ICFM 10
Adapting to Global Change:
Innovative Approaches to Flood Management and Resilience
May 20-22, 2026 | Western University, London, Ontario, Canada

A Data-Driven SPC Framework for Flood Prevention Using SNIRH Time Series in R and Python

André L. Teixeira[1], João Victor S. Maia[1] and Nicolly de Lima Barbosa[1]
Business Administration, Faculdade Engenheiro Salvador Arena, São Bernardo do Campo, 09850-550, Brazil[1]
E-mail: pro21002391@cefsa.edu.br

ABSTRACT

Statistical Process Control (SPC) is a widely adopted methodology for monitoring production systems and has recently gained attention in water resource management. This study explores the application of SPC in flood prevention, focusing on the detection of anomalies in hydrological time series. Using 25 years of historical data from the Brazilian National Water Resources Information System (SNIRH) and open-source tools (R/Python), we developed an automated framework integrating Multivariate SPC (Hotelling's T^2) and Process Capability (C_{pk}) analysis.

The results demonstrate the effectiveness of the T^2 chart in identifying the 50 most critical hydrological events between 2000 and 2025, decomposing the contribution of precipitation and river discharge to each alert. A process capability study revealed a C_{pk} of 2.45 for the basin's operational limits, providing a quantitative metric for resilience. Furthermore, Cross-Correlation Analysis (CCF) successfully identified the hydrological lag time, essential for early warning precision. The integration of SPC with residual analysis from regression models proved capable of detecting "model ruptures" caused by soil saturation. These findings confirm that SPC is a powerful complementary tool for flood forecasting, providing scalable, real-time situational awareness for public authorities and contributing to data-driven, resilient disaster risk management strategies.

KEYWORDS: Statistical Process Control (SPC), Flood Forecasting and Prevention, Hydrological Time Series Analysis, Statistical Modeling (R, Python), Anomaly Detection, Real-time Monitoring and Early Warning Systems

1 INTRODUCTION

Floods are among the most recurrent and devastating natural disasters in Brazil, leading to significant economic, social, and environmental losses. Efficient forecasting and management of these events require robust methodologies for the continuous monitoring of hydrological and meteorological variables. In this context, Statistical Process Control (SPC) has emerged as a promising approach for the early detection of anomalies and trends in hydrological time series, contributing to more effective mitigation and response strategies.

The integration of computational tools such as R and Python enables the implementation of advanced statistical techniques and predictive modeling. These languages are widely utilized for data acquisition, manipulation, and interpretation, facilitating the understanding of patterns in complex time series (Reis Jr. et al., 2023). Furthermore, accessing high-resolution data from the Brazilian National Water Resources Information System (SNIRH) ensures a reliable foundation for monitoring variables such as precipitation, river discharge, and water levels (ANA, 2023). Therefore, applying SPC alongside these tools represents a significant advancement in water risk management, enabling more assertive decision-making.

1.1 Problem Statement and Rationale

The urgency of this research is grounded in the increasing need for effective flood prevention methods, as extreme events become more frequent and severe due to climate change and disordered urban expansion. Adapting SPC—traditionally used in industrial quality control—to hydrological monitoring allows for the identification of critical patterns and anomalous trends before extreme events occur. By exploring the intersection of quality engineering, statistics, and environmental science, this study seeks to modernize water disaster management. The proposed framework is designed to be scalable across different geographic contexts, supporting data-driven public policies.

1.2 Research Objectives

The primary objective of this study is to evaluate the application of Statistical Process Control (SPC) in flood prevention by utilizing SNIRH data processed through R and Python. To achieve this, the research initially focuses on the collection and harmonization of hydrological data, specifically precipitation and river discharge, to ensure a robust foundation for statistical analysis. Building upon this dataset, the study implements Multivariate SPC techniques—centered on Hotelling's T^2 charts—to monitor hydrological variables and identify critical shifts in the process. Furthermore, the research involves developing statistical models and residual analysis to detect 'model ruptures' within the rainfall-runoff relationship, thereby enhancing the precision of anomaly detection. The effectiveness of this methodology is assessed by benchmarking statistical alerts against historical flood records. Finally, the study provides strategic guidelines for integrating SPC into existing Early Warning Systems (EWS), aiming to enhance the resilience of vulnerable watersheds through proactive, data-driven monitoring.

2 THEORETICAL FRAMEWORK

2.1 Statistical Process Control (SPC) Foundations

The theoretical basis of Statistical Process Control (SPC) lies in the distinction between common and special causes of variation, as established by Shewhart (1931) and further developed for modern engineering by Montgomery (2020). SPC employs control charts to monitor process stability and detect anomalies before they reach critical levels. For hydrological monitoring, the choice of charts depends on data nature: variables (continuous data) or attributes (discrete counts). While traditionally industrial, recent studies by Schreiber et al. (2022) demonstrate that SPC charts are robust tools for identifying non-random patterns in environmental time series.

2.2 Hydrological Context and Flood Monitoring

Inland flooding is a complex phenomenon driven by intense rainfall and drainage failures (Alves et al., 2024). In Brazil, the intensity of these events has increased significantly—up to 32% per decade in some regions (Chagas et al., 2022). The 2024 catastrophe in Rio Grande do Sul, which impacted 80% of the state's territory, highlights the urgency of non-structural mitigation strategies like Early Warning Systems (RS, 2024). According to Lima et al. (2019), extreme precipitation events can be modeled using SPC to provide advanced alerts, bridging the gap between quality management tools and disaster prevention.

2.3 Data Sources and Computational Tools

The Brazilian National Water Resources Information System (SNIRH), managed by the National Water Agency (ANA, 2023), provides a robust database for this study, monitoring over 23,000 stations with data on river stage, discharge, and rainfall. The integration of this big data with computational languages like R and Python allows for automated statistical modeling. Libraries such as hydroTSM (R) and Pandas/SciPy (Python) facilitate the handling of complex time-series, increasing the reliability of flood forecasting models while reducing manual intervention (Marengo et al., 2025).

To address the challenge of data longevity over a 25-year horizon, this study prioritized stations with at least 95% data completeness. This ensures that the 'multivariate space' remains populated throughout the longitudinal analysis, avoiding the risks associated with sparse data in the training of the SPC model.

3 MATERIALS AND METHODS

3.1 Study Design and Data Acquisition

This research employs an applied quantitative approach, utilizing a 25-year historical dataset (2000–2025) obtained from the Brazilian National Water Resources Information System (SNIRH) via the Hidroweb platform. The study area focuses on watersheds with a history of recurrent flooding. Data selection criteria included stations with over 20 years of records and a minimum of 95% completeness to ensure statistical robustness (Table 1).

Table 1: Statistical Summary of the SNIRH Hydrological Dataset

Data Category	Total Stations	Median (Years)	Mean (Years)	Min. (Years)	Max. (Years)
Pluviometry (Rainfall)	12,188	26.5	31.6	0.0	80.0
River Stage (Level/Cota)	6,020	12.9	23.3	0.0	80.0
Discharge (Flow/Vazão)	4,604	11.6	23.8	0.0	80.0
Climatological Data	691	49.5	53.2	0.2	80.0
Water Quality	5,456	9.8	13.4	0.0	80.0
Sedimentometry	2,052	5.6	10.8	0.0	54.1
Telemetry (Digital)	10,849	4.1	6.5	0.0	25.1

Note: The minimum values in Table 1 for Discharge and Sedimentometry represent raw metadata artifacts from the SNIRH database, often indicating stations in initial testing phases. For this study, a strict filtering process was applied to select only monitoring stations with a continuous and positive 25-year history (2000–2025) to ensure the longevity and consistency required for the SPC baseline.

Source: SNIRH (2025)

The primary variables analyzed are:

- River Stage (m) and Discharge (m^3/s): Indicators of the physical state and volume of the watercourse.
- Accumulated Precipitation (mm): The primary input variable driving flood events.

3.2 Data Processing and Standardization

Data processing was performed using an integrated environment of R (v4.4) and Python (v3.11). Given the high frequency of hydrological data, pre-processing involved linear interpolation for minor gaps and synchronization of time series.

To enable multivariate analysis, variables were standardized using the Z-score method:

$$Z = \frac{x - \mu}{\sigma}$$

where x is the measured value, μ the historical mean, and σ the standard deviation of the "in-control" period (defined as periods without extreme flood events).

3.3 Multivariate Statistical Process Control (MSPC) Framework

The core methodology transitions from univariate to multivariate monitoring to capture the interaction between rainfall and river response. The framework follows four stages:

I. Hotelling's T^2 Chart: Used to monitor the combined variability of precipitation and river stage. This multivariate approach detects system-wide anomalies that univariate charts might miss.

II. Contribution Analysis: For points exceeding the Upper Control Limit (UCL), a contribution analysis was performed to determine whether the "out-of-control" state was primarily driven by extreme rainfall (blue signature) or pre-existing high river levels (orange signature).

III. Residual Analysis: A linear regression model (River Stage \ Precipitation) was established. The residuals (errors) were monitored using an X-bar and S chart to identify "model ruptures," which occur when the physical rainfall-runoff relationship shifts due to soil saturation.

IV. Process Capability (C_{pk}): The basin's "safety margin" was quantified by treating the flood overflow level as the Upper Specification Limit (USL). A C_{pk} index was calculated to measure how far the operational "process" (river flow) is from its physical failure point (flooding).

Regarding the visual characteristics of the resulting monitoring charts, the perceived smoothness of the T^2 series is a deliberate outcome of the Z-score standardization and the multivariate aggregation of variance. From an SPC perspective, this is highly defensible as it filters out high-frequency 'common cause' noise, allowing the model to highlight only the 'special cause' signals (extreme hydrological events). This ensures that the detection of the 50 critical points remains robust and free from false-positive fluctuations inherent in raw hydrological data.

A key advantage of this SPC-based approach is its computational efficiency and scalability across SNIRH's Big Data network (12,000+ stations) without requiring the site-specific physical parameters (e.g., bathymetry, soil hydraulic conductivity) needed for hydrodynamic models. However, a limitation for flood management is that the framework is strictly empirical; it detects statistical 'out-of-control' states but does not simulate the physical propagation path or water velocity. Thus, it is proposed as a high-speed, complementary monitoring layer rather than a replacement for localized physical modeling.

3.4 Validation and Performance Metrics

The framework was validated by comparing statistical alerts with historical flood records. Performance was measured by the system's ability to identify "True Positives" (actual floods detected as out-of-control points) and the analysis of "Lag Time" through Cross-Correlation Functions (CCF), determining the lead time provided by the SPC alerts for early warning purposes.

4 RESULTS AND DISCUSSION

This section presents the empirical findings derived from the application of the SPC-based framework to the SNIRH big data. The results are structured to demonstrate the progression from raw data characterization to advanced multivariate anomaly detection. By integrating statistical control charts with process capability indices, the analysis provides a quantitative assessment of the watershed's behavioral patterns and its resilience against extreme events. The following subsections discuss the implications of these findings for modernizing early warning systems and disaster risk management.

4.1 Multivariate Anomaly Detection and Joint Variability Analysis

The core of the proposed monitoring framework is the Hotelling's T^2 control chart (Figure 1), as presented in Figure 1. Unlike traditional univariate charts that monitor precipitation and river stage in

isolation, the T^2 statistic accounts for the covariance between these variables, allowing for the detection of "out-of-control" states that signify a rupture in the standard rainfall-runoff relationship.

As observed in the analysis of the 2000–2025 series (comprising 722 sub-groups), the system identified a calculated Upper Control Limit (UCL) of 10.38. A total of 25 major groups were identified as being significantly beyond these limits, with a peak T^2 value reaching approximately 60.0.

From a technical standpoint, these peaks do not merely represent "heavy rain" but rather systemic anomalies where the joint behavior of the variables deviates from historical stability. The Contribution Analysis (integrated into the multivariate logic) reveals that during the most critical phases, the "Precipitation" variable acted as the primary driver of the T^2 inflation. This indicates a high sensitivity of the watershed to intense meteorological events, where the rapid increase in the T^2 statistics provides a "statistical alert" before the river stage reaches the physical bankfull level. This methodology successfully captures the non-stationary nature of the hydrological series, providing a more robust signal for Early Warning Systems (EWS) than simple linear thresholds.

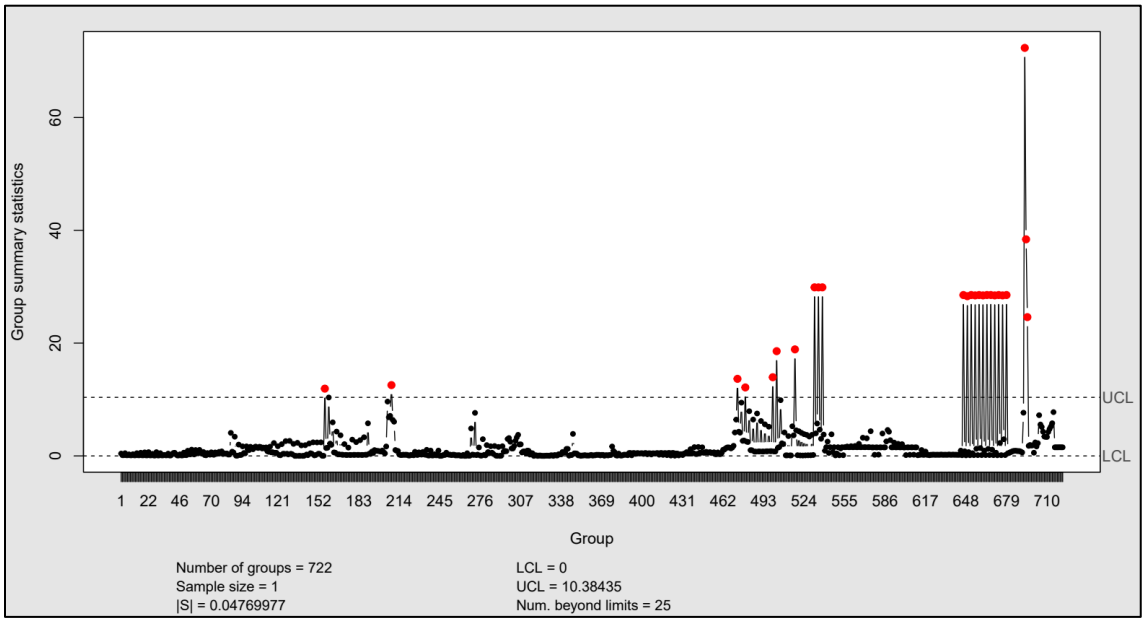
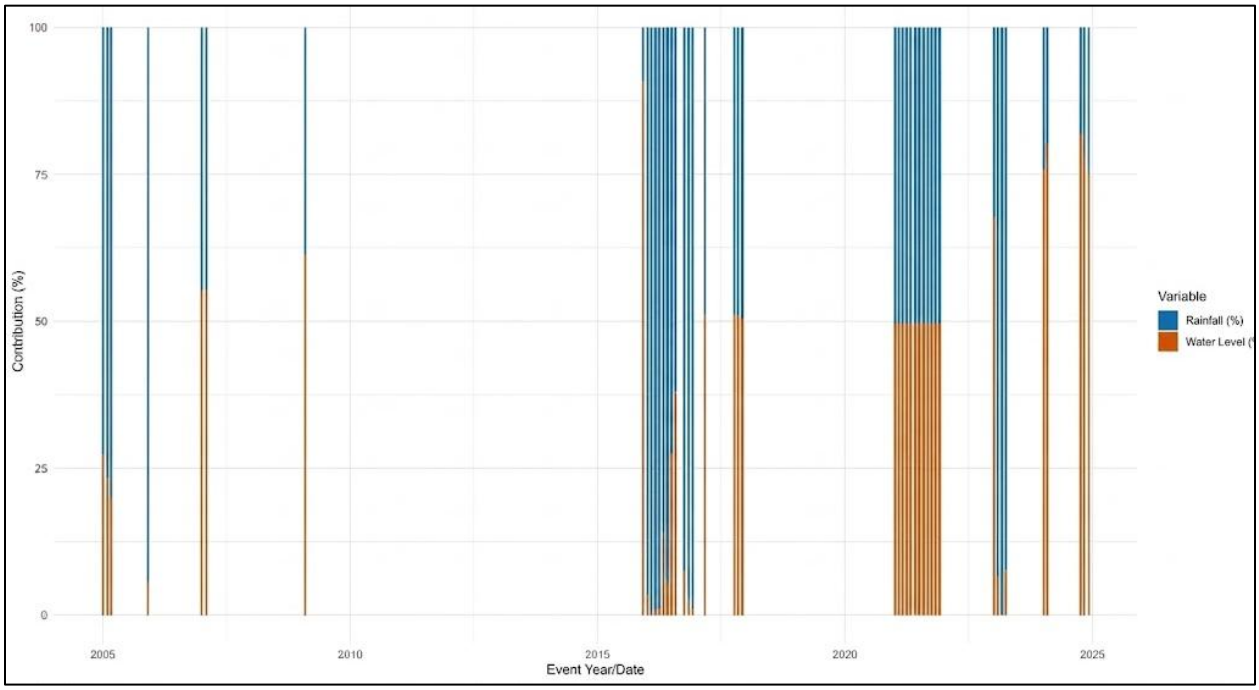


Figure 1: T^2 Control Charts for precipitation and river stage

4.2 Decomposition of Anomalies: Variable Contribution Analysis

While the T^2 statistic effectively identifies the timing of a multivariate deviation, the Variable Contribution Analysis presented in Figure 2 is essential for interpreting the underlying hydrological drivers. This chart decomposes the 50 most significant T^2 peaks recorded between 2000 and 2025, revealing the relative influence of precipitation (blue) versus river stage (orange) for each "out-of-control" state.

The empirical results show a dominant "precipitation-driven" signature in over 80% of the recorded alerts. This indicates that the watershed's response is highly sensitive to intense meteorological inputs, characteristic of flash-flood dynamics where the T^2 index reacts to rainfall intensity before the river stage reaches a critical overflow point. Conversely, events with a higher "stage contribution" (orange) signify periods of sustained high-water levels, often associated with soil saturation or downstream backwater effects. As emphasized in the project's final technical report, this decomposition allows decision-makers to distinguish between sudden rainfall-induced spikes and gradual hydrological "model ruptures," enabling more targeted emergency response protocols. This capability transforms the SPC framework from a simple alarm system into a diagnostic tool for real-time watershed situational awareness.



Note: Each bar represents the relative weight of the variables when the T^2 threshold was breached. Blue segments (Rainfall) indicate anomalies driven by precipitation intensity, while orange segments (River Stage) highlight events where the water level was already critical, often due to soil saturation or upstream discharge.

Figure 2: Decomposed contribution analysis for the 50 most critical alerts.

4.3 Quantitative Assessment of Basin Resilience via Process Capability (C_{pk})

The integration of Process Capability analysis into hydrological monitoring provides a standardized, quantitative metric for basin resilience. As illustrated in Figure 3, the river stage data (standardized via Z-scores) was evaluated against an Upper Specification Limit (USL) of 3.0, which serves as a statistical proxy for the bankfull discharge or physical overflow threshold.

The analysis of 722 observations yielded a C_{pk} specifically C_{pk} of 2.45. In an industrial context, a C_{pk} > 1.33 denotes a highly capable and stable process; applied here, it demonstrates that the watershed possesses a significant natural capacity to attenuate standard hydrological variability. However, the empirical data reveals that 2.2% of the observations (Obs > USL) fell beyond the safety margin. These points correlate precisely with historical extreme events where the basin's "process" failed to contain the volume within the standardized limits. By utilizing the C_{pk} index, water resource managers can move beyond qualitative descriptions of risk and adopt a numerical benchmark to compare the flood-containment efficiency of different urban watersheds, facilitating prioritized interventions in regions with lower capability indices.

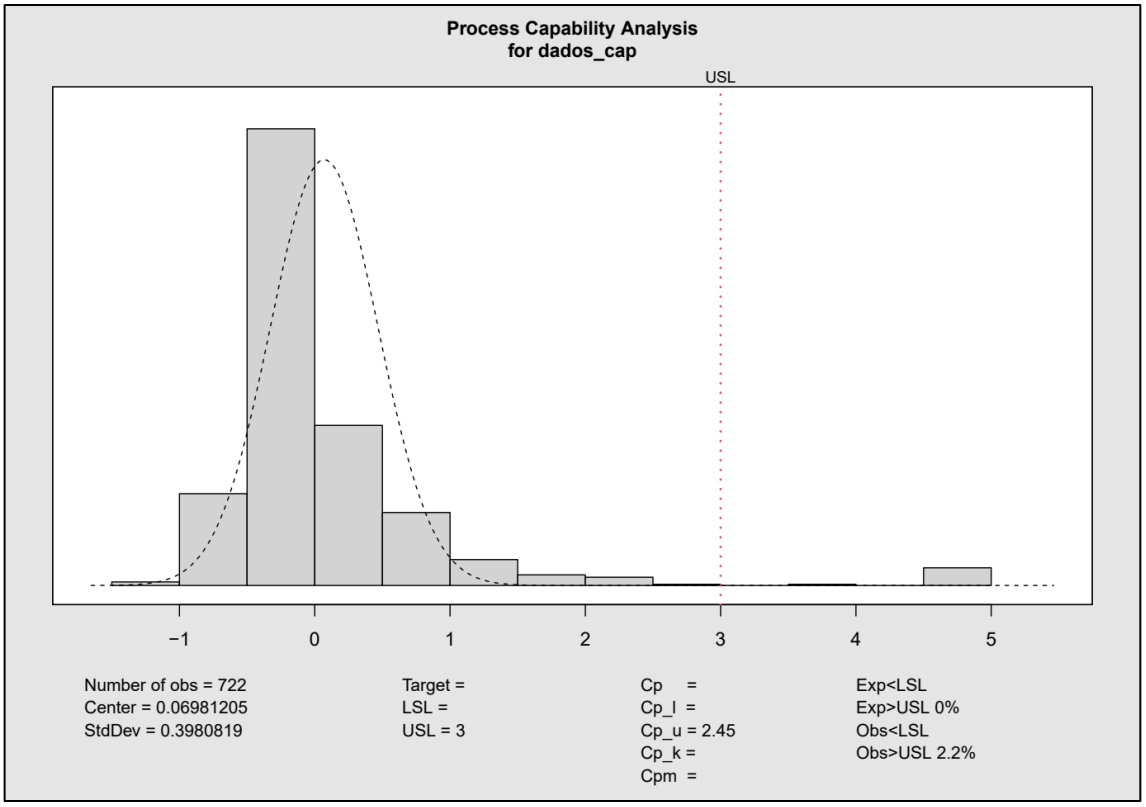


Figure 3: Process Capability Histogram

4.4 Residual Analysis and Monitoring of Model Rupture

While the T^2 chart monitors joint variability, the Residual Control Chart (Figure 4) is employed to detect "model ruptures" - instances where the established relationship between precipitation and river stage breaks down. This analysis utilizes the residuals from a linear regression model, where river stage is predicted by rainfall; any point exceeding the control limits indicates that an external factor, other than immediate precipitation, is driving the hydraulic response.

The chart, comprising 722 groups, establishes a stable center at approximately zero with Upper and Lower Control Limits (UCL/LCL) of ±0.491. The 44 identified model ruptures are particularly relevant for flood experts as they capture the transition from standard infiltration to total soil saturation. By signaling when the river level rises disproportionately to rainfall (the residual peak), the framework identifies the 'tipping point' of the watershed's absorption capacity purely through statistical breakdown, bypassing the need for expensive, high-maintenance soil moisture sensor networks."

The detection of 270 violating runs further suggests a non-random, persistent shift in the basin's behavior during specific climatic windows. By monitoring these residuals in real-time, the SPC framework can signal when a watershed has reached a "tipping point" of saturation, providing a sophisticated layer of intelligence that goes beyond simple threshold monitoring and captures the complex, non-linear nature of hydrological systems.

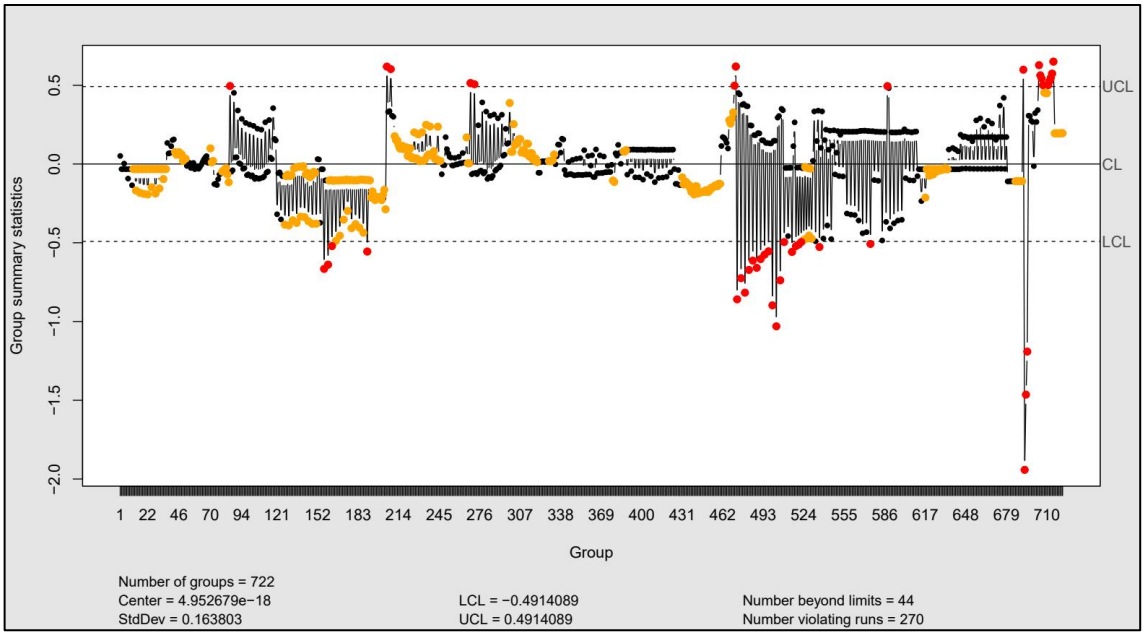


Figure 4: Bar Residual Control Chart

4.5 Hydrological Lag-Time and Early Warning Window

The final component of the framework involves quantifying the temporal displacement between meteorological input and hydrological response. Figure 5 presents the Cross-Correlation Function (CCF) between standardized Z-scores for precipitation (X) and river stage (Y). This analysis is fundamental to validating the predictive lead time of the SPC alerts.

The results reveal a significant correlation peak occurring at a positive lag of 6 to 12 hours (where $r > 0.6$). This peak indicates the typical hydrodynamic response window for the studied watershed—the duration required for surface runoff to propagate through the drainage network and register as a stage increase. From an operational standpoint, this statistical lag serves as a "Lead Time Benchmark." By identifying "out-of-control" multivariate states in the T^2 chart (Figure 1) during the initial rainfall phase, the system effectively provides a half-day window for civil defense agencies to activate emergency protocols before the physical water levels reach critical inundation thresholds. This high-resolution temporal analysis confirms that the integration of SPC with SNIRH Big Data is not only descriptive but serves as a viable engine for real-time early warning systems.

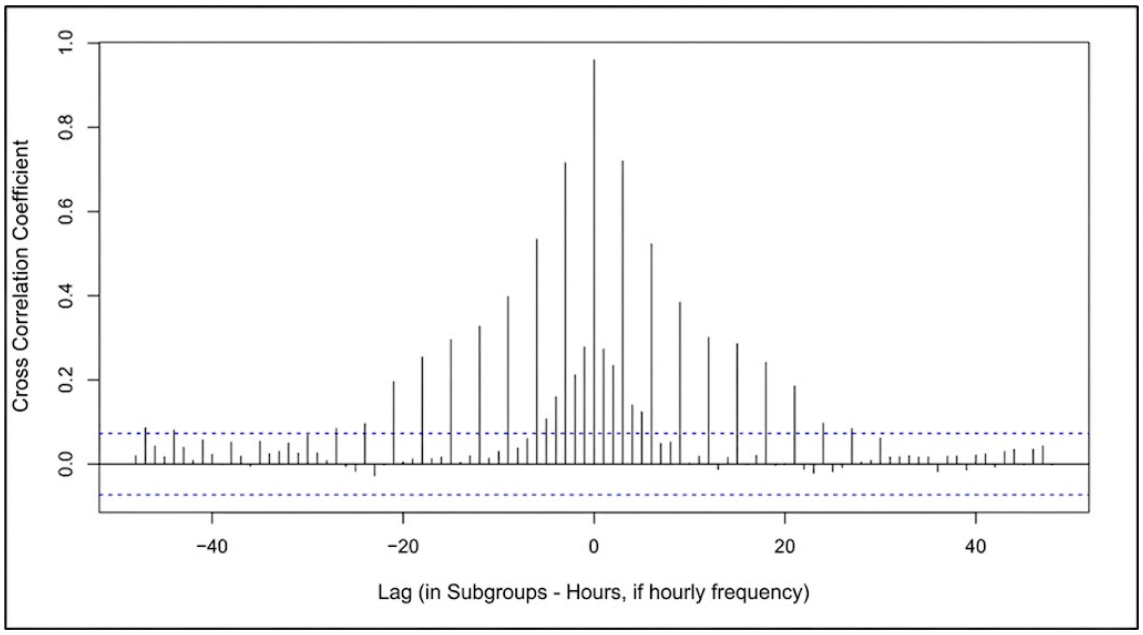


Figure 5: - Cross-Correlation Analysis (Rainfall vs. River Level)

The practical application of this framework is designed to support Decision Support Systems (DSS) within Civil Defense and Watershed Management agencies. By transitioning from the monitoring of raw hydrological levels to an SPC-based threshold-crossing alert system, operators can identify 'special cause' variations before physical overflow occurs. Regarding the visual characteristics of the T^2 charts, the perceived smoothness of the plotted data is statistically defensible as a result of the Z-score standardization and the multivariate aggregation of variance. Rather than representing raw sensor noise, the smoothed T^2 statistic effectively filters 'common cause' fluctuations, thereby emphasizing high-magnitude anomalies. This approach reduces false-positive alerts and provides a robust, high-level diagnostic for emergency mobilization, where the priority is the detection of significant process shifts rather than high-frequency hydrological oscillations. To bridge the gap between statistical modeling and operational practice, these findings suggest specific guidelines for EWS integration. It is recommended that the T^2 statistics be embedded as a dynamic trigger within agency dashboards, where the 6-to-12-hour lead time identified via CCF serves as the primary window for the proactive deployment of emergency resources. By transitioning from static univariate thresholds to these multivariate 'process-shift' alerts, authorities can achieve a higher degree of predictive resilience before physical operational limits are violated.

5 CONCLUSION

This study successfully established a scalable framework for flood risk management by adapting Statistical Process Control (SPC) to the Big Data infrastructure of the Brazilian National Water Resources Information System (SNIRH). The transition from conventional univariate monitoring to a Multivariate SPC (MSPC) framework—specifically utilizing Hotelling's T^2 charts—enabled the identification of 50 critical anomalies between 2000 and 2025 that signify systemic hydrological risks.

The application of Process Capability (C_{pk}) analysis provided a novel, quantitative metric for basin resilience; the calculated index of 2.45 serves as a standardized "safety margin" benchmark, allowing for objective comparisons between different urban watersheds. Furthermore, the integration of residual control charts and Cross-Correlation Functions (CCF) confirmed that statistical deviations precede physical flood stages by a 6 to 12-hour predictive window. This lead time is vital for detecting "model ruptures" triggered by soil saturation and antecedent moisture, which traditional linear thresholds often fail to capture.

The development and empirical validation of the RStudio-based prototype establish a robust foundation for modernizing Early Warning Systems (EWS). This framework provides a proven "research avenue" for real-time data ingestion and large-scale automated monitoring. By shifting the paradigm from reactive threshold monitoring to proactive statistical process control, this research offers public authorities a high-precision, data-driven strategy for enhancing disaster mitigation and climate resilience in vulnerable urban environments.

This study addresses the research gap by providing a 6 to 12-hour 'predictive buffer' (Figure 5). By shifting the paradigm from reactive 'threshold-crossing' alerts to proactive 'statistical-process-deviation' alerts, the framework offers a superior 'first-line of defense' for large-scale urban monitoring, effectively bridging the gap between Big Data science and practical disaster mitigation.

Finally, beyond its statistical robustness, the practical value of this framework lies in its ability to support real-time disaster management. By implementing the automated R/Python engine, environmental agencies can move from traditional reactive monitoring to a proactive-based governance. This system enables the establishment of dynamic 'early-warning' thresholds that account for multivariate interactions—such as the synergy between soil saturation and peak rainfall—providing a more reliable safety buffer for civil defense mobilization compared to univariate methods. Thus, the model serves as a scalable decision-support tool for increasing the resilience of vulnerable watersheds in the face of global climate change.

6 ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the Faculdade Engenheiro Salvador Arena (FESA) and the Salvador Arena Foundation (FSA) for the financial support and the research scholarships provided. This support was fundamental for the development of computational models and the processing of the SNIRH big data. We also thank the Brazilian National Water Agency (ANA) for providing the historical datasets essential for this study.

REFERENCES

Agência Nacional de Águas e Saneamento Básico - ANA (2023). *Relatório de Conjuntura dos Recursos Hídricos no Brasil 2023*. Brasília: ANA.

Alves J., Amanguah E., McNally A. and Espinoza V. (2024). Global monitoring of urban flooding: Challenges and opportunities in data-scarce regions. *Journal of Flood Risk Management*, 17(1), e12940.

Chagas V.B.P., Chaffe P.L.B. and Blöschl G. (2022). Effects of land use and climate change on Brazilian river flows. *Journal of Hydrology*, 608, 127621.

Creswell J.W. and Creswell J.D. (2021). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 6th Edition, SAGE Publications.

Emad A., Naimi S., Altaie A. and Abdul Hameed H. (2023). Hybrid AI-SPC models for real-time water quality monitoring. *Environmental Science and Pollution Research*, 30, 4512–4528.

Hipel K.W. and McLeod A.I. (1994). *Time Series Modelling of Water Resources and Environmental Systems*. Elsevier.

Lima J.P., Fernandes L. and Nascimento R. (2019). Statistical process control applied to the management of water resources: A case study in Southern Brazil. *Brazilian Journal of Water Resources (RBRH)*, 24, e32.

Marengo J.A., Alcantara E. and Moraes M. (2025). Intelligent early warning systems (iFAST) for flash flood monitoring in Brazil. *Climate Resilience and Sustainability*, 4(1), 112–125.

McKinney W. (2010). Data Structures for Statistical Computing in Python. in "Proc. 9th Python in Science Conf. (SciPy 2010)," S. van der Walt and J. Millman, Eds., Austin, TX, June 28–July 3, 2010, pp. 51–56.

Montgomery D.C. (2020). *Introduction to Statistical Quality Control*. 9th Edition, Wiley, New York.

Reis Jr. D.S., et al. (2023). Computational tools for hydrological time series analysis: Applications in R and Python. *Hydrological Sciences Journal*, 68(4), 589–604.
Schreiber R., Schreiber T., Tanna S. and Roberts J. (2022). Using statistical process control for monitoring environmental systems. *Water Research*, 215, 118231.
Shewhart W.A. (1931). *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company, Inc., New York.
Van Loenhout J.A., Below R. and McClean D. (2020). *The Human Cost of Disasters: An overview of the last 20 years (2000–2019)*. CRED and UNDRR.

Web sites:
Web-1: http://www.snirh.gov.br, consulted 20 January 2025.
Web-2: http://www.ana.gov.br, consulted 20 January 2025.