINGI, K. Chaotic time series forecasting approaches using machine learning techniques: A review. **Symmetry**, Switzerland, v. 14, n. 5, p. 955, 2022. DOI: https://doi.org/10.3390/sym14050955. Disponível em: https://www.mdpi.com/2073-8994/14/5/955. Acesso em 6 jul. 2023.

RAMARAO, M.V.S.; SANJAY, J.; MUJUMDAR, K.R.M.; BAZAZ, A.; REVI, A. On observed aridity changes over the semiarid regions of India in a warming climate. **Theoretical and applied climatology**, Austria, v. 136, n. 1-2, p. 693-702, 2019.

DOI: https://doi.org/10.1007/s00704-018-2513-6. Disponível em: https://link.springer.com/article/10.1007/s00704-018-2513-6#:~:text=It%20can%20also%20be%20noticed,the%20semiarid%20regions%20of%20India. Acesso em: 6 jul. 2023.

REMESAN, R.; MATHEW, J. **Hydrological data driven modelling**. Switzerland: Springer International Publishing, 2016.

REN, Y.; SUGANTHAN, P.N.; SRIKANTH, N. A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods. **IEEE Transactions on Sustainable Energy**, New York , v. 6, n. 1, p. 236-244, 2014. DOI: https://doi.org/10.1109/TSTE.2014.2365580. Disponível em: https://ieeexplore.ieee.org/document/6979269. Acesso em: 6 jul. 2023.

RICHMAN, J.S.; MOORMAN, J.R. Physiological time-series analysis using approximate entropy and sample entropy. **American journal of physiology-heart and circulatory physiology**, United States, v. 278, n. 6, p. H2039-H2049, 2000. DOI: https://doi.org/10.1152/ajpheart.2000.278.6.H2039. Disponível em: https://journals.physiology.org/doi/full/10.1152/ajpheart.2000.278.6.H2039. Acesso em: 6 jul. 2023.

RICKLES, Dean; HAWE, Penelope; SHIELL, Alan. A simple guide to chaos and complexity. **Journal of Epidemiology & Community Health**, United Kingdom, v. 61, n. 11, p. 933-937, 2007.
DOI: https://doi.org/10.1136%2Fjech.2006.054254. Disponível em: https://jech.bmj.com/content/61/11/933.short?casa_token=092QgigRclYAAAAA:0GBKn7i2x8qXRS93IxqWw-uK7WzGjkbvKi5-LbQ0_dOnTUqfuZJCwU_4ORKupfh5WDEV_rdei5w. Acesso em: 6 jul. 2023.

ROCHA, R.V.; SOUZA FILHO, F.A.; SILVA, S.M.O. Análise da Relação entre a Precipitação Média do Reservatório Orós, Brasil-Ceará, e os índices PDO e AMO Através da Análise de Changepoints e Transformada de Ondeletas. **Revista Brasileira de Meteorologia**, São Paulo, v. 34, n. 1, p. 139-149, 2019. DOI: https://doi.org/10.1590/0102-77863340034. Disponível em: https://www.scielo.br/j/rbmet/a/hLfS7GJCcLZVQyRbNVR37VJ/. Acesso em: 6 jul. 2023.

RODRIGUEZ-ITURBE, I.; FEBRES DE POWER, B.; SHARIFI, M.B.; GEORGAKAKOS, K.P. Chaos in rainfall. **Water resources research**, United States, v. 25, n. 7, p. 1667-1675, 1989. DOI: https://doi.org/10.1029/WR025i007p01667. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR025i007p01667. Acesso em: 6 jul. 2023.

ROLIM, L.Z.R.; DE SOUZA FILHO, F.A. Shift detection in hydrological regimes and pluriannual low-frequency streamflow forecasting using the hidden Markov model. **Water**, Switzerland, v. 12, n. 7, p. 2058, 2020. DOI: https://doi.org/10.3390/w12072058. Disponível em: https://www.mdpi.com/2073-4441/12/7/2058. Acesso em: 6 jul. 2023.

ROLIM, L.Z.R.; SILVA, S.M.O.; DE SOUZA FILHO, F.A. Analysis of precipitation dynamics at different timescales based on entropy theory: an application to the State of Ceara, Brazil. **Stochastic Environmental Research and Risk Assessment**, New York, v. 36, n. 8, p. 2285-2301, 2022. DOI: https://doi.org/10.1007/s00477-021-02112-y. Disponível em: https://link.springer.com/article/10.1007/s00477-021-02112-y. Acesso em: 6 jul. 20203.

RÖSCH, Angi; SCHMIDBAUER, Harald. **WaveletComp 1.1**: A guided tour through the R package, Germany, 2016. Disponível em: http://www.hs-stat.com/projects/WaveletComp/WaveletComp_guided_tour.pdf. Acesso em 6 jul. 2023.

ROSENSTEIN, M.T.; COLLINS, J.J.; DE LUCA, C.J. A practical method for calculating largest Lyapunov exponents from small data sets. **Physica D: Nonlinear Phenomena**, Netherlands, v. 65, n. 1-2, p. 117-134, 1993. DOI: https://doi.org/10.1016/0167-2789(93)90009-P. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/016727899390009P. Acesso em: 6 jul. 2023.

SALAS, J.D.; OBEYSEKERA, J.T.B. ARMA model identification of hydrologic time series. **Water Resources Research**, United States, v. 18, n. 4, p. 1011-1021, 1982. DOI: https://doi.org/10.1029/WR018i004p01011. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR018i004p01011. Acesso em: 6 jul. 2023.

SANG, Y.-F. A review on the applications of wavelet transform in hydrology time series

analysis. **Atmospheric research**, Netherlands, v. 122, n. 3, p. 8-15, 2013. DOI: https://doi.org/10.1016/j.atmosres.2012.11.003. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169809512003924?casa_token=4Jc61u2F5gEAAAAA:JKr0UGdn7VJCfeBBiD8fBrPj8wJi8N3w87zFoT_UYfdp6Likwg8eEUtKu9cEandNicnkV93-kAI. Acesso em: 6 jul. 2023.

SANTANA, L.I.T.; DA SILVA, A.S.A.; MENEZES, R.S.C.; STOSIC, T. Análise de quantificação de recorrência de séries temporais mensais de chuva em Pernambuco, Brasil. **Research, Society and Development**, São Paulo, v. 9, n. 9, p. e637997737-e637997737, 2020a. DOI: https://doi.org/10.33448/rsd-v9i9.7737. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/7737. Acesso em 6 jul. 2023.

SANTANA, L.I.T.; BARRETO I.D.C.; ARAÚJO, L.S.; STOSIC, T. Recurrence quantification analysis of São Francisco river flow: hydrological alterations caused by the construction of Sobradinho dam. **Research, Society and Development**, São Paulo, v. 9, n. 11, p. e87491110467-e87491110467, 2020b. DOI: https://doi.org/10.33448/rsd-v9i11.10467. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/10467. Acesso em: 6 jul. 2023.

SARAIVA, S.V.; CARVALHO, F.O.; SANTOS, C.A.G.; BARRETO, L.C.; FREIRE, P.K.D.M.M. Daily streamflow forecasting in Sobradinho Reservoir using machine learning models coupled with wavelet transform and bootstrapping. **Applied Soft Computing**, Netherlands, v. 102, n. 4, p. 107081, 2021. DOI: https://doi.org/10.1016/j.asoc.2021.107081. Disponível em: https://www.sciencedirect.com/science/article/pii/S1568494621000041?casa_token=V3_eHwZ7DicAAAAA:zLqbMRR3nl6sGFzOR_tDdXqYFctuD9LcuTIlPJ09YKnpMevmKlm7l-PPfnq0hJBFV-uapHiY9sM. Acesso em: 6 jul. 2023.

SEN, P.K. Estimates of the regression coefficient based on Kendall's tau. **Journal of the American statistical association**, United Kingdom, v. 63, n. 324, p. 1379-1389, 1968. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/01621459.1968.10480934. Acesso em: 6 jul. 2023.

SHU, Z.; JESSON, M.; STERLING, M. Nonlinear dynamic analysis of daily rainfall variability across the UK from 1989 to 2018. **Journal of Hydrology**, Amsterdam, v. 603, n.

12, p. 126849, 2021. DOI: https://doi.org/10.1016/j.jhydrol.2021.126849. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169421008994. Acesso em: 7 jul. 2023.

SIEK, M.B.L.A.; SOLOMATINE, D.P. Nonlinear chaotic model for predicting storm surges. **Nonlinear Processes in Geophysics**, Germany, v. 17, n. 5, p. 405-420, 2010. DOI: https://doi.org/10.5194/npg-17-405-2010. Disponível em: https://npg.copernicus.org/articles/17/405/2010/. Acesso em: 7 jul. 2023.

SILVA, A. S. A., BARRETO, I.D.C.; CUNHA-FILHO, M.; MENEZES, R.S.C.; STOSIC, B.; STOSIC, T. Multiscale Complexity Analysis of Rainfall in Northeast Brazil. **Water**, Switzerland, v. 13, n. 22, p. 3213, 2021. DOI: https://doi.org/10.3390/w13223213. Disponível em: https://www.mdpi.com/2073-4441/13/22/3213. Acesso em: 7 jul. 2023.

SILVA, S. M. O. da, SOUZA, F. de A., AQUINO, S. H. S. Avaliação do risco da alocação de água em período de escassez hídrica: o caso do Sistema Jaguaribe-Metropolitano. **Engenharia Sanitária e Ambiental**, São Paulo, v. 22, p. 749-760, 2017. DOI: https://doi.org/10.1590/S1413-41522017161303. Disponível em: https://www.scielo.br/j/esa/a/XGtJXr84w83JkSnpxvwWGwS/?lang=pt. Acesso em: 7 jul. 2023.

SINGH, P.K.; CHUDASAMA, H. Pathways for climate change adaptations in arid and semi-arid regions. **Journal of cleaner production**, United Kingdom, v. 284, n. 2, p. 124744, 2021. DOI: https://doi.org/10.1016/j.jclepro.2020.124744. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652620347880. Acesso em: 7 jul. 2023.

SINGH, S.K. Long-term streamflow forecasting based on ensemble streamflow prediction technique: a case study in New Zealand. **Water resources management**, Netherlands, v. 30, n. 7, p. 2295-2309, 2016. DOI: https://doi.org/10.1007/s11269-016-1289-7. Disponível em: https://link.springer.com/article/10.1007/s11269-016-1289-7. Acesso em: 7 jul. 2023.

SIVAKUMAR, B. Chaos theory in hydrology: important issues and interpretations. **Journal of hydrology**, Amsterdam, v. 227, n. 1-4, p. 1-20, 2000. DOI: https://doi.org/10.1016/S0022-1694(99)00186-9. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169499001869. Acesso em: 7 jul.

2023.

SIVAKUMAR, B. Correlation dimension estimation of hydrological series and data size requirement: myth and reality. **Hydrological sciences journal**, United Kingdom, v. 50, n. 4, 2005. DOI: https://doi.org/10.1623/hysj.2005.50.4.591. Disponível em: https://www.tandfonline.com/doi/abs/10.1623/hysj.2005.50.4.591. Acesso em: 7 jul. 2023.

SIVAKUMAR, B. Nonlinear dynamics and chaos in hydrologic systems: latest developments and a look forward. **Stochastic Environmental Research and Risk Assessment**, New York, v. 23, n. 7, p. 1027-1036, 2009. DOI: https://doi.org/10.1007/s00477-008-0265-z. Disponível em: https://link.springer.com/article/10.1007/s00477-008-0265-z. Acesso em: 7 jul. 2023.

SIVAKUMAR, B. WOLDEMESKEL, F. M., PUENTE, CE. Nonlinear analysis of rainfall variability in Australia. **Stochastic Environmental Research and Risk Assessment**, New York, v. 28, n. 1, p. 17-27, 2014. DOI: https://doi.org/10.1007/s00477-013-0689-y. Disponível em: https://link.springer.com/article/10.1007/s00477-013-0689-y. Acesso em: 7 jul. 2023.

SIVAKUMAR, B. **Chaos in hydrology:** bridging determinism and stochasticity. Dordrecht, Springer, 2016.

SIVAKUMAR, B.; BERNDTSSON, R.; OLSSON, J.; JINNO, K. Evidence of chaos in the rainfall-runoff process. **Hydrological Sciences Journal**, United Kingdom, v. 46, n. 1, p. 131-145, 2001. DOI: https://doi.org/10.1080/02626660109492805. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/02626660109492805. Acesso em: 7 jul. 2023.

SIVAKUMAR, B.; PERSSON, M.; BERNDTSSON, R.; UVO, C.B. Is correlation dimension a reliable indicator of low-dimensional chaos in short hydrological time series?. **Water resources research**, United States, v. 38, n. 2, p. 3-1-3-8, 2002. DOI: https://doi.org/10.1029/2001WR000333. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2001WR000333. Acesso em: 7 jul. 2023.

SIVAKUMAR, B.; SINGH, V.P. Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework. **Hydrology and Earth System Sciences**, Germany, v. 16, n. 11, p. 4119-4131, 2012. DOI: https://doi.org/10.5194/hess-16-4119-2012. Disponível em: https://hess.copernicus.org/articles/16/4119/2012/. Acesso em: 7 jul. 2023.

SOUZA FILHO, F.A.; LALL, U. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm. **Water Resources Research**, United States, v. 39, n. 11, p. 1307, 2003. DOI: https://doi.org/10.1029/2002WR001373. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2002WR001373. Acesso em: 7 jul. 2023.

STEINWART, I.; CHRISTMANN, A. **Support vector machines**. New York: Springer Science & Business Media, 2008.

STONE, L.; SAPARIN, P.I.; HUPPERT, A.; PRICE, C. El Nino chaos: the role of noise and stochastic resonance on the ENSO cycle. **Geophysical research letters**, United States, v. 25, n. 2, p. 175-178, 1998. DOI: https://doi.org/10.1029/97GL53639. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97GL53639. Acesso em: 7 jul. 2023.

STROZZI, F.; ZALDÍVAR, J.-M.; ZBILUT, J.P. Application of nonlinear time series analysis techniques to high-frequency currency exchange data. **Physica A: Statistical Mechanics and its Applications**, Netherlands, v. 312, n. 3-4, p. 520-538, 2002. DOI: https://doi.org/10.1016/S0378-4371(02)00846-4. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0378437102008464. Acesso em: 7 jul. 2023.

SUGIHARA, G.; MAY, R.M. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. **Nature**, New York, v. 344, n. 6268, p. 734-741, 1990. DOI: https://doi.org/10.1038/344734a0. Disponível em: https://www.nature.com/articles/344734a0. Acesso em: 7 jul. 2023.

SUN, Y.; NIU, J.; SIVAKUMAR, B. A comparative study of models for short-term streamflow forecasting with emphasis on wavelet-based approach. **Stochastic Environmental Research and Risk Assessment**, New York, v. 33, n. 10, p. 1875-1891, 2019. DOI: https://doi.org/10.1007/s00477-019-01734-7. Disponível: https://link.springer.com/article/10.1007/s00477-019-01734-7. Acesso em: 7 jul. 2023.

SZOLGAYOVA, E.; PARAJKA, J.; BLÖSCHL, G.; BUCHER, C. Long term variability of the Danube River flow and its relation to precipitation and air temperature. **Journal of Hydrology**, Amsterdam, v. 519, n. 11, p. 871-880, 2014. DOI:

https://doi.org/10.1016/j.jhydrol.2014.07.047. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169414005812. Acesso em: 7 jul. 2023.

TAKENS, F. Detecting strange attractors in turbulence. *In:* Dynamical Systems and Turbulence, Warwick 1980, 2006, Berlin. **Proceedings of a symposium held at the University of Warwick 1979/80**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 366-381.

TAN, X.; GAN, T.Y. Multifractality of Canadian precipitation and streamflow. **International Journal of Climatology**, United Kingdom, v. 37, n. 1, p. 1221-1236, 2017. DOI: https://doi.org/10.1002/joc.5078. Disponível em: https://rmets.onlinelibrary.wiley.com/doi/epdf/10.1002/joc.5078. Acesso em: 7 jul. 2023.

TANG, C.; CHEN, D.; CROSBY, B.T.; PIECHOTA, T.C.; WHEATON, J.M. Is the PDO or AMO the climate driver of soil moisture in the Salmon River Basin, Idaho?. **Global and Planetary Change**, Netherlands, v. 120, n. 7, p. 16-23, 2014. DOI: https://doi.org/10.1016/j.gloplacha.2014.05.008. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0921818114001039. Acesso em: 7 jul. 2023.

TANG, Y.; KURTHS, J.; LIN, W.; OTT, E.; KOCAREV, L. Introduction to focus issue: When machine learning meets complex systems: Networks, chaos, and nonlinear dynamics. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, New York, v. 30, n. 6, p. 063151, 2020. Disponível em: https://pubs.aip.org/aip/cha/article-abstract/30/6/063151/1030253/Introduction-to-Focus-Issue-When-machine-learning?redirectedFrom=fulltext. Acesso em: 7 jul. 2023.

TAYLOR, K.E. Summarizing multiple aspects of model performance in a single diagram. **Journal of geophysical research: atmospheres**, Hoboken, v. 106, n. D7, p. 7183-7192, 2001. DOI: https://doi.org/10.1029/2000JD900719. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JD900719. Acesso em: 7 jul. 2023.

TEHRANY, M.S.; PRADHAN, B.; JEBUR, M.N. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. **Journal of**

hydrology, Amsterdam, v. 512, n. 5, p. 332-343, 2014. DOI: doi:10.1016/j.jhydrol.2014.03.008. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169414001826. Acesso em: 7 jul. 2023.

THERNEAU, T.M.; ATKINSON, E.J. An introduction to recursive partitioning using the RPART routines, Minnesota: Mayo Foundation, 2015. v. 67. Disponível em: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf. Acesso em: 7 jul. 2023.

THOMAS, A; FIERING, M. Mathematical synthesis of streamflow sequences for the analysis of river basin by simulation. **Design of water resources-systems**, Cambridge, p. 459-493, 1962. Disponível em: https://www.degruyter.com/document/doi/10.4159/harvard.9780674421042.c15/html. Acesso em: 7 jul. 2023.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, United Kingdom, v. 58, n. 1, p. 267-288, 1996. Disponível em: https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x. Acesso em: 7 jul 2023.

TIEN BUI, D.; PRADHAN, B.; LOFMAN, O.; REVHAUG, I. Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naive Bayes Models. **Mathematical problems in Engineering**, United States, v. 2012, p. 16, 2012. DOI: doi:10.1155/2012/974638. Disponível em: https://www.hindawi.com/journals/mpe/2012/974638/. Acesso em: 7 jul. 2023.

TONGAL, H.; BERNDTSSON, R. Phase-space reconstruction and self-exciting threshold modeling approach to forecast lake water levels. **Stochastic environmental research and risk assessment**, New York, v. 28, n. 4, p. 955-971, 2014. DOI: https://doi.org/10.1007/s00477-013-0795-x. Disponível em: https://link.springer.com/article/10.1007/s00477-013-0795-x. Acesso em: 7 jul. 2023.

TONGAL, H.; SIVAKUMAR, B. Entropy analysis for spatiotemporal variability of seasonal, low, and high streamflows. **Stochastic Environmental Research and Risk Assessment**, New York, v. 33, n. 1, p. 303-320, 2019. DOI: https://doi.org/10.1007/s00477-018-1615-0. Disponível: https://link.springer.com/article/10.1007/s00477-018-1615-

0#:~:text=The%20entropy%20method%20is%20used,we%20employ%20the%20kriging%20method. Acesso em: 7 jul. 2023.

TONGAL, H. Comparison of local and global approximators in multivariate chaotic forecasting of daily streamflow. **Hydrological Sciences Journal**, United Kingdom, v. 65, n. 7, p. 1129-1144, 2020. DOI: https://doi.org/10.1080/02626667.2020.1732983. Disponível em: https://www.tandfonline.com/doi/full/10.1080/02626667.2020.1732983. Acesso em: 7 jul. 2023

TORRENCE, C.; COMPO, G.P. A practical guide to wavelet analysis. **Bulletin of the American Meteorological society**, United States, v. 79, n. 1, p. 61-78, 1998. Disponível em: https://journals.ametsoc.org/view/journals/bams/79/1/1520-0477_1998_079_0061_apgtwa_2_0_co_2.xml. Acesso em 7 jul. 2023.

TORRENCE, C.; WEBSTER, P.J. Interdecadal changes in the ENSO–monsoon system. **Journal of climate**, United States, v. 12, n. 8, p. 2679-2690, 1999. Disponível em: https://journals.ametsoc.org/view/journals/clim/12/8/1520-0442_1999_012_2679_icitem_2.0.co_2.xml. Acesso em 7 jul. 2023.

TORRES, M.E.; COLOMINAS, M.A.; SCHLOTTHAUER, G.; FLANDRIN, P. A complete ensemble empirical mode decomposition with adaptive noise. *In*: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2011, Prague, Czech Republic. **Proceedings [...]**. Czech Republic, 2011. p. 4144-4147. DOI: https://doi.org/10.1109/ICASSP.2011.5947265. Disponível em: https://ieeexplore.ieee.org/document/5947265. Acesso em: 7 jul. 2023.

TURNER, John R.; BAKER, Rose M. Complexity theory: An overview with potential applications for the social sciences. **Systems**, Switzerland, v. 7, n. 1, p. 4, 2019. DOI: https://doi.org/10.3390/systems7010004. Disponível em: https://www.mdpi.com/2079-8954/7/1/4. Acesso em: 7 jul. 2023.

VAPNIK, V. **The nature of statistical learning theory**. New York: Springer science & business media, 1995.

VIGNESH, R.; JOTHIPRAKASH, V.; SIVAKUMAR, B. Streamflow variability and classification using false nearest neighbor method. **Journal of Hydrology**, Amsterdam, v.

531, n. 12, p. 706-715, 2015. DOI: https://doi.org/10.1016/j.jhydrol.2015.10.056. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169415008276. Acesso em: 7 jul. 2023.

VLACHAS, P.R.; BYEON, W.; WAN, Z.Y.; SAPSIS, T.P.; KOUMOUTSAKOS, P. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. **Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences**, United Kingdom, v. 474, n. 2213, p. 20170844, 2018. DOI: https://doi.org/10.1098/rspa.2017.0844. Disponível em: https://royalsocietypublishing.org/doi/10.1098/rspa.2017.0844. Acesso em: 7 jul. 2023.

WANG, F.; HUANG, G.H.; FAN, Y.; LI, Y.P. Development of clustered polynomial chaos expansion model for stochastic hydrological prediction. **Journal of Hydrology**, Amsterdam, v. 595, n. 4, p. 126022, 2021. DOI: https://doi.org/10.1016/j.jhydrol.2021.126022. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S002216942100069X. Acesso em: 7 jul. 2023.

WANG, L.; LI, X.; MA, C.; BAI, Y. Improving the prediction accuracy of monthly streamflow using a data-driven model based on a double-processing strategy. **Journal of Hydrology**, Amsterdam, v. 573, n. 6, p. 733-745, 2019. DOI: https://doi.org/10.1016/j.jhydrol.2019.03.101. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169419303294. Acesso em: 7 jul. 2023.

WANG, K. et al. Performance improvement of machine learning models via wavelet theory in estimating monthly river streamflow. **Engineering Applications of Computational Fluid Mechanics**, United Kingdom, v. 16, n. 1, p. 1833-1848, 2022. DOI: https://doi.org/10.1080/19942060.2022.2119281. Disponível em: https://www.tandfonline.com/doi/full/10.1080/19942060.2022.2119281. Acesso em: 7 jul. 2023.

WEN, X.; FENG, Q.; DEO, R.C.; WU, M.; YIN, Z.; YANG, L.; SINGH, V.P. Two-phase extreme learning machines integrated with the complete ensemble empirical mode decomposition with adaptive noise algorithm for multi-scale runoff prediction problems. **Journal of hydrology**, Amsterdam, v. 570, n. 3, p. 167-184, 2019. DOI:

https://doi.org/10.1016/j.jhydrol.2018.12.060. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169419300381. Acesso em: 7 jul. 2023.

WOLF, A.; SWIFT, J.B.; SWINNEY, H.L.; VASTANO, J.A. Determining Lyapunov exponents from a time series. **Physica D: nonlinear phenomena**, Netherlands, v. 16, n. 3, p. 285-317, 1985. DOI: https://doi.org/10.1016/0167-2789(85)90011-9. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/0167278985900119. Acesso em: 7 jul. 2023.

WU, Y.; LIU, S.; YAN, W.; XIA, J.; XIANG, W.; WANG, K.; LUO, Q.; FU, W.; YUAN, W. Climate change and consequences on the water cycle in the humid Xiangjiang River Basin, China. **Stochastic environmental research and risk assessment**, New York, v. 30, n. 1, p. 225-235, 2016. DOI: https://doi.org/10.1007/s00477-015-1073-x. Disponível em: https://link.springer.com/article/10.1007/s00477-015-1073-x. Acesso em: 7 jul. 2023.

WU, Z.; HUANG, N.E. Ensemble empirical mode decomposition: a noise-assisted data analysis method. **Advances in adaptive data analysis**, Singapore, v. 1, n. 01, p. 1-41, 2009. DOI: https://doi.org/10.1142/S1793536909000047. Disponível em: https://www.worldscientific.com/doi/10.1142/S1793536909000047. Acesso em: 7 jul. 2023.

XAVIER, S.F.A.; DA SILVA JALE, J.; STOSIC, T.; DOS SANTOS, C.A.C.; SINGH, V.P. An application of sample entropy to precipitation in Paraíba State, Brazil. **Theoretical and Applied Climatology**, Austria, v. 136, n. 1-2, p. 429-440, 2019. DOI: https://doi.org/10.1007/s00704-018-2496-3. Disponível em: https://link.springer.com/article/10.1007/s00704-018-2496-3. Acesso em: 7 jul. 2023.

XU, J.; CHEN, Y.; LI, W.; JI, M.; DONG, S. The complex nonlinear systems with fractal as well as chaotic dynamics of annual runoff processes in the three headwaters of the Tarim River. **Journal of Geographical Sciences**, China, v. 19, n. 1, p. 25-35, 2009. DOI: https://doi.org/10.1007/s11442-009-0025-0. Disponível em: https://link.springer.com/article/10.1007/s11442-009-0025-0. Acesso em: 7 jul. 2023.

XU, T.; LIANG, F. Machine learning for hydrologic sciences: An introductory overview. **Wiley Interdisciplinary Reviews: Water**, United States, v. 8, n. 5, p. e1533, 2021. DOI: https://doi.org/10.1002/wat2.1533. Disponível em:

https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wat2.1533?casa_token=2vRypA8YJaoAAAAA:ZQQQU2MAywQ7DtD-3hCSUS214bJp9mC4dhHKPZDKPUM-iJTgd1zuYqKkD1pbcUi7ZOtO0fTIbiai1It3. Acesso em: 7 jul. 2023.

YAN, B.; CHAN, P.W.; LI, Q.; HE, Y.; SHU, Z. Dynamic analysis of meteorological time series in Hong Kong: A nonlinear perspective. **International Journal of Climatology**, United Kingdom, v. 41, n. 10, p. 4920-4932, 2021. DOI: https://doi.org/10.1002/joc.7106. Disponível em: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.7106. Acesso em 7 jul. 2023.

YANG, S.; YANG, D.; CHEN, J.; SANTISIRISOMBOON, J.; LU, W.; ZHAO, B. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. **Journal of Hydrology**, Amsterdam, v. 590, n. 11, p. 125206, 2020. DOI: https://doi.org/10.1016/j.jhydrol.2020.125206. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169420306661. Acesso em: 7 jul. 2023.

YANG, T.; GAO, X.; SOROOSHIAN, S.; LI, X. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. **Water Resources Research**, United States, v. 52, n. 3, p. 1626-1651, 2016. DOI: https://doi.org/10.1002/2015WR017394. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2015WR017394. Acesso em: 7 jul. 2023.

ZBILUT, J.P.; THOMASSON, N.; WEBBER, C.L. Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals. **Medical engineering & physics**, Netherlands, v. 24, n. 1, p. 53-60, 2002. DOI: https://doi.org/10.1016/S1350-4533(01)00112-6. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1350453301001126. Acesso em: 7 jul. 2023.

ZBILUT, J.P.; WEBBER JR, C.L. Embeddings and delays as derived from quantification of recurrence plots. **Physics letters A**, Netherlands, v. 171, n. 3-4, p. 199-203, 1992. DOI: https://doi.org/10.1016/0375-9601(92)90426-M. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/037596019290426M. Acesso em: 7 jul. 2023.

ZHANG, L.; LI, H.; LIU, D.; FU, Q.; LI, M.; FAIZ, M.A., KHAN, M.I., LI, T. Identification and application of the most suitable entropy model for precipitation complexity measurement. **Atmospheric Research**, Netherlands, v. 221, n. 6, p. 88-97, 2019. DOI: https://doi.org/10.1016/j.atmosres.2019.02.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169809518315886. Acesso em: 7 jul. 2023.

ZHANG, Z.; ZHANG, Q.; SINGH, V.P.; SHI, P. River flow modelling: comparison of performance and evaluation of uncertainty using data-driven models and conceptual hydrological model. **Stochastic environmental research and risk assessment**, New York, v. 32, n. 9, p. 2667-2682, 2018. DOI: https://doi.org/10.1007/s00477-018-1536-y. Disponível em: https://link.springer.com/article/10.1007/s00477-018-1536-y. Acesso em: 7 jul. 2023.

ZHANG, Q.; ZHOU, Y.; SINGH, V.P.; CHEN, X. The influence of dam and lakes on the Yangtze River streamflow: long-range correlation and complexity analyses. **Hydrological Processes**, United Kingdom, v. 26, n. 3, p. 436-444, 2012. DOI: https://doi.org/10.1002/hyp.8148. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.8148. Acesso em: 7 jul. 2023.

ZHAO, X.; CHEN, X.; XU, Y.; XI, D.; ZHANG, Y.; ZHENG, X. An EMD-based chaotic least squares support vector machine hybrid model for annual runoff forecasting. **Water**, Switzerland, v. 9, n. 3, p. 153, 2017. Disponível em: https://www.mdpi.com/2073-4441/9/3/153. Acesso em: 7 jul. 2023.

ZHENG, Y.; HE, Y.; CHEN, X. Spatiotemporal pattern of precipitation concentration and its possible causes in the Pearl River basin, China. **Journal of Cleaner Production**, United Kingdom, v. 161, n. 9, p. 1020-1031, 2017. DOI: https://doi.org/10.1016/j.jclepro.2017.06.156. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652617313264. Acesso em: 7 jul. 2023.

ZHOU, J.; PENG, T.; ZHANG, C.; SUN, N. Data pre-analysis and ensemble of various artificial neural networks for monthly streamflow forecasting. **Water**, Switzerland, v. 10, n. 5, p. 628, 2018. DOI: https://doi.org/10.3390/w10050628. Disponível em: https://www.mdpi.com/2073-4441/10/5/628. Acesso em: 7 jul. 2023.

ZHOU, P.; LI, C.; LI, Z.; CAI, Y. Assessing uncertainty propagation in hybrid models for daily streamflow simulation based on arbitrary polynomial chaos expansion. **Advances in Water Resources**, United Kingdom, v. 160, n. 2, p. 104110, 2022. DOI: https://doi.org/10.1016/j.advwatres.2021.104110. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0309170821002591. Acesso em: 7 jul. 2023.

ZHOU, X.; LEI, W. Spatial patterns of sample entropy based on daily precipitation time series in China and their implications for land surface hydrological interactions. **International Journal of Climatology**, United Kingdom, v. 40, n. 3, p. 1669-1685, 2020. DOI: https://doi.org/10.1002/joc.6294. Disponível em: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.6294. Acesso em 7 jul. 2023

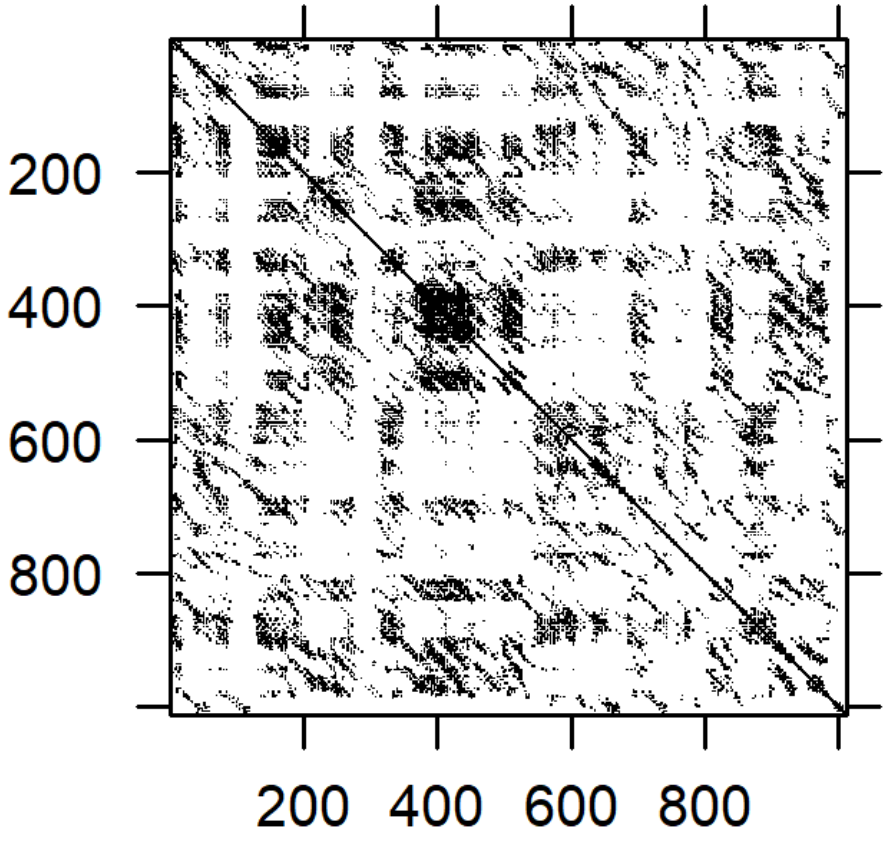
ZUCCHINI, W.; MACDONALD, I.L. **Hidden Markov models for time series: an introduction using R**. United States: Chapman and Hall/CRC, 2016.

APPENDIX A – SUPPLEMENTAL FILES OF CHAPTER 4

This section presents the additional figures mentioned in the main article. The Pacific Decadal Oscillation (PDO) monthly time series from 1931 to 2016 to analyze its relation to rainfall patterns. Through the reconstruction of phase space, which required a five-month delay time and five dimensions, we discovered significant patterns. The recurrence plot (RP) of PDO showed noticeable white stripes around the years 1972 and 1997, matching with a breakpoint observed in the rainfall time series around 1971. The shape of this RP was found to be comparable to patterns identified in previous research (Ogunjo & Fuwape, 2020).

These findings strengthen the association between PDO and regional rainfall patterns, offering a better understanding of climatic variability and its potential impact on weather-dependent activities. Given the marked breakpoints, future research should explore these periods more closely, seeking potential triggers and consequences.

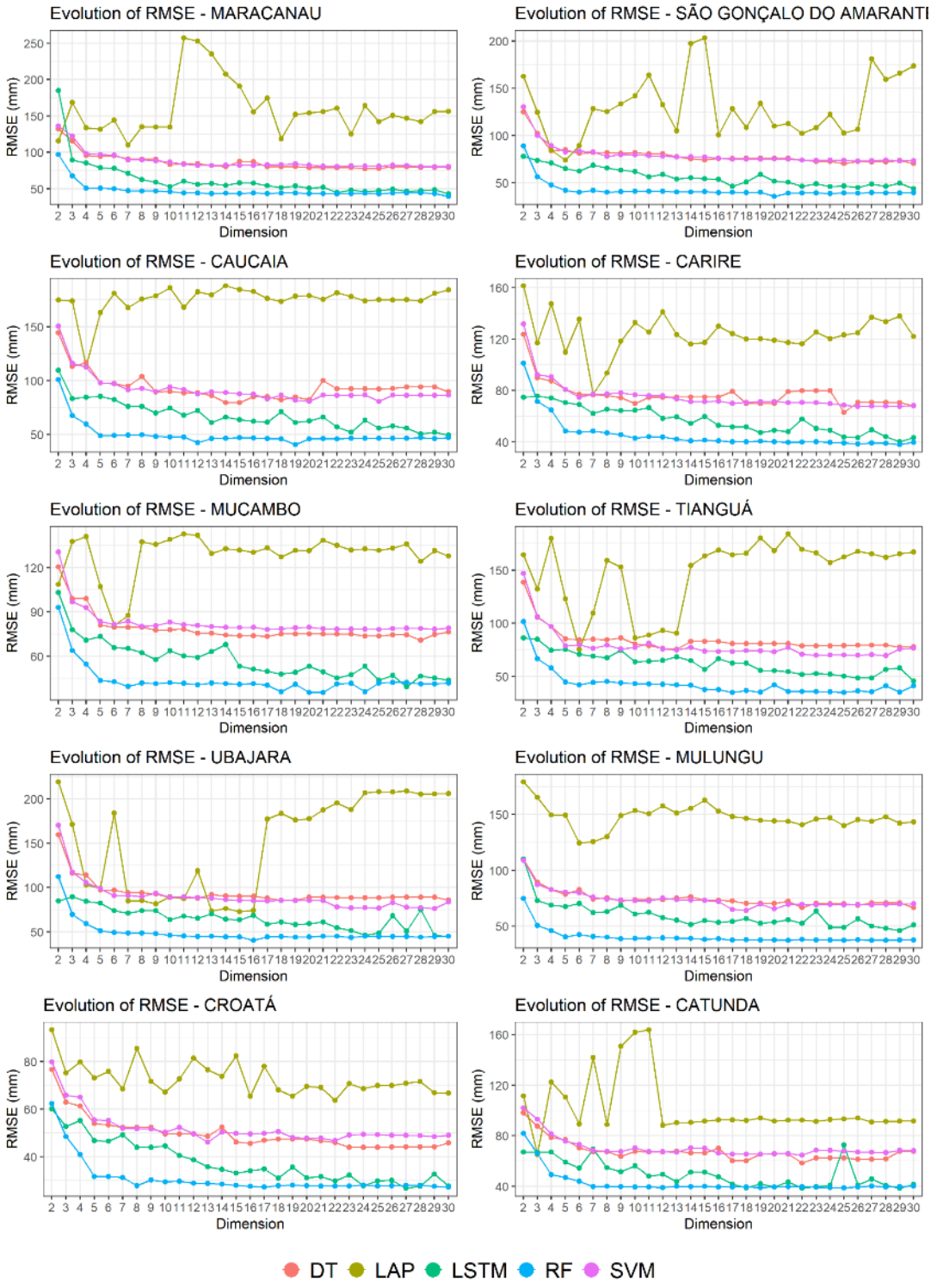
Figure A1 – PDO recurrence plot

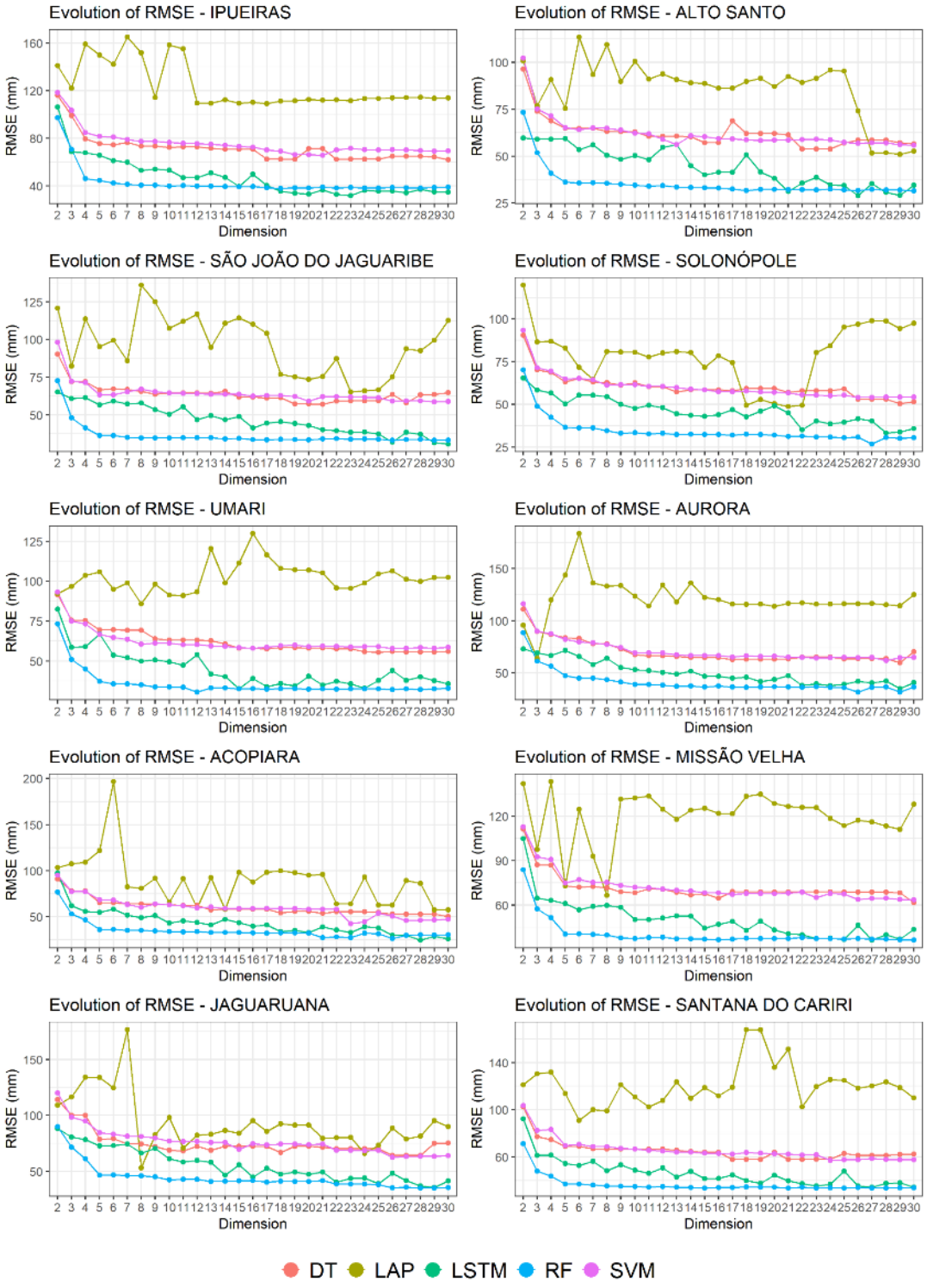


APPENDIX B – SUPPLEMENTAL FILES OF CHAPTER 6

The interplay between delay time and embedding dimensions influences the precision of predictions. The selection of dimensions for each station was determined based on the stabilization of model errors in most machine learning models, guided by the RMSE values (B1). The RF and SVM methodologies showed marginal prediction accuracy enhancements with increasing dimensions, as evidenced by the RMSE. Other methods, such as LAP, exhibited substantial RMSE variation with embedding dimension changes.

Figure B1 – Evolution of RMSE values across model dimensions ranging from 2 to 30 for the 20 rainfall stations.




Evolution of RMSE - IPUEIRAS
Evolution of RMSE - ALTO SANTO
Evolution of RMSE - SÃO JOÃO DO JAGUARIBE
Evolution of RMSE - SOLONÓPOLE
Evolution of RMSE - UMARI
Evolution of RMSE - AURORA
Evolution of RMSE - ACOPIARA
Evolution of RMSE - MISSÃO VELHA
Evolution of RMSE - JAGUARUANA
Evolution of RMSE - SANTANA DO CARIRI
RMSE (mm)
Dimension
DT
LAP
LSTM
RF
SVM