UNIVERSIDADE FEDERAL DE LAVRAS

ALEJANDRA BUSTILLOS VEGA

DATA-DRIVEN HYDROLOGY: : RAINFALL MODELING IN FOREST ECOSYSTEMS USING MACHINE LEARNING

LAVRAS – MG

2026

ALEJANDRA BUSTILLOS VEGA

DATA-DRIVEN HYDROLOGY: : RAINFALL MODELING IN FOREST ECOSYSTEMS USING MACHINE LEARNING

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Recursos Hídricos, área de concentração em Hidrologia, para a obtenção do título de Doutora.

Prof. DSc. Carlos Rogério de Mello
Orientador

Prof. DSc. Gernot Heisenberg
Coorientador

LAVRAS – MG
2026

Ficha Catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Vega, Alejandra Bustillos Vega.
Data-Driven Hydrology: Rainfall Modeling In Forest Ecosystems Using Machine Learning / Alejandra Bustillos Vega Vega. - 2026.
226 p. : il.

Orientador: Carlos Rogério de Mello Mello
Coorientador: Gernot Heisenberg Heisenberg

Tese (Doutorado) - Universidade Federal de Lavras, 2026.
Bibliografia.

1. Hydrology. 2. Machine Learning. 3. Rainfall modeling. I. Mello, Carlos Rogério de Mello. II. Heisenberg, Gernot Heisenberg. III. Universidade Federal de Lavras. IV. Título.

ALEJANDRA BUSTILLOS VEGA

DATA-DRIVEN HYDROLOGY: : RAINFALL MODELING IN FOREST ECOSYSTEMS USING MACHINE LEARNING

HIDROLOGIA ORIENTADA POR DADOS: : MODELAGEM DE PRECIPITAÇÃO EM ECOSSISTEMAS FLORESTAIS USANDO APRENDIZADO DE MÁQUINA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Recursos Hídricos, área de concentração em Hidrologia, para a obtenção do título de Doutora.

APROVADA em 06 de março de 2026.

Prof. DSc. Gilberto Coelho	UFLA
Prof. DSc. André Ferreira Rodrigues	UFMG
Prof. DSc. Samuel Beskow	UFPel
Prof. DSc. Gernot Heisenberg	TH Köln

gov.br
Documento assinado digitalmente
CARLOS ROGERIO DE MELLO
Data: 26/06/2026 10:26:04-0300
Verifique em https://validar.iti.gov.br

Prof. DSc. Carlos Rogério de Mello
Orientador

Prof. DSc. Gernot Heisenberg
Co-Orientador

LAVRAS – MG
2026

À minha família — raiz que sustenta, mesmo quando a distância parece intransponível. Cada página desta tese carrega, de alguma forma, a sua presença..

AGRADECIMENTOS

Nenhuma tese nasce sozinha. Este trabalho é, antes de tudo, o resultado de encontros.

Ao Prof. Dr. Carlos Rogério de Mello, orientador desta pesquisa, pela confiança depositada desde o primeiro momento, pela rigorosidade científica que eleva o trabalho e pela generosidade que torna o caminho possível. Mais do que orientação técnica, agradeço a presença constante e a disposição para construir junto.

Ao Prof. Dr. Gernot Heisenberg, coorientador, pela abertura intelectual, pela perspectiva internacional que enriqueceu esta pesquisa e pela disponibilidade que transcendeu fronteiras geográficas.

À banca de qualificação, pelas críticas que, ao invés de diminuir o trabalho, o tornaram mais robusto. O rigor acadêmico, quando exercido com seriedade, é também uma forma de cuidado.

Ao Programa de Pós-Graduação em Recursos Hídricos da Universidade Federal de Lavras, pela estrutura e pelo ambiente acadêmico que tornaram este doutorado possível. Ao experimento da “Matinha” — e a todos que ao longo de mais de uma década saíram a campo, sob chuva ou sob sol, para garantir que os dados existissem. Sem esse trabalho silencioso, esta tese não teria substância. À minha família, por compreender as ausências, por celebrar cada avanço e por lembrar, nos momentos de dúvida, por que vale a pena continuar.

Este trabalho foi financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, processo n.º 402066/2023-5), cujos recursos viabilizaram a coleta e disponibilização dos dados experimentais da Matinha que sustentam esta tese. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Código de Financiamento 001), pelo apoio ao Programa de Pós-Graduação em Recursos Hídricos da UFLA e pela concessão da bolsa de doutorado sanduíche que tornou possível a colaboração internacional. À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela bolsa concedida que possibilitou a dedicação integral a esta pesquisa.

A todos que, de alguma forma, deixaram marca neste percurso — obrigada

A natureza não faz perguntas — ela simplesmente registra, ciclo após ciclo, a sua própria lógica. Cabe à ciência aprender a escutá-la com a humildade de quem sabe que os dados contam mais do que as certezas.

(Autor Desconhecido)

RESUMO

Este estudo desenvolve um Marco Metodológico Integrado baseado em Aprendizado de Máquina (Machine Learning – ML) para a reconstrução, modelagem e transformação da precipitação em ecossistemas florestais sazonais, abordando conjuntamente disponibilidade de dados, variabilidade hidrológica e os processos ecohidrológicos fundamentais que controlam a entrada de água no solo. Primeiramente, propõe-se uma metodologia robusta para reconstrução de séries horárias e diárias contínuas em regiões com redes de monitoramento incompletas. A abordagem combina controle automatizado de qualidade por meio de detecção não supervisionada de anomalias, correção de viés de produtos satelitais e de reanálise, e identificação ótima de modelos com base em critérios multiobjetivo que equilibram acurácia estatística e consistência hidrológica.

Os resultados demonstram que modelos de ML treinados com dados submetidos a controle de qualidade superam significativamente métodos estatísticos tradicionais, preservando a variabilidade convectiva essencial para aplicações hidrológicas em ecossistemas tropicais. O estudo avança na modelagem da precipitação ao incorporar explicitamente processos de interceptação pelo dossel florestal. Utilizando ensembles de ML e um conjunto de preditores hidrologicamente interpretáveis, observa-se forte variabilidade espaço-temporal na interceptação, inadequadamente representada por coeficientes constantes.

De forma geral, o estudo demonstra que a reconstrução da precipitação e sua transformação em precipitação efetiva são processos inseparáveis na ecohidrologia de florestas sazonais, contribuindo com ferramentas reproduzíveis para análise do ciclo hidrológico e gestão de recursos hídricos.

Palavras-chave: hidrologia orientada por dados; aprendizado de máquina; reconstrução de precipitação; interceptação florestal; ecohidrologia; florestas sazonais tropicais.

ABSTRACT

This study develops an Integrated Methodological Framework based on Machine Learning (ML) for the reconstruction, modeling, and transformation of rainfall in seasonal forest ecosystems, jointly addressing data availability, hydrological variability, and the fundamental ecohydrological processes controlling water input to the soil. First, a robust methodology is proposed for reconstructing continuous hourly and daily rainfall series in regions with incomplete monitoring networks. The approach combines automated quality control through unsupervised anomaly detection, bias correction of satellite and reanalysis products, and optimal model identification based on multi-objective criteria that balance statistical accuracy and hydrological consistency.

Results show that ML models trained on quality-controlled data significantly outperform traditional statistical methods while preserving convective variability, which is essential for hydrological applications in tropical ecosystems. The study advances rainfall modeling by explicitly incorporating forest canopy interception processes. Using Machine Learning ensembles and hydrologically interpretable predictors, the models reveal strong spatiotemporal variability in interception, which cannot be adequately represented by constant coefficients.

Overall, the study demonstrates that rainfall reconstruction and its transformation into effective precipitation are inseparable processes in the ecohydrology of seasonal forests, providing reproducible tools for hydrological cycle analysis, ecohydrological modeling, and water resources management in tropical forest ecosystems.

Keywords: data-driven hydrology; machine learning; rainfall reconstruction; forest interception; ecohydrology; tropical seasonal forests.

INDICADORES DE IMPACTO

O presente trabalho apresenta impactos tecnológicos significativos ao estabelecer um arcabouço robusto para o controle de qualidade automático e a reconstrução de séries pluviométricas de alta resolução — a aplicação de algoritmos de aprendizado de máquina, como *Isolation Forest* e modelos multiobjetivo, permite a recuperação de dados em escalas horárias, reduzindo custos operacionais de manutenção e perdas históricas em redes de monitoramento; do ponto de vista socioambiental, a pesquisa contribui para a segurança hídrica ao propor modelos validados para a estimativa da precipitação efetiva em tempo real, utilizando a infraestrutura experimental única da Reserva Florestal "Matinha" (Mata Atlântica) e gerando subsídios científicos para políticas de conservação e manejo de bacias hidrográficas; o projeto possui caráter extensionista potencial através da futura disponibilização de ferramentas de código aberto para técnicos e pesquisadores, classificando-se nas áreas temáticas de **5 - Meio ambiente** e **7 - Tecnologia e produção**, e alinhando-se aos Objetivos de Desenvolvimento Sustentável (ODS) 6 (Água Potável e Saneamento), 13 (Ação Contra a Mudança Global do Clima) e 15 (Vida Terrestre).

IMPACT INDICATORS

This doctoral research delivers substantial technological impacts by establishing a robust framework for automated quality control and the reconstruction of high-resolution rainfall time series — by leveraging advanced machine learning algorithms such as *Isolation Forest* and multi-objective selection models, this work enables the recovery of sub-daily data, thereby mitigating operational maintenance costs and preventing the loss of historical records in hydrological monitoring networks; from a socio-environmental perspective, the study enhances water security by proposing validated models for real-time effective precipitation estimation, utilizing the unique experimental infrastructure of the "Matinha" Atlantic Forest reserve and providing scientific evidence to support forest conservation policies and watershed management; the project holds potential extensionist value through the future dissemination of open-source computational tools for practitioners and researchers, and is classified under the thematic areas of **5 - Environment** and **7 - Technology and Production**, aligning with the United Nations Sustainable Development Goals (SDGs): SDG 6 (Clean Water and Sanitation), SDG 13 (Climate Action), and SDG 15 (Life on Land).

LIST OF FIGURES

LIST OF TABLES

LIST OF BOXES

LIST OF ABBREVIATIONS

Apr.	April
Aug.	August
cf.	compare (*confer*)
DBH	Diameter at Breast Height (1.30 m)
Dec.	December
Dr.	Doctor
DSc.	Doctor of Science
e.g.,	for example (*exempli gratia*)
Eq.	Equation
et al.	and others (*et alii*)
Feb.	February
Fig.	Figure
ha	hectare
i.e.,	that is (*id est*)
Jan.	January
Jul.	July
Jun.	June
km	kilometre
Mar.	March
May	May
MG	Minas Gerais
mm	millimetre
mm h^{-1}	millimetres per hour
mm yr^{-1}	millimetres per year
MSc.	Master of Science
Nov.	November
Oct.	October
p.	page
pp.	pages
Prof.	Professor

Sep.	September
SP	São Paulo
Tab.	Table

LIST OF ACRONYMS

ABNT	Associação Brasileira de Normas Técnicas (Brazilian Technical Standards Association)
ACF	Autocorrelation Function
AI	Artificial Intelligence
API	Antecedent Precipitation Index
ARIMA	AutoRegressive Integrated Moving Average
CAMELS	Catchment Attributes and Meteorology for Large-sample Studies
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Brazilian Higher Education Personnel Improvement Coordination)
CatBoost	Categorical Boosting
CDF	Cumulative Distribution Function
CDR	Climate Data Record
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station data
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazilian National Council for Scientific and Technological Development)
CV	Coefficient of Variation
DBH	Diameter at Breast Height
DL	Deep Learning
ECMWF	European Centre for Medium-Range Weather Forecasts
EDA	Exploratory Data Analysis
ENSO	El Niño–Southern Oscillation
ERA5	ECMWF Reanalysis v5
iForest	Isolation Forest
INMET	Instituto Nacional de Meteorologia (Brazilian National Meteorological Institute)
IQR	Interquartile Range
ITCZ	Intertropical Convergence Zone
KGE	Kling–Gupta Efficiency
KNN	*k*-Nearest Neighbors

LAI	Leaf Area Index
LightGBM	Light Gradient Boosting Machine
LLM	Large Language Model
LOF	Local Outlier Factor
LOPO	Leave-One-Point-Out
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MERGE	Model for Precipitation Estimation by Ensemble of Radar and Stations (*Modelo de Estimativa de Precipitação por Ensemble de Radar e Estações*, INPE/INMET blended product)
ML	Machine Learning
MSE	Mean Squared Error
NSE	Nash–Sutcliffe Efficiency
PACF	Partial Autocorrelation Function
PDF	Portable Document Format
PERSIANN-CDR	Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks – Climate Data Record
PROEC	Pró-Reitoria de Extensão e Cultura (Office of Extension and Culture – UFLA)
Prophet	Additive time-series forecasting model (Meta)
PRPG	Pró-Reitoria de Pós-Graduação (Office of Graduate Studies – UFLA)
QC	Quality Control
QM	Quantile Mapping
R^2	Coefficient of Determination
RBF	Radial Basis Function (kernel)
RF	Random Forest
RF-FESM	Forest Reserve of Montane Semideciduous Seasonal Forest
RMSE	Root Mean Square Error
SACZ	South Atlantic Convergence Zone
SARIMA	Seasonal AutoRegressive Integrated Moving Average

SDG	Sustainable Development Goals
SHAP	SHapley Additive exPlanations
SIRGAS	Geocentric Reference System for the Americas (*Sistema de Referência Geocêntrico para as Américas*)
STL	Seasonal-Trend Decomposition via LOESS
SVR	Support Vector Regression
SWC	Soil Water Content
TSCV	Time-Series Cross-Validation
UFLA	Universidade Federal de Lavras (Federal University of Lavras)
UFMG	Universidade Federal de Minas Gerais (Federal University of Minas Gerais)
UFPel	Universidade Federal de Pelotas (Federal University of Pelotas)
UN	United Nations
URL	Uniform Resource Locator
UTM	Universal Transverse Mercator
XGBoost	Extreme Gradient Boosting

LIST OF SYMBOLS

P	Total gross precipitation above the canopy, mm
ET	Actual evapotranspiration, mm
Q	Surface and subsurface runoff to watercourses (streamflow), mm
R	Groundwater recharge, mm
ΔS	Change in soil water storage, mm
P_g	Gross rainfall above the canopy per event, mm
P_i	Throughfall — fraction of rainfall passing through the canopy, mm
P_{eff}	Effective precipitation: $P_{\text{eff}} = P_i + S_f$, mm
P_e	Total rainfall depth per event (event gross rainfall), mm
P_{corr}	Bias-corrected precipitation via quantile mapping, mm
$T_{(j,t)}$	Throughfall at spatial point j at time t, mm
S_f	Stemflow — water routed down the trunk to the soil, mm
I	Canopy interception loss: $I = P_g - P_{\text{eff}}$, mm
C	Effective precipitation coefficient: $C = P_{\text{eff}}/P_e$ (dimensionless)
S	Canopy storage capacity — maximum water retained before drainage begins, mm
P'_g	Minimum rainfall required to saturate the canopy in the Gash model (saturation threshold), mm
$\bar{c}$	Mean canopy cover fraction (dimensionless)
$\bar{E}$	Mean evaporation rate from a saturated canopy, mm h^{-1}
$\bar{R}$	Mean rainfall intensity during rain periods, mm h^{-1}
p	Free throughfall coefficient — fraction of rainfall reaching the ground without contacting vegetation (dimensionless)
W_t	Canopy water storage at time t, mm
D_t	Canopy drainage rate at time t, mm h^{-1}
D_s	Drainage rate under saturated canopy conditions, mm h^{-1}
κ	Drainage decay parameter in the Rutter model
KGE	Kling–Gupta Efficiency: $\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$
NSE	Nash–Sutcliffe Efficiency: $\text{NSE} = 1 - \frac{\sum_t (Q_{\text{obs,t}} - Q_{\text{sim,t}})^2}{\sum_t (Q_{\text{obs,t}} - \bar{Q}_{\text{obs}})^2}$

RMSE	Root Mean Square Error: $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(C_i - \hat{C}_i)^2}$, mm
MAE	Mean Absolute Error: $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert C_i - \hat{C}_i\rvert$, mm
R^2	Coefficient of Determination: $R^2 = 1 - \frac{\sum(C_i-\hat{C}_i)^2}{\sum(C_i-\bar{C})^2}$
r	Pearson correlation coefficient between observed and simulated values (temporal synchrony term in KGE)
α	Ratio of simulated to observed standard deviation — variability component of KGE: $\alpha = \sigma_{\mathrm{sim}}/\sigma_{\mathrm{obs}}$
β	Ratio of simulated to observed mean — bias component of KGE: $\beta = \mu_{\mathrm{sim}}/\mu_{\mathrm{obs}}$
γ	Third KGE component (CV-based variability variant): $\gamma = \mathrm{CV}_{\mathrm{sim}}/\mathrm{CV}_{\mathrm{obs}}$
$\mathrm{Borda}(m)$	Borda score for model m: $\mathrm{Borda}(m) = \sum_{k=1}^{K} r_k(m)$ — sum of ranks across K metrics
DOY	Day of Year, ranging from 1 to 365
$\mathrm{DOY}_{\mathrm{sin}}$	Sine-transformed day of year: $\sin\left(\frac{2\pi\cdot\mathrm{DOY}}{365}\right)$
$\mathrm{DOY}_{\mathrm{cos}}$	Cosine-transformed day of year: $\cos\left(\frac{2\pi\cdot\mathrm{DOY}}{365}\right)$
$y(t)$	Prophet decomposed time series: $y(t) = g(t) + s(t) + h(t) + \varepsilon_t$
$g(t)$	Trend component of the Prophet decomposition
$s(t)$	Seasonality component of the Prophet decomposition
$h(t)$	Holidays and special-event component of the Prophet decomposition
ε_t	Residual error (noise) at time t
$\Phi_P(B^s)\phi_p(B)\nabla^d\nabla_s^D Y_t$	SARIMA model equation: seasonal (Φ_P, B^s) and non-seasonal (ϕ_p, B) operators, differencing (∇^d, ∇_s^D) applied to series Y_t; errors ε_t with moving-average operators $\Theta_Q(B^s)\theta_q(B)$
$\hat{x}_t$	Bias-corrected value of variable x via quantile mapping at time t
$F_P(\cdot)$	Empirical cumulative distribution function (CDF) of the predictor (satellite or model)
$F_O^{-1}(\cdot)$	Quantile function (inverse CDF) of the observed reference data
$\hat{P}$	Precipitation estimated by Inverse Distance Weighting (IDW), mm

d_i	Distance between the estimation point and neighbouring station i (IDW), km
β	Distance-decay exponent in IDW (typically $\beta = 2$)
N	Number of neighbouring stations used in IDW interpolation
$\hat{Y}$	Value predicted by multiple linear regression
β_j	Regression coefficient for predictor X_j in multiple linear regression
X_j	Predictor variable j in the regression model
$\hat{Y}_{\text{RF}}(x)$	Random Forest prediction — mean of B decision trees $T_b(x)$
B	Number of decision trees in the Random Forest ensemble
$T_b(x)$	Decision tree b trained on a bootstrap subsample of the data
$F_m(x)$	Boosting model at iteration m: $F_m(x) = F_{m-1}(x) + \eta\, h_m(x)$
η	Learning rate in Gradient Boosting (dimensionless, $0 < \eta \leq 1$)
$h_m(x)$	Weak learner (regression tree) added at iteration m of boosting
$\mathscr{L}(\phi)$	XGBoost loss function: $\mathscr{L}(\phi) = \sum_i \ell(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$
G_j	First-order gradient of the loss function at leaf j (XGBoost)
H_j	Second-order gradient (Hessian) of the loss function at leaf j (XGBoost)
λ	L_2 regularisation parameter in XGBoost
w_j^*	Optimal leaf weight at leaf j in XGBoost: $w_j^* = -G_j/(H_j + \lambda)$
h_t	Hidden state of the LSTM cell at time t
x_t	Input vector to the LSTM at time t
f_t	Forget gate of the LSTM — controls how much of the previous cell state is retained
i_t	Input gate of the LSTM — controls how much new information enters the cell
o_t	Output gate of the LSTM — controls what is read from the cell state
c_t	LSTM cell state at time t: $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$

$\tilde{c}_t$	LSTM cell state candidate: $\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$
$\sigma(\cdot)$	Sigmoid function: $\sigma(z) = 1/(1+e^{-z})$ — used in LSTM gates
W_f, W_i, W_c, W_o	Weight matrices for the forget, input, cell, and output gates of the LSTM
b_f, b_i, b_c, b_o	Bias vectors for the LSTM gates
ϕ_j	SHAP value of variable j: weighted average marginal contribution over all coalitions $S \subseteq \mathscr{F} \setminus \{j\}$
$\mathscr{F}$	Complete set of predictor variables in the model (feature set)
$s(x,n)$	Isolation Forest anomaly score: $s(x,n) = 2^{-\mathbb{E}[h(x)]/c(n)}$; $s \approx 1$ indicates an anomaly
$h(x)$	Path length of observation x in an isolation tree
$c(n)$	Expected average path length for n samples — normalisation factor of iForest
$\mathrm{LOF}_k(x)$	Local Outlier Factor: ratio of the mean local density of the k-neighbours to the local density of point x; LOF > 1 indicates a potential anomaly
$\mathrm{lrd}_k(x)$	Local Reachability Density — inverse of the mean reachability distance to the k-nearest neighbours of x
$N_k(x)$	Set of k nearest neighbours of point x

CONTENTS

1 INTRODUCTION

Water is the defining environmental challenge of our time. Not as an abstract concept, but as a concrete problem that shapes where human societies can prosper and which ecosystems survive. This is especially true in tropical regions, where water moves through landscapes in intense, unpredictable ways. Tropical forests capture moisture from the air, hold it in soils and roots, and release it in patterns that influence climate and biodiversity across continents (HAMID; SAMANTARAY, 2025; SILVA; COSTA; CHRISTOFFERSEN, 2019; ZARGAR et al., 2025; ANDRADE et al., 2022). When these forests are degraded or destroyed, the consequences ripple through entire biogeochemical systems (RODRIGUES et al., 2021a; SANTOS; DURÁN-QUESADA; SÁNCHEZ-MURILLO, 2023; SILVA; MELLO, 2021; ANDRADE et al., 2024; MONEGO; ANOCHI; VELHO, 2023).

The Brazilian Atlantic Forest exemplifies this challenge. This biome is extraordinary: it harbors more species per unit area than the Amazon and sustains freshwater supplies for more than 125 million people (GUAUQUE-MELLADO et al., 2022; GUAUQUE-MELLADO et al., 2026). Yet it has been reduced to fragments, retaining less than 15% of its original extent. Understanding how this forest manages water—how it captures rainfall, stores it in soil, and channels it to aquifers and streams—is not merely academic. It is essential for protecting what remains and restoring what was lost. Yet tropical forests do not behave like the temperate forests that traditional hydrology was built to describe. Rainfall is intense and heterogeneous. The forest responds to dry seasons by shedding leaves. These nonlinearities challenge conventional models (RODRIGUES et al., 2021a; DEPIN et al., 2014; LEMOS et al., 2023).

This doctoral research is structured around a database: more than a decade of continuous, high-resolution hydrological monitoring—covering the period from September 2012 to December 2023—in a semideciduous Atlantic Forest experimental area known as “Matinha”, located on the campus of the UFLA (Minas Gerais) (RODRIGUES et al., 2021a; Junqueira Junior et al., 2019). This archive represents not only a historical record but a living laboratory that enables testing new hypotheses about how tropical forests partition rainfall and how it is possible to reconstruct rainfall time series with physical fidelity using AI tools.

Facing this challenge, climate scientists and water managers have turned increasingly to machine learning. There is good reason: ML algorithms can learn from data without the rigidity of traditional models. They can capture nonlinear relationships and adapt as conditions change—properties that are essential in convective tropical rainfall, which violates the linear assumptions baked into conventional hydrology (SHEN, 2018; BAşAğAOğLU et al., 2022).

Yet ML in hydrology is not a panacea. The field faces real obstacles. Tropical regions have sparse monitoring networks, so training data are scarce. Models can overfit—learning noise rather than signal. Most critically, ML models are often "black boxes": they generate predictions without explaining the physical processes underlying those predictions. For hydrology, interpretability matters. Water managers need not only accurate forecasts but also understanding: which rainfall variables matter most? How does forest structure influence water partitioning? (SCHMIDT et al., 2020; BAşAğAOğLU et al., 2022).

The thesis is articulated around two axes corresponding to two critical gaps in the current literature.

The first axis concerns the **reconstruction of precipitation time series**. Discontinuity in rainfall records—present even in the September 2012 to December 2023 series collected in "Matinha"—is not a bookkeeping issue — gaps in convective tropical series mean the loss of real hydrological events (ABOUZIED et al., 2025). Traditional imputation methods (arithmetic mean, inverse distance weighting, linear regression) assume stationarity and linearity, assumptions systematically violated in convective tropical rainfall (DEPIN et al., 2014; BLEIDORN et al., 2024). This research proposes to overcome these limitations through a methodological structure that merges global atmospheric reanalyses (ERA5) with nonlinear ML algorithms, prioritizing hydrological consistency and error minimization.

The second axis addresses the **modeling of rainfall interception and effective precipitation**. Upon reaching the forest canopy, gross precipitation is partitioned into three components: throughfall, stemflow, and interception, which corresponds to the fraction retained by the canopy and evaporated back to the atmosphere (Junqueira Junior et al., 2019; MUZYLO et al., 2009). The sum of the first two is effective precipitation (P_{eff}), which represents the volume of water that effectively reaches the soil and sustains infiltration and recharge processes.

Despite decades of research, interception continues to be represented in a simplified manner in many widely used hydrological models. Distributed models such as SWAT (*Soil and Water Assessment Tool*) and VIC (*Variable Infiltration Capacity*) calculate interception as a constant fraction of gross precipitation, limited by a maximum canopy storage capacity that does not vary with phenology (MUZYLO et al., 2009). The analytical Gash model (BY; GASH, 1979) and its revised version (GASH; LLOYD; LACHAUD, 1995) represent important conceptual advances by describing the process in terms of discrete events, but they require the estimation of physiological parameters that are difficult to obtain (canopy storage capacity, evaporation rate during rainfall) and assume that these parameters are stationary throughout the year—a critical limitation in semideciduous forests where up to 50% of species lose leaves during the dry season (RODRIGUES et al., 2021a; MUZYLO et al., 2009; VÉLIZ-CHÁVEZ et al., 2014). In summary, interception is treated in these models as a simplistic function of gross precipitation, ignoring its dependence on rainfall intensity thresholds and antecedent canopy moisture status.

A central scientific problem of this thesis, which deserves emphasis from the outset, is the absence of validated models capable of estimating effective precipitation in real time or under future scenarios from readily available climatic variables. This gap has direct implications for water resources management: without the ability to predict how much rainfall effectively reaches the soil under different canopy and climatic conditions, it is impossible to construct reliable water balances, model aquifer recharge, or anticipate the effects of extreme events on water availability in forested watersheds (RODRIGUES et al., 2021a; Junqueira Junior et al., 2019). This thesis proposes to overcome this limitation by demonstrating that ensembles of decision algorithms (RF and Gradient Boosting) can capture interception nonlinearities and canopy spatial heterogeneity (MOHAMMED; MENGISTE; GEBREMICHAEL, 2025; WU et al., 2024), providing operational estimates of effective precipitation from routinely monitored climatic variables.

In summary, this research aspires to establish a new methodological standard for tropical hydrology, demonstrating that “Data-Driven Hydrology”, when rigorously anchored in physical principles and field validation, can reveal patterns and processes that are invisible to conventional methodologies (SHEN, 2018).

1.1 Justification

The completion of this thesis is grounded in the convergence of a critical need for reliable hydrological information, the availability of new computational tools, and the existence of a long-term experimental infrastructure in the Brazilian context.

1.1.1 Ecohydrological Relevance and Water Security in the Atlantic Forest

The Atlantic Forest is one of the most threatened biomes and, paradoxically, one of the most vital for Brazil's water security (Junqueira Junior et al., 2019). Streamflow regulation, aquifer recharge, and flood mitigation intrinsically depend on the forest's capacity to intercept and redistribute precipitation (RODRIGUES et al., 2021a; OLIVEIRA et al., 2021). However, existing studies often rely on short-term campaigns or on models not validated under the region's marked seasonality conditions (Köppen Cwa) (Junqueira Junior et al., 2019). Understanding the fine dynamics of interception—how much water is lost through canopy evaporation and how much effectively reaches the soil as effective precipitation (P_{eff})—is crucial for modeling ecosystem resilience under climate change scenarios that project more intense rainfall events and longer droughts (RODRIGUES et al., 2021a).

1.1.2 Methodological Innovation Beyond Traditional Statistics

The methodological justification lies in overcoming the limitations of conventional approaches on two fronts:

- **Data reconstruction:** Conventional gap-filling methods assume stationarity and linearity, assumptions systematically violated in convective tropical rainfall (DEPIN et al., 2014; BLEIDORN et al., 2024). The use of DL algorithms, such as LSTM, and Gradient Boosting (XGBoost) enables capturing complex temporal dependencies and nonlinear relationships with large-scale atmospheric variables, offering a reconstruction that respects local climate physics (ABOUZIED et al., 2025; MONEGO; ANOCHI; de Campos Velho, 2022; WANG; TIAN; CARROLL, 2023; PINTHONG et al., 2025). Few studies address rainfall reconstruction in Atlantic Forest environments by integrating reanalysis and satellite data with bias correction and multiobjective model selection (ANDRADE et al., 2024; ASCENSO et al., 2024).
- **Interception:** Classical physical models (Gash, Rutter) require physiological parameterization that is difficult to obtain, involving intensive calibration that often does not reflect the real vegetative characteristics of forests (VÉLIZ-CHÁVEZ et al., 2014; MUZYLO et al., 2009). The proposed ML models overcome this limitation (SCHMIDT et al., 2020) by

learning the dynamics directly from data and explicitly incorporating variables such as rainfall intensity and antecedent moisture. Moreover, few investigations systematically evaluate models under simultaneous temporal and spatial validation schemes, a particularly critical gap in the Atlantic Forest, where previous studies focused on isolated sites or short periods (Junqueira Junior et al., 2019).

1.1.3 Experimental Infrastructure

The existence of a long-term monitoring infrastructure such as “Matinha” (UFLA) is central to this thesis. A systematic search of the Brazilian literature indicates that, although there are interception monitoring initiatives in biomes such as the Amazon, the Cerrado, and Atlantic Forest fragments in other states, none simultaneously presents: (i) continuous series longer than ten years (2012–2023); (ii) monitoring of rainfall partitioning at 32 spatial throughfall points and in 32 trees of 15 distinct species for stemflow; (iii) an hourly temporal resolution; and (iv) complementary soil moisture data at 32 points and five depths. This combination of temporal resolution, spatial density, and duration confers a singular character to the dataset for Brazilian forest ecohydrology, enabling investigations that shorter or less dense series would not permit.

1.1.4 Relationship with the Sustainable Development Goals (SDGs)

This project aligns directly with the United Nations 2030 Agenda (IPCC, 2021; NATIONS, 2015):

SDG 6 (Clean Water and Sanitation): The tools developed to reconstruct rainfall data and estimate effective precipitation improve the information base for water balance calculations and watershed management.

SDG 13 (Climate Action): Reliable modeling of extreme events and the use of global reanalyses (ERA5) allow local climate to be contextualized within global change.

SDG 15 (Life on Land): Interception modeling elucidates the forest’s functional role in regulating the water cycle, providing quantitative evidence for conservation and restoration policies.

1.2 Research Problem and Scientific Questions

1.2.1 Problem Definition

Research in the Atlantic Forest, and by extension in tropical forest hydrology, faces two interconnected barriers:

1. **Integrity of precipitation time series:** The historical precipitation series collected in "Matinha" (UFLA) between September 2012 and December 2023—totaling more than eleven years of continuous monitoring at hourly resolution—contains unavoidable gaps due to instrumental failures, power issues, and field maintenance. Conventional filling methods produce "smoothed" series that underestimate convective extremes. Furthermore, the detection of spurious values in high-frequency series is often manual and subjective. There is a lack of an automated protocol capable of cleaning and reconstructing the series while preserving its natural statistical structure and extreme events.

2. **Complexity of rainfall partitioning and lack of operational predictive models:** Despite extensive interception data collection (32 spatial points, September 2012 to December 2023), traditional models (Gash, Liu) did not allow event-by-event evaluation or an integrated spatiotemporal treatment. Critically, **there is no validated model that enables the estimation of effective precipitation in real time or under future scenarios from readily available climatic variables**—such as event magnitude, antecedent moisture, and seasonal indicators—directly limiting the operational applicability of these data for water resources management in forested watersheds.

1.2.2 Research Questions

General Question: In what way can the incorporation of advanced ML algorithms improve the accuracy, physical consistency, and operational utility of hydrological information (precipitation and interception) in seasonal forest ecosystems?

Specific Questions:

- **P1 (Reconstruction):** To what extent does the fusion of ERA5 reanalysis data with Gradient Boosting and DL algorithms, subjected to bias correction via Quantile Mapping, outperform purely statistical and satellite-based methods in reconstructing daily and hourly precipitation variability observed between September 2012 and December 2023?

- **P2 (Quality Control):** Are isolation-based anomaly detection algorithms (iForest) more effective than density-based methods (LOF) for identifying instrumental errors in rainfall series with many zeros, without removing valid extreme events?
- **P3 (Interception):** Which physical variables (intensity, duration, antecedent moisture, seasonality) dominate interception efficiency in the Atlantic Forest at the event scale observed between September 2012 and December 2023, and how can ensemble models (RF and Gradient Boosting) capture these nonlinear relationships to improve effective precipitation estimation?

1.3 Hypotheses

- **H1:** Precipitation time series reconstruction based on learning nonlinear physical relationships between large-scale atmospheric variables (ERA5) and local rainfall, using XGBoost and LSTM networks, produces synthetic series with significantly higher KGE efficiency metrics than those generated by SARIMA models, due to the ability of ML algorithms to incorporate the physics of atmospheric convection.
- **H2:** The use of unsupervised algorithms such as iForest enables more robust data cleaning in rainfall series than density-based methods (LOF) or fixed thresholds, because anomalies are isolated based on decision-tree path length, which is better suited to asymmetric distributions with many zeros.
- **H3:** Rainfall partitioning in Atlantic Forest ecosystems does not exhibit a constant proportion, but rather event-dependent dynamics driven by magnitude and antecedent moisture conditions. ML models, through hydrologically interpretable feature engineering, can represent the asymptotic behavior of interception by capturing short-term variations that static parameters fail to describe.

1.4 Objectives

1.4.1 General Objective

To develop and validate an integrated methodological framework for hydrology in a semideciduous Atlantic Forest remnant that, through the use of AI, enables: (i) robust reconstruction of historical precipitation series (September 2012 to December 2023); and (ii) accurate modeling of rainfall interception and effective precipitation processes at the event scale, ensuring hydrological coherence, spatiotemporal robustness, and physical interpretability of the results.

1.4.2 Specific Objectives

Linked to Article 1 (Reconstruction):

1. Implement and compare unsupervised outlier detection techniques to clean the historical precipitation database (September 2012 to December 2023).

2. Evaluate the performance of multiple exogenous sources (ERA5, CHIRPS, PERSIANN-CDR) and the effectiveness of Quantile Mapping for bias correction relative to local climatology.

3. Develop, optimize, and validate ML models (XGBoost, CatBoost, LSTM, Prophet) for daily and hourly imputation, using multiobjective selection (Pareto–Borda) balancing accuracy (RMSE) and structural consistency (KGE).

Linked to Article 2 (Interception):

4. Characterize the spatiotemporal variability of interception (I) and effective precipitation (P_{eff}) at the event scale in the Atlantic Forest, using the intensive monitoring network (September 2012 to December 2023).

5. Build and validate ensemble predictive models (RF, XGBoost, CatBoost, LightGBM) to estimate the effective precipitation coefficient ($C_{P_{\text{eff}}}$), incorporating feature engineering variables that represent rainfall intensity and system moisture memory.

6. Determine the relative importance of physical drivers of interception using interpretability techniques (SHAP values), differentiating behavior between dry and wet seasons.

2 THEORETICAL FRAMEWORK: DATA-DRIVEN HYDROLOGY FOR EFFECTIVE PRECIPITATION ESTIMATION IN TROPICAL FORESTS

Three conceptual axes organize the argument developed here. **Forest ecohydrology** describes how vegetation actively reshapes the water cycle by partitioning precipitation, regulating evapotranspiration, and modulating soil infiltration (BRASIL et al., 2022; CAU et al., 2025; RODRIGUES et al., 2020). It provides the physical grounding without which any predictive model risks learning statistically convenient but physically meaningless correlations. **Pluviometric data integrity** addresses the problem of gaps, outliers, and instrumental biases endemic to tropical monitoring networks (HERRERA; FOWLER, 2023; PAN et al., 2016; SEGOVIA-CARDOZO; BERNAL-BASURCO; RODRÍGUEZ-SINOBAS, 2023): such defects propagate through both physical model calibration and ML training, making quality control a methodological decision rather than a preparatory routine (SINGH et al., 2024; WAQAS, 2024). **Machine learning applied to hydrology** supplies the computational tools for capturing the nonlinear relationships and high spatiotemporal variability that classical physical models handle poorly; the decision to use *ensemble learning* algorithms and recurrent neural networks follows directly from documented limitations in the existing literature and from accumulated evidence that these methods outperform linear alternatives on processes with memory (SHEN, 2018; WILLARD et al., 2022; KRATZERT et al., 2019b).

The articulation between ecohydrology and data science

The convergence of forest ecohydrology and data science has been formalized as "data-driven hydrology", a field that complements rather than displaces process-based modeling. Willard et al. (2022) describe the *Physics-aware Machine Learning* (PaML) paradigm as admitting four modes of integration: physically data-guided learning, physics-informed learning, physics-embedded architecture, and coupled hybrid learning. Across all four, the operating principle is the same: physical knowledge constrains ML generalization, and ML algorithms surface patterns that process equations cannot extract from noisy, multivariate observations. This thesis sits at that intersection. The ecohydrology of the Atlantic Forest supplies the physical vocabulary and the hypotheses; the high-frequency time series from "Matinha" (UFLA) supply the data; and the ML algorithms supply the analytical language to connect both while preserving physical plausibility and operational relevance.

Logical structure of the chapter

The chapter moves from the concrete to the methodological. It opens with the physical processes governing water balance and rainfall partitioning in tropical forests (Sections 2.1 and 2.2), then shifts to the data problems those processes create: series integrity and quality control (Section 2.3), followed by the ML foundations used to address them (Section 2.4), anomaly detection algorithms (Section 2.5), and model evaluation criteria (Section 2.6).

2.1 The Hydrological Cycle in Tropical Forest Ecosystems

2.1.1 Foundations of Water Balance in Forests

The water balance of a tropical forest is a mass conservation equation. Water entering through precipitation must leave through evaporation, drainage to rivers, or groundwater recharge, or else it accumulates in storage. At basin scale, over a period long enough to average out seasonal swings, the relationship takes the form:

$$P = \mathrm{ET} + Q + R \pm \Delta S \tag{2.1}$$

where P is total precipitation (mm), ET actual evapotranspiration, Q surface and subsurface runoff to watercourses, R groundwater recharge, and ΔS the change in soil water storage (BRUIJNZEEL, 2004). In tropical forest ecosystems, the ET term is consistently dominant: humid forest evapotranspiration can consume between 60 % and 80 % of annual precipitation (BRUIJNZEEL, 2004), making it the primary pathway by which the forest recycles water to the atmosphere and, in doing so, modulates regional climate.

The regulating role of forest cover

The forest is an active participant in the hydrological cycle, operating across scales. At the event scale, the canopy intercepts a fraction of rainfall that evaporates before reaching the soil, with measurable consequences for aquifer recharge and surface runoff (Section 2.2). At the seasonal scale, phenology shifts the interception budget: leaf fall in semideciduous forests reduces storage capacity and transpiration, altering the dry-to-wet-season water balance. At the basin scale, root systems maintain macroporous, structured soils with high hydraulic conductivity that sustain infiltration even during intense convective events (Junqueira Junior et al., 2019). Atlantic Forest studies show that forest cover is positively associated with minimum flows (Q_7, the lowest mean seven-day discharge with a ten-year return period), confirming the forest's role in sustaining river baseflow through the dry season (ROCHA; CERQUEIRA et al., 2018). The water-security implications are straightforward: deforestation raises flood peaks and erosion rates while cutting the aquifer recharge that keeps streams perennial during drought.

Canopy structure and vertical water fluxes

Canopy structure, operationally described by leaf area index (LAI, $m^2\ m^{-2}$), crown architecture, vertical stratification, and species phenology, governs vertical water fluxes (LEVIA; CARLYLE-MOSES; TANAKA, 2011). A high LAI increases the surface area available for interception and evaporation. A dense canopy also reduces radiation and wind speed in the understory, sustaining a humid microclimate that suppresses evapotranspiration from lower strata. Branch orientation and bark roughness control how efficiently water held by the crown is routed down the trunk as stemflow. That component is volumetrically modest, typically 0.5–6 % of gross precipitation in tropical forests (LEVIA; CARLYLE-MOSES; TANAKA, 2011), but biogeochemically important: it concentrates crown-leached nutrients at the tree base and drives preferential infiltration in that zone, with lasting effects on soil microbiology (KEIM; SKAUGSET; WEILER, 2006).

Evergreen forests versus semideciduous forests

The contrast between evergreen and seasonally deciduous forests matters directly for the objectives of this thesis. Amazonian dense ombrophilous forests maintain a broadly stable canopy throughout the year, with consistently high LAI and near-constant interception capacity. Semideciduous forests, such as the Montane Seasonal Semideciduous Forest (FESM) that is the experimental area here, follow a different rhythm: during the dry season, between 20 % and 50 % of species shed their foliage partially or completely (RODRIGUES et al., 2021a). In the UFLA fragment, LAI ranges from 3.7 to 5.0 m^2 m^{-2} in the wet season and falls substantially through winter (RODRIGUES et al., 2021a). That seasonal shift is not merely a change in magnitude. Leaf fall alters crown architecture, reduces canopy storage capacity S (the rainfall depth required to saturate the canopy before drainage begins), and changes the aerodynamics that govern evaporation from intercepted water. A model calibrated during the wet season and applied unchanged to winter conditions will generate systematic errors, because the parameters themselves have changed. Classical static-parameter models cannot adjust for this without full recalibration, which is one concrete reason this thesis adopts ML algorithms that learn seasonal parameter variability directly from the data (Junqueira Junior et al., 2019).

2.1.2 Seasonality and Spatiotemporal Variability of Tropical Precipitation

Modeling effective precipitation in a semideciduous tropical forest requires, first and foremost, understanding the structure and variability of the input signal: gross precipitation. Southeastern Brazil — and in particular southern Minas Gerais, where the Matinha experimental site is located — is subject to a combination of atmospheric mechanisms that produces a markedly seasonal, spatially heterogeneous, and temporally non-stationary rainfall regime.

The South Atlantic Convergence Zone (SACZ)

The main system modulating precipitation during the austral summer in southeastern Brazil is the **South Atlantic Convergence Zone (SACZ)** (QUADRO, 1994; CARVALHO; JONES; LIEBMANN, 2004). The SACZ manifests as a band of cloudiness, deep convection, and moisture convergence oriented northwest–southeast, extending from the interior of the

Amazon basin to the Atlantic Ocean off the Brazilian coast, covering a meridional extent of approximately 5,000 km. Its formation is associated with the transport of heat and moisture from the Amazon basin by the low-level jet, which interacts with frontal systems originating in extratropical latitudes (CARVALHO; JONES; LIEBMANN, 2004). From a quantitative standpoint, the SACZ is responsible for a substantial fraction of total precipitation during the rainy season in southeastern Brazil. Nielsen et al. (2019) showed that the system accounts for between 25 % and 50 % of monthly totals during the rainy season, with the largest contributions in December and January; analyses specific to March indicate that the SACZ can explain up to 56 % of observed precipitation (ZILLI; HART, 2019). More than totals, it is important to recognize that the SACZ is intrinsically associated with extreme events: persistent and intensified SACZ episodes are linked to floods and mass movements that affect forested basins, making their correct representation in predictive models a risk-management issue (QUADRO, 1994; CARVALHO; JONES; LIEBMANN, 2004).

Complementary systems: frontal systems, easterly waves, and local convection

Beyond the SACZ, temporal variability in precipitation is modulated by synoptic and mesoscale systems operating at distinct time and space scales (QUADRO, 1994). Frontal systems coming from the south, when interacting with tropical moisture during transition months, can remain stationary for days and anchor SACZ formation, amplifying its effects. Easterly Wave Disturbances — vorticity waves that propagate from east to west along the lower troposphere — are responsible for precipitation events during the transition period between seasons in eastern Minas Gerais and on the slopes of the Serra da Mantiqueira. In winter, the absence of these systems produces a marked dry season, with water deficit imposing pronounced physiological stress on semideciduous vegetation. In this period, the few rainfall events that occur have a distinct dynamics: they fall on a canopy with reduced LAI and lower storage capacity S, producing interception values proportionally different from those observed in the rainy season — one of the reasons why constant-parameter models fail to adequately reproduce the annual behavior of the system (RODRIGUES et al., 2021a).

Non-stationarity and implications for modeling

Tropical convective precipitation is intrinsically non-stationary at multiple time scales. At the intraseasonal scale, alternation between active episodes and pauses of the SACZ produces intradecadal to interannual variability. At the interannual scale, phenomena such as the El Niño–Southern Oscillation (ENSO) and sea surface temperature anomalies in the tropical Atlantic modulate the position and intensity of the SACZ, directly affecting annual rainfall totals (CARVALHO; JONES; LIEBMANN, 2004). At the multidecadal scale, there is evidence of a progressive southward shift in the mean position of the SACZ in recent decades, with measurable impacts on regional precipitation patterns in southeastern Brazil (JONES; CARVALHO, 2019). From a statistical standpoint, tropical convective events are characterized by high intensities, short duration, high spatial heterogeneity, and strongly right-skewed distributions with heavy tails (*heavy-tailed*). These properties violate the assumptions of Gaussianity and stationarity required by classical statistical methods such as ARIMA and linear regression, and result in time series with complex, nonlinear autocorrelation structure. For interception modeling, this non-stationarity has a direct consequence: the canopy moisture state before an event — which determines its residual storage capacity — is a function not only of the immediately previous event, but of a time window whose length varies with atmospheric conditions, a “memory” that static and linear models cannot represent.

Cwa climate and the hydrological dynamics of the study area

The local climate of Lavras (MG), where the Matinha experimental site is located, falls within the Köppen Cwa classification (humid subtropical with rainy summer and dry winter), with mean annual temperature around 19.4 °C and mean annual precipitation of approximately 1,511 mm (Junqueira Junior et al., 2019). About 85 % of annual totals concentrate in the rainy period (October to March), while winter is characterized by an almost complete absence of precipitation between June and August. This marked seasonality imposes on the FESM forest ecosystem an annual cycle of alternation between saturation (when LAI is maximum and soils are near field capacity) and severe water deficit (when vegetation draws on deep water reserves to maintain metabolism during drought). Understanding this cycle is essential to interpret interception data collected over more than a decade in the UFLA fragment (RODRIGUES et al., 2021a).

2.1.3 Hydrological Ecosystem Services of Tropical Forests

The Atlantic Forest retains hydrological functions that extend well beyond its ecological boundaries. With remnants covering roughly 12 % of the biome's original extent, the forest still supplies water resources to more than 125 million Brazilians (SOS Mata Atlântica; INPE, 2021). That supply depends on the continued delivery of hydrological ecosystem services by scattered fragments embedded in agricultural and urban landscapes.

Water regulation services

The first and most recognized hydrological service of forests is flow regulation. By intercepting part of precipitation and promoting infiltration through the root system, the forest attenuates flood peaks (reducing downstream flood risk) and maintains baseflow during drought (ensuring the perenniality of watercourses). This regulating function depends on canopy integrity and soil structure: a fragment with reduced LAI due to water stress or degradation intercepts less rain, generates more surface runoff, and recharges aquifers less (BRUIJNZEEL, 2004). Aquifer recharge is favored by the high macroporosity and hydraulic conductivity of the Distroferric Red Latosols predominant in the UFLA region — soils that, under intact forest cover, show infiltration rates substantially higher than those of compacted soils under pasture or cultivation. Studies indicate that the forest allows recharge even during periods of intense convective precipitation by temporally distributing water flux within the soil profile (Junqueira Junior et al., 2019). The observable result is the forest's capacity to maintain river baseflow (Q_7) even during prolonged droughts (ROCHA; CERQUEIRA et al., 2018). Mitigation of water erosion is another service of high social value. Forest interception reduces the kinetic energy of raindrops reaching the soil, decreasing surface sealing and particle detachment rates. The result is a drastic reduction in sediment transport compared to deforested areas, with implications for water quality, reservoir siltation, and geomorphic slope stability (BRUIJNZEEL, 2004).

Threats from climate change

Projected climate change scenarios for southeastern Brazil are concerning from the perspective of hydrological services. The Intergovernmental Panel on Climate Change (IPCC, 2021) projects an intensification of the hydrological cycle in the region, with precipitation more concentrated in time (more intense events of shorter mean duration), longer dry spells, and increased interannual variability. For Atlantic Forest semideciduous forests, this implies larger oscillations between seasonal canopy saturation and water deficit, with the potential to alter species phenology, compromise aquifer recharge, and eventually trigger state-transition thresholds (*tipping points*) in ecosystems. Long-term monitoring at Matinha (2012–2023) has already documented this pressure: extreme drought episodes in the region produced measurable shifts in rainfall partitioning dynamics, with the forest maintaining core functions but exhibiting changes in interception patterns that static-parameter models cannot track (RODRIGUES et al., 2021a). Capturing that kind of adaptive, non-stationary behavior is exactly what motivates the ML algorithms developed in this thesis.

The indispensability of reliable models

Reliable estimation of effective precipitation matters beyond the laboratory. Basin water management, environmental licensing, land-use planning, and calibration of large-scale hydrological models all depend on it. Errors in interception estimation are not contained: overestimating interception leads to underestimating effective precipitation, resulting in consistently low runoff predictions, low recharge estimates, and exaggerated projections of water stress. Model quality here has direct consequences for management quality (SHEN, 2018; MUZYLO et al., 2009).

2.2 Rainfall Partitioning by the Forest Canopy: Physical and Conceptual Foundations

2.2.1 The Partitioning Process: Concepts and Components

When a raindrop leaves the clouds and begins its fall toward the ground, the forest is not necessarily its final destination. Upon reaching the canopy, water can take three distinct paths, and the proportion among them — which varies with each event, each species, and each season — determines how much water effectively supplies the soil and, consequently, subsurface hydrological processes. These three paths are the three components of rainfall partitioning (BY; GASH, 1979; LEVIA; CARLYLE-MOSES; TANAKA, 2011).

Throughfall (*throughfall*, P_i)

Throughfall is the fraction of rainfall that passes through the canopy and reaches the forest floor. It is composed of two distinct mechanisms, although often indistinguishable in the field: *free throughfall*, corresponding to drops that pass directly through canopy openings (canopy openness fraction p) without contact with vegetation; and *drip throughfall*, which is water first intercepted by leaves and branches and, after surface storage reaches the limit of capillary retention, is released as drops of larger diameter and higher kinetic energy (BY; GASH, 1979; LEVIA; CARLYLE-MOSES; TANAKA, 2011). Quantitatively, P_i is typically the dominant component of partitioning: in tropical forests, it represents between 60 % and 90 % of gross precipitation (LEVIA; CARLYLE-MOSES; TANAKA, 2011). In an FESM remnant in the northeastern Atlantic Forest, for example, Filho et al. (2019) reported throughfall of 52.9 % of gross precipitation — a relatively low value reflecting the high density and thickness of the canopy at that site.

Stemflow (*stemflow*, S_f)

Stemflow is the fraction of rainfall that, intercepted by crowns and conveyed by branches and twigs toward the bole, runs down the tree bark until reaching the soil at the base of the trunk. Although volumetrically modest — typically between 0.5 % and 6 % of gross precipitation in tropical forests (LEVIA; CARLYLE-MOSES; TANAKA, 2011) — *stemflow* has hydrological and biogeochemical relevance disproportionate to its magnitude, because it concentrates nutrients leached from bark and crowns and deposits them locally in a zone of high root activity, favoring preferential infiltration in this region (KEIM; SKAUGSET; WEILER, 2006). The efficiency of S_f generation varies among species: species with smooth bark and upward-oriented branches produce more *stemflow* than species with rough bark and horizontal or downward branches (LEVIA; CARLYLE-MOSES; TANAKA, 2011).

Interception (I) and effective precipitation (P_{eff})

Interception is the fraction of gross precipitation temporarily retained on vegetative surfaces — leaves, branches, bark, lichens, epiphytes — that is returned to the atmosphere by evaporation during and after the rainfall event, without ever reaching the soil. In a strict sense, interception is not a flux (a transfer of mass from one compartment to another), but a loss: the water never completes the transition from the atmosphere to the soil (BY; GASH, 1979). Its magnitude can be substantial: in the northeastern Atlantic Forest, Filho et al. (2019) reported interception losses of 44.9 % of gross precipitation. In FESM fragments in southeastern Brazil, interception values tend to be lower — Junqueira Junior et al. (2019) reported approximately 12.1 % interception — reflecting differences in canopy structure, floristic composition, and local climatic conditions. The mass balance of these three components can be formulated as:

$$P_g = P_i + S_f + I \tag{2.2}$$

where P_g is gross precipitation (*gross rainfall*). **Effective precipitation** is defined as the sum of throughfall and stemflow:

$$P_{\text{eff}} = P_i + S_f \tag{2.3}$$

representing the total volume of water that effectively reaches the forest floor and is available for infiltration, surface runoff, and groundwater recharge (BY; GASH, 1979; LEVIA; CARLYLE-MOSES; TANAKA, 2011). Table 2.1 synthesizes values reported in the literature for the three partitioning components across different Atlantic Forest forest typologies, illustrating variability among studies and sites.

The variability across studies reflects real differences in canopy structure, species composition, phenology, and local rainfall characteristics. The fact that no fixed set of parameters travels well from one site to another, or from one season to the next, is part of what motivates the machine learning approach developed here.

2.2.2 Physical Factors Controlling Interception

Interception efficiency at any given event results from the interaction between atmospheric variables (rainfall characteristics and evaporative demand) and canopy variables (structure,

Table 2.1 – Values of rainfall partitioning components across different Atlantic Forest typologies and global tropical averages.

Study/Site	Typology	$P_i\,(\%)$	$S_f\,(\%)$	$I\,(\%)$
Filho et al. (2019)	FESM — Northeastern Brazil	52.9	2.2	44.9
Junqueira Junior et al. (2019)	Montane FESM — Southeastern Brazil	87.3	0.65	12.1
Rodrigues et al. (2021a)	Montane FESM — UFLA/MG (7-year period)	78.3	0.5	21.2
Mello et al. (2024)	Global meta-analysis — Tropical forests	70–75	< 1.0	10–30

P_i: throughfall; S_f: stemflow; I: interception.

Fonte: The author (2026).

phenology, and storage capacity). Identifying which of these factors dominate, and under what conditions, is a prerequisite for physically sound modeling.

Climatic factors: magnitude, intensity, duration, and wind

Among climatic variables, the **total event magnitude** (P_g, in mm) is the first-order control factor. The relationship between P_g and I is nonlinear and asymptotic: small-magnitude events can be almost entirely intercepted (when $P_g < S$, all rainfall is retained by the canopy without reaching the soil); as P_g increases, the interception proportion decreases, until sufficiently large events produce an increasingly dominant throughfall fraction. By e Gash (1979) showed that this asymmetry is captured by the three discrete phases of its model: wetting (where $I \approx P_g$), saturation (where I increases at an approximately constant rate $\bar{E}/\bar{R}$), and drying (where evaporation depletes residual storage after rainfall ends). **Rainfall intensity** (R, in mm h^{-1}) is especially determinant during the saturation phase: at a constant evaporation rate $\bar{E}$, a higher intensity R implies that the rainfall required to saturate the canopy (P'_g) is reached more rapidly and with lower accumulated volume, increasing the throughfall proportion (BY; GASH, 1979). On the other hand, very high intensities, by saturating the canopy within seconds, can in practice reduce proportional interception by preventing evaporation from occurring during the event. **Event duration** and internal rainless intervals (*rainless gaps*) are likewise relevant: pauses within an event allow a partially dried canopy to regenerate storage capacity, increasing total accumulated interception in long, fragmented events (BY; GASH, 1979). **Wind speed** (u) acts in two opposing directions on interception: on the one hand, it increases the evaporation rate of the wet canopy by renewing humid air at the leaf–atmosphere interface, increasing I losses; on the other hand, it can drive oblique rainfall to penetrate more deeply into the canopy and reach the soil before being intercepted, reducing I. The resulting balance depends on canopy structure and the specific conditions of the event (LEVIA; CARLYLE-MOSES; TANAKA, 2011).

Canopy storage capacity (S) and the free drainage coefficient (p)

On the canopy side, the two central physical parameters in interception models are canopy storage capacity S (mm) and the free drainage coefficient p (dimensionless) (BY; GASH, 1979; RUTTER; MORTON; ROBINS, 1975). S represents the maximum volume of water the canopy can retain by capillary retention — on leaves, branches, bark, epiphytes — before gravitational drainage begins. In tropical forests, S typically ranges between 0.5 and 3.0 mm, depending on leaf density and surface texture, the presence of epiphytes and mosses, and LAI. The coefficient p describes the fraction of rainfall that passes directly through canopy gaps without contact with vegetation (canopy openness fraction); typical values range between 0.05 and 0.45 (MUZYLO et al., 2009). Both parameters vary seasonally in semideciduous forests. At Matinha (UFLA), S decreases significantly during the dry season, reflecting LAI reduction with leaf fall, while p increases because canopy openness expands with foliage loss (RODRIGUES et al., 2021a). Treating these parameters as constant throughout the year — as in the original formulations of the classical Gash and Rutter models — introduces systematic errors that manifest as overestimation of interception in the dry season and underestimation in the rainy season (MUZYLO et al., 2009).

Nonlinearity and dependence on antecedent states

Rainfall intensity and **antecedent canopy wetness** are considered the main *drivers* of nonlinear interception behavior (BY; GASH, 1979; LEVIA; CARLYLE-MOSES; TANAKA, 2011). When the canopy is already wet from a recent event, its residual storage capacity — $S_{res} = S - W$, where W is the water currently retained — is reduced. A subsequent event, even of small magnitude, may exceed this residual threshold more quickly, generating throughfall that would not occur under dry-canopy conditions. This state dependence configures a system with **hydrological memory**: the canopy response to a rainfall event depends on the conditions of preceding events, which poses a fundamental challenge for static-parameter models and justifies the use of neural networks capable of representing temporal states (LSTM), developed in Section 2.4.

Spatial heterogeneity of the canopy

The three-dimensional structure of the forest canopy is not spatially homogeneous: variation in crown density among species, diameter at breast height (DBH), foliar architecture, and the sociological position of trees produces a spatially heterogeneous distribution of throughfall on the forest floor. Studies using dense rain gauge networks have documented the existence of throughfall "hot spots" (*hot spots*) — located beneath canopy gaps or preferential drip points from thick branches — and "rain shadows" under dense crowns (LEVIA; CARLYLE-MOSES; TANAKA, 2011). The Matinha monitoring network, with 32 understory rain gauges distributed under different species, was designed to capture this variability: previous work at the same site showed that species such as *Miconia willdenowii* record the highest P_{eff} values, whereas *Xylopia brasiliensis* presents the lowest, reflecting differences in foliar architecture and crown density (RODRIGUES et al., 2021a).

Phenology and seasonality

In semideciduous forests, phenology represents the main factor controlling temporal variation of interception at the seasonal scale. Leaf fall during the dry season reduces LAI and, with it, both the surface available for interception and the storage capacity S. This means that, during the dry season, even small-magnitude rainfall events — which in the rainy season would be almost entirely intercepted — generate throughfall proportions larger than expected by models calibrated with rainy-period data. This dual seasonal variation — higher interception in the rainy season (high LAI, high S) and lower in the dry season (low LAI, low S) — invalidates the use of constant coefficients throughout the year and requires approaches that explicitly represent phenological dynamics (Junqueira Junior et al., 2019; RODRIGUES et al., 2021a).

2.2.3 Classical Physical Models of Interception: Gash, Rutter, and Derived Approaches

Analytical and numerical modeling of forest interception has a history spanning more than five decades, with two foundational models that still dominate the literature: the analytical model of Gash (BY; GASH, 1979) and the continuous-process model of Rutter (RUTTER; MORTON; ROBINS, 1975). Both seek to describe the same physical processes, but they do so with different approximations and temporal resolutions, which condition their applicability in different contexts.

The Gash Model (1979) and its revised versions

The Gash model (BY; GASH, 1979) is the most widely applied in forest hydrology, with extensive use worldwide including multiple Atlantic Forest studies (MUZYLO et al., 2009). The formulation rests on treating precipitation as a series of discrete, independent events, each passing through three physically distinct phases. In the **wetting phase**, the canopy is still below its saturation capacity ($W < S$), and all precipitation is retained by the canopy without drainage to the ground. The precipitation volume required to saturate the canopy, P'_g, is obtained from the balance between rainfall input and evaporation occurring during this process:

$$P'_g = -\frac{S}{\bar{c}} \ln\left(1 - \frac{\bar{E}}{\bar{c}\bar{R}}\right) \approx \frac{S/\bar{c}}{1 - \bar{E}/(\bar{c}\bar{R})} \tag{2.4}$$

where $\bar{c}$ is the canopy cover fraction (dimensionless), $\bar{E}$ is the mean evaporation rate during the event (mm h^{-1}), and $\bar{R}$ is the mean rainfall intensity (mm h^{-1}) (BY; GASH, 1979). The ratio $\bar{E}/\bar{R}$ expresses the proportion of rainfall evaporated during wetting — a central parameter that, in the original model, is assumed constant for all events. In the **saturation phase**, the canopy is wet and the drainage rate to the ground is approximately equal to rainfall rate minus evaporation rate. The accumulated interception during this phase is proportional to the amount of rainfall and to the ratio $\bar{E}/\bar{R}$. In the **drying phase**, precipitation has ceased, but the canopy still contains stored water that will be gradually evaporated until it returns to a dry state, ready for a new event. The general equation for total canopy interception over a set of m small events (with $P_{g,j} < P'_g$) and n large events (with $P_{g,j} \geq P'_g$) in the formulation revised by Gash, Lloyd e Lachaud (1995) is:

$$
\begin{aligned}
I \;=\; \bar{c}\sum_{j=1}^{m} P_{g,j} \;+\; \bar{c}\sum_{j=1}^{n}\left[P'_g + \frac{\bar{E}}{\bar{R}}\left(P_{g,j} - P'_g\right)\right] \\
+\; \bar{c}\,S\,(n - \text{number of events in which the canopy does not dry between events})
\end{aligned}
\tag{2.5}
$$

where the first term represents total interception in small events, the second in large events, and the third drying losses (BY; GASH, 1979; GASH; LLOYD; LACHAUD, 1995). The revised version by Valente e David (1997) extends the model to discontinuous (*sparse canopy*) cover, explicitly differentiating canopy and trunk components and making the model more suitable for forest fragments and savanna formations. This version is used in the global model GLEAM (*Global Land Evaporation Amsterdam Model*) to estimate interception losses at the global scale from remotely sensed LAI products (MIRALLES et al., 2010).

The Rutter Model (1975)

The Rutter model (RUTTER; MORTON; ROBINS, 1975) adopts a radically different approach: rather than treating events as discrete objects, it solves the canopy water balance continuously in short time steps (typically hourly). The canopy is modeled as a reservoir with maximum capacity *S*; at each time step, precipitation increases storage, evaporation reduces it, and drainage occurs when storage exceeds *S* according to an exponential function:

$$
D_t \;=\; D_s \exp[\kappa\,(W_t - S)] \quad \text{if } W_t > S \tag{2.6}
$$

where D_t is the drainage rate at time t (mm h^{-1}), D_s and κ are scale and shape parameters of the exponential function, and W_t is current canopy storage (RUTTER; MORTON; ROBINS, 1975). Evaporation during the event is calculated using the Penman–Monteith equation with zero surface resistance (fully wet canopy). The main advantage of the Rutter model over Gash is higher physical fidelity for individual events with variable intensity and internal interruptions: by solving the balance in short time steps, the model captures canopy filling and emptying dynamics in much greater detail than Gash's discrete-phase approximation. Its main limitation is that it requires high-temporal-resolution meteorological data (hourly precipitation, wind speed, radiation, temperature, and humidity) continuously available — a requirement that is often not met in operational monitoring networks in tropical regions (MUZYLO et al., 2009).

Parameter calibration

Both models require empirical estimation of parameters that are not directly observable: S (canopy storage capacity), p (free drainage coefficient), $\bar{E}$ (evaporation rate during the event), and $\bar{c}$ (cover fraction). These parameters are typically obtained by regression of accumulated throughfall versus accumulated gross precipitation for selected events: the slope of the fitted line for large events estimates $1-\bar{c}$ (or $1-\bar{E}/\bar{R}$ in the Gash formulation), while the intercept provides an estimate of $\bar{c} \cdot S$ (GASH; LLOYD; LACHAUD, 1995). Studies such as Rodrigues et al. (2021a) and Junqueira Junior et al. (2019) reported substantial variation in parameters calibrated for Matinha across different seasons, showing that static calibration underestimates the true variability of the system.

Comparative performance of the models

The performance of the Gash and Rutter models is generally satisfactory in temperate conifer forests and in humid tropical forests with relatively homogeneous and persistent canopy structure (MUZYLO et al., 2009). In a comprehensive review of nearly 80 studies, Muzylo et al. (2009) found that the Gash models (original and revised sparse versions) were the most frequently applied, preferred over the Rutter approach for their ease of use, lower data requirements, and consistently reported errors — with both models achieving similar predictive accuracy in most tested contexts. However, in forests with strong phenological seasonality and climatic conditions that favor additional energy sources for evaporation — such as Atlantic Forest Montane FESM fragments during prolonged dry periods — both models tend to exhibit substantial systematic errors, whose physical origin is discussed in the next section.

2.2.4 Limitations of Classical Models in Seasonal Tropical Forests

The Gash and Rutter models work well in many settings, but they carry a set of structural limitations that accumulate in seasonal tropical forests, which is the environment of the UFLA experimental site. The following subsections describe those limitations and their hydrological consequences.

The problem of parameter nonstationarity in seasonal environments

Interception models by Gash and Rutter use canopy parameters — S, p, and $\bar{E}$ — assumed constant throughout the simulation period (BY; GASH, 1979; RUTTER; MORTON; ROBINS, 1975). This assumption is appropriate when applied to hydrologically homogeneous event sets or to periods with approximately stable vegetation structure. However, in ecosystems with marked phenological seasonality, this condition is no longer satisfied. In tropical semideciduous forests, seasonal variation in leaf area index (LAI) substantially alters canopy structure, modifying storage capacity (S), cover fraction, and preferential drainage pathways (p) (MUZYLO et al., 2009). Thus, the relationship between gross precipitation and interception becomes nonstationary over the hydrological year. In the Matinha Montane FESM (UFLA), LAI varies approximately between 3.7 and 5.0 $m^2\ m^{-2}$ in the rainy season, decreasing during the dry season (RODRIGUES et al., 2021a). Because S is directly related to foliar biomass, a model calibrated under leaf-on conditions and applied without recalibration to the dry period tends to overestimate interception and underestimate throughfall, introducing systematic bias in the annual water balance. Studies at the site show marked seasonal variation in Gash model parameters, indicating that the static-parameter assumption is not adequate to represent the full phenological cycle of the ecosystem (RODRIGUES et al., 2021a).

Simplified representation of interception in catchment hydrological models

Widely used catchment hydrological models, such as SWAT, HBV, and VIC, do not explicitly implement canopy interception schemes based on detailed physical processes, such as the Gash or Rutter models. Instead, interception is represented by a conceptual canopy storage reservoir, in which incident precipitation first fills a maximum vegetation storage capacity (S), and the excess is converted into effective precipitation (throughfall). Conceptually, the process can be described by:

$$I = \min(P_g,\ S - W_{t-1}) \tag{2.7}$$

$$P_{\text{eff}} = P_g - I \tag{2.8}$$

where P_g is gross precipitation, I is interception, S is the maximum canopy storage capacity, and W_{t-1} is previously stored water in the canopy. This formulation reproduces the threshold behavior of the process (i.e., small events can be almost entirely intercepted whereas larger events generate a higher throughfall proportion). However, it remains an aggregated approximation, because it does not explicitly resolve progressive canopy drainage, evaporation during the event, nor the dependence of interception on rainfall intensity and intra-event pauses, processes explicitly represented in physical interception models such as those of Gash and Rutter. In even more simplified water-balance formulations, a constant proportional loss is sometimes adopted:

$$P_{\text{eff}} = (1-\alpha)P_g \tag{2.9}$$

where α represents a fixed fraction of interception loss. This hypothesis violates two fundamental physical properties of the process: (i) interception exhibits threshold behavior, because events with $P_g < S$ are almost entirely intercepted, whereas large events generate an increasing throughfall proportion; and (ii) the P_g–I relationship depends on antecedent canopy state (W), which varies continuously with the sequence of events and atmospheric conditions (BY; GASH, 1979; MUZYLO et al., 2009). Consequently, the use of a constant α tends to underestimate interception in small and frequent events and to overestimate it in intense events, introducing systematic errors in the water balance.

Practical difficulties of parameterization in tropical ecosystems

Estimating S by regression requires a sufficiently large set of events with an initially dry canopy, which is rarely guaranteed in regions with frequent rainfall. The evaporation rate $\bar{E}$ is particularly difficult to measure directly: when estimated as a residual of the mass balance, it often assumes values higher than the potential evapotranspiration calculated by Penman–Monteith — a physically paradoxical result that has been attributed to the presence of additional energy sources not accounted for in the standard formulation (MUZYLO et al., 2009; LEVIA; CARLYLE-MOSES; TANAKA, 2011). These additional energy sources for wet-canopy evaporation include: heat stored in biomass and air (ΔQ_b), which sustains evaporation even when net radiation is low (such as during nighttime events or on cloudy days); advection of sensible heat from drier areas surrounding the forest fragment (Q_A), which can be particularly intense in forest remnants embedded within degraded and warmed landscapes; and downward sensible heat flux (H) that occurs when the overlying air is warmer than the saturated canopy, reversing the usual gradient and transferring energy to sustain evaporation (MUZYLO et al., 2009; LEVIA; CARLYLE-MOSES; TANAKA, 2011). Table 2.1 synthesizes these mechanisms and their impacts on interception.

Box 2.1 – Additional energy sources for wet-canopy evaporation not accounted for in the standard Penman–Monteith formulation.

Energy source	**Physical mechanism**	**Impact on interception**
Stored heat (ΔQ_b)	Heat stored in air and biomass (trunks and branches) released during and after the event	Maintains high evaporation rates even under low radiation or during nighttime events
Energy advection (Q_A)	Transport of sensible heat from unvegetated and warmed areas surrounding the forest fragment	Intensifies evaporation in isolated fragments within agricultural matrices; becomes dominant during prolonged dry periods
Sensible heat flux (H)	Overlying air warmer than the saturated canopy; reversal of the usual gradient	Transfers energy from the atmosphere to the canopy, sustaining evaporation even in the absence of solar radiation
Soil heat flux (G)	Heat flux from the subsurface warmed during the day	Smaller, but measurable, contribution to the total energy balance

Sources: Dijk et al. (2015), Michiles e Gielow (2008), Rodrigues et al. (2021a).
Fonte: The author (2026).

Empirical evidence of nonlinearity and threshold dependence

The literature reports evidence that interception in tropical forests exhibits nonlinear and threshold-dependent behavior that classical static-parameter models cannot adequately reproduce. By e Gash (1979) already noted, in the original formulation of the model, that the precipitation–interception relationship changes character at the transition points between phases (wetting, saturation, drying). More recent studies have shown that event intensity and antecedent canopy wetness determine when and how quickly these transitions occur — and that the event-to-event variability of these thresholds is not well represented by mean parameters calibrated over long series (MUZYLO et al., 2009; LEVIA; CARLYLE-MOSES; TANAKA, 2011). At Matinha

(UFLA), more than a decade of monitoring data reveal that the relationship between P_g and I is not monotonic and linear, but displays abrupt transitions associated with canopy saturation that vary with the time of year and with the recent history of events (RODRIGUES et al., 2021a; Junqueira Junior et al., 2019). Junqueira Junior et al. (2019) showed that the Gash model applied to this site produced systematic discrepancies between estimates and observations, particularly during seasonal transitions, when LAI variation is most rapid and canopy storage capacity S is changing continuously.

Specific gaps in interception modeling for semideciduous Atlantic Forest

The literature review allows four specific and critical gaps to be identified that motivate the approach of this thesis. The first is the **absence of joint spatiotemporal validation**: most studies validate interception models only temporally (at one or a few measurement points), without assessing the model's spatial generalization capacity to non-instrumented positions. The second is the **lack of models that explicitly represent seasonal parameter variation**: although the *Variable Storage Gash* model exists (MUZYLO et al., 2009), its systematic application in Atlantic Forest Montane FESM across complete annual cycles remains scarce. The third is the **scarcity of studies that evaluate antecedent canopy wetness at the event scale**: the system's "memory" is rarely quantified systematically, although it is recognized as a determining factor in nonlinear behavior. The fourth — and most directly relevant to this thesis — is the **nonexistence of operational models capable of predicting effective precipitation in real time from readily obtainable climatic variables**, without reliance on high-cost canopy sensors that are difficult to maintain in the field (MUZYLO et al., 2009; RODRIGUES et al., 2021a; Junqueira Junior et al., 2019). Machine learning offers a concrete response to these gaps. By learning the relationships between measurable climatic variables and effective precipitation from data rather than from physical equations, algorithms such as XGBoost and LSTM avoid the need to explicitly estimate S, $\bar{E}$, and $\bar{c}$, and can represent nonlinearity, threshold dependence, and hydrological memory without manual recalibration, provided the input data are of sufficient quality and validation respects the temporal structure of the series.

2.3 Integrity of Pluviometric Time Series: A Structural Problem in Tropical Hydrology

The regions where rainfall matters most, where cycles are most intense and most variable, are also the regions where long, gap-free, high-quality series are hardest to come by. In tropical hydrology this asymmetry is well documented but rarely treated as a methodological problem in its own right. This section treats it as such: data integrity is not a preprocessing step that precedes the real analysis, but a constraint that shapes what can be known and what cannot. Without reliable input data, ML algorithms learn noise rather than hydrological signal.

2.3.1 The Problem of Gaps in Monitoring Networks

Structural causes of failures in rainfall series

Gaps in pluviometric time series arise from a heterogeneous set of causes that can be organized into three categories (TEEGAVARAPU, 2012; MENNE et al., 2012). The first is **instrumental**: tipping-bucket rain gauges suffer clogging by organic debris, especially in forest environments with high deposition of leaves and plant material over the collector; the tipping mechanism may wear out or become miscalibrated with prolonged use; data acquisition and telemetry systems are subject to hardware failures, memory corruption, and communication interruptions. The second category is **operational**: problems in electrical power supply (power outages, discharged batteries, lightning damage to power sources) are frequent causes of missing data periods in remote stations, which in the Atlantic Forest are often installed in fragments of difficult access (RODRIGUES et al., 2021a). The third category is **anthropogenic**: vandalism and equipment theft — especially of solar panels and batteries — constitute a relevant cause of data loss in isolated stations in developing tropical countries, with estimated impacts of up to 15 % annual record loss in some Brazilian networks (Agência Nacional de Águas e Saneamento Básico, 2021). Quadro 2.2 synthesizes these causes, their expression patterns in the data, and available mitigation strategies.

Box 2.2 – Causes of failures in pluviometric series, data expression patterns, and mitigation strategies.

Category	Specific cause	Data pattern	Mitigation
Instrumental	Collector clogging	Unrecorded accumulation followed by sudden discharge	Periodic cleaning; use of protective screens
	Tipping-bucket wear	Gradual drift in values; rainfall recorded during dry weather	Semiannual calibration
	Datalogger failure	Continuous block of missing data (*NaN*)	Redundant storage; cloud backup
Operational	Power outage	Gaps coinciding with intense convective events	Solar panels with redundant battery banks
	Telecommunication failure	Data available locally but absent on the central server	Dual telemetry (cellular + satellite)
Anthropogenic	Vandalism/equipment theft	Abrupt long-duration gaps	Fencing; cameras; incident reporting
	Local environmental modification	Gradual drift due to changes in surrounding cover near the sensor	Monitoring of surroundings; field metadata

Sources: Teegavarapu (2012), Menne et al. (2012), Agência Nacional de Águas e Saneamento Básico (2021).

Fonte: The author (2026).

The specific impact of gaps in high-frequency series

Most quality-control studies in hydrology focus on daily or monthly data. However, modern operational applications — real-time flood warnings, hourly runoff forecasting, calibration of catchment models at sub-hourly resolution — require high-frequency time series (typically 5 to 60 minutes), where a 24-hour gap may correspond to the absence of 288 records (TEEGAVARAPU, 2012). At this scale, the impact of failures is qualitatively different: an intense and short-duration convective event (30–60 min), responsible for 80 % of the runoff of a flood event, may be entirely missed if it occurs during a datalogger gap. The loss is qualitative as much as quantitative: a model trained on gap-filled data will never see the extreme events that matter most — the short, intense bursts that occur when atmospheric stress is highest, which are also the conditions most likely to trigger station failures. This correlation between instrumental failure and hydrologically relevant events is particularly severe because it introduces a systematic exclusion bias of extremes in reconstructed series (MENNE et al., 2012; KIM; PACHEPSKY, 2010).

Scarcity of meteorological stations in the tropics

The problem of missing data is exacerbated by the low density of meteorological stations in strategic tropical regions. The World Meteorological Organization (WMO) recommends a minimum density of one station per 600 km² in mountainous regions with high spatial variability of precipitation (World Meteorological Organization, 2008). In Brazil, analysis of the National Water Agency (ANA) rainfall network reveals that strategic basins of the Atlantic Forest and the Amazon often exhibit densities far below this standard, with vast areas under insufficient monitoring (Agência Nacional de Águas e Saneamento Básico, 2021). In addition to insufficient density, there is a severe problem of temporal discontinuity: the Brazilian rainfall network lost more than 2,000 active stations between 1990 and 2010, resulting in series with prolonged inactivity periods that drastically reduce the effective length of historical records available for model calibration (Agência Nacional de Águas e Saneamento Básico, 2021).

Hydrological consequences of poor data quality

The consequences of using incomplete or low-quality series in hydrological modeling are systemic and difficult to detect *a posteriori*: they propagate silently from data to model, from model to forecast, and from forecast to management decision (TEEGAVARAPU, 2012). In hydrological model calibration, series with high proportions of gaps (>10 % of records) produce parameters calibrated on a sample that is not representative of true climatic variability; the resulting model will perform well under the underrepresented calibration conditions and degrade under the extreme conditions that matter most from a risk-management perspective (KIM; PACHEPSKY, 2010). In the estimation of derived hydrological indices (coefficient of variation, extreme percentiles, frequency of events above thresholds), missing data introduce biases that are difficult to detect without careful analysis of the temporal structure of the gaps.

2.3.2 Instrumental Errors and Quality Control in Rainfall Data

Conceptual foundations of automated quality control

Modern quality control (QC) systems in hydrometeorological networks have evolved from case-by-case manual review toward integrated, layered frameworks. The most influential architecture in this domain, developed for the Global Historical Climatology Network–Daily (GHCN-Daily) database, organizes QC into three sequential levels of increasing analytical complexity (DURRE et al., 2010). The first level encompasses absolute range tests: is the recorded value within the physically possible limits for the region and time of year? The second level applies internal consistency tests: is the value coherent with other records from the same station during the same event? The third level implements spatial consistency tests: is the value coherent with simultaneous records at neighboring stations, given the expected spatial correlation structure for that event type? This stratified architecture is conceptually important because it establishes a logical hierarchy — filtering the most obvious errors (physically impossible values) before applying more sophisticated algorithms that could confuse subtle errors with valid climate signal. Applying computationally expensive algorithms to data that contains trivially flagged values wastes processing resources and may propagate errors into subsequent detection stages (DURRE et al., 2010)

A conceptual distinction central to the design of any hydrological QC system was formalized by (HAWKINS, 1980): outliers in a dataset may be of two radically different types. *Error outliers* are values produced by instrumental failures, transcription errors, or sensor drift; they must be removed because they introduce non-informative noise. *Innovative outliers* are genuinely extreme values that reflect exceptional hydrometeorological events — intense convective storms, South Atlantic Convergence Zone (SACZ) episodes, orographically enhanced rainfall — and must be preserved because they carry the greatest hydrological information content for flood modeling, interception dynamics, and extreme event analysis (HAWKINS, 1980). This distinction has direct practical consequences: a QC algorithm that systematically removes the highest values in the distribution is not "cleaning" the series; it is amputating its capacity to inform about high-impact events. Preserving innovative outliers while removing error outliers is the central epistemic challenge of automated QC in tropical pluviometric series, where the asymmetric, zero-inflated distribution of precipitation makes standard parametric tests (e.g., three standard deviations above the mean) structurally inappropriate as the sole criterion.

————————————————————————-

Typology of errors in tipping-bucket rain gauges

The tipping-bucket rain gauge is the dominant instrument in automatic hydrological monitoring networks worldwide. Its principle is simple: each time a standardized volume (typically 0.1 mm or 0.2 mm) accumulates, the bucket tips and records an electronic pulse, allowing calculation of rainfall intensity and totals for any desired time interval. However, this simple principle generates a specific typology of errors that requires careful attention. The first type is the **point spurious value**: a pulse recorded in the absence of real precipitation, which may result from mechanical vibration (heavy traffic or strong winds shaking the bucket), the entry of insects or organic material into the mechanism, or electromagnetic interference in the datalogger. The result is rainfall records during completely dry conditions, easily identifiable through comparison with neighboring stations or satellite data (MENNE et al., 2012). The second type is **accumulation followed by sudden discharge**: when the collector is partially obstructed, internally retained water is suddenly released after a period of unrecorded accumulation. The resulting pattern is characteristic: a long sequence of zeros followed by an abnormally high value in a single record, which may exceed historical maxima of the series and be erroneously identified as a real extreme event (MENNE et al., 2012; TEEGAVARAPU, 2012). This pattern is particularly insidious because it mimics short-duration intense convective events, which are exactly the records one cannot afford to discard. The third type is **false inflated zeros**: records of zero precipitation during periods when the station should clearly have recorded rainfall, detectable through comparison with neighboring stations or satellite products. These may result from complete clogging of the collector, failure in the electronic pulse-reading circuit, or unintended data discard by the datalogger firmware (FUNK et al., 2015).

Specific challenges of outlier detection in precipitation

Outlier detection in pluviometric series is substantively more complex than in temperature or discharge series, for reasons that require explicit clarification (MENNE et al., 2012). First, precipitation is a variable with a **strongly right-skewed distribution** and a point mass at zero: in semi-arid regions or during dry seasons, 90 % or more of records may be zero, and non-zero values follow heavy-tailed distributions (gamma, lognormal, generalized Pareto), where the probability of very high values is non-trivial (MENNE et al., 2012). A naive threshold of "three standard deviations above the mean" applied to this distribution would automatically eliminate genuine extreme events — the records most needed for calibrating flood-warning models. Second, there is the problem of **spatiotemporal variability of the true signal**: a mesoscale convective event may produce 80 mm in 30 minutes at one location and only 5 mm at a distance of 3 km. Comparison with neighboring stations — the gold standard of quality control in temperature networks — may therefore erroneously identify as an outlier an event that is simultaneously extreme and real (KIM et al., 2024). This limitation is even more pronounced in regions of complex topography, where orographic effects amplify precipitation at sub-kilometric scales.

2.3.3 Traditional Gap-Filling Approaches and Their Limitations

Historical evolution of gap-filling methods

The evolution of gap-filling methods for pluviometric series tracks the computational tools and epistemological priorities of each era. Each generation of methods exposed the limitations of the one before it, and the case for ML-based reconstruction rests, in part, on the accumulated evidence of where those earlier approaches fall short.

Pre-computational period (before 1970). The earliest gap-filling approaches were deterministic and based on spatial analogy. The arithmetic mean and the *normal ratio* method — which weights neighboring station values by the ratio of their mean annual precipitations — represented the operational standard documented by (PAULHUS; KOHLER, 1952) for the United States Weather Bureau. Their virtue was implementability without computing infrastructure; their fundamental limitation was the incapacity to incorporate structured spatial correlations or to quantify the uncertainty of the reconstructed value.

The double mass curve method (SEARCY; HARDISON, 1960) deserves particular attention because it was for decades the standard procedure for detecting inconsistencies in long series. The method consists of plotting accumulated values of one station against accumulated values of one or more reference stations: a break in slope signals an artificial — non-climatic — change in the record, typically attributable to a gauge relocation, observer change, or land-use modification in the station surroundings. Its critical limitation is that it is visual and subjective, and cannot formally distinguish between an artificial bias shift and a genuine regional climate change signal (SEARCY; HARDISON, 1960).

Era of multivariate statistics (1970–2000). The availability of computers enabled the incorporation of more sophisticated statistical methods. Multiple linear regression with neighboring stations as predictors became widespread (EISCHEID et al., 1995). Its natural extension was principal component analysis (PCA) as a dimensionality reduction tool prior to gap-filling, which improved performance when many correlated stations were available.

A conceptually significant advance was the development of geostatistical methods, particularly **kriging** in its ordinary, universal, and external drift variants (MATHERON, 1963; JOURNEL; HUIJBREGTS, 1978). Kriging offers a property that no previous method possessed: it provides both an estimated value and an estimation variance, allowing quantification of spatial uncertainty. Formally, ordinary kriging estimates the value at an ungauged location $\mathbf{x}_0$ as:

$$\hat{P}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i P(\mathbf{x}_i), \quad \text{subject to} \quad \sum_{i=1}^{N} \lambda_i = 1 \tag{2.10}$$

where the weights λ_i are obtained by solving the kriging system that minimizes the estimation variance σ_K^2, derived from the experimental variogram $\gamma(h)$ — the variance of differences between values separated by lag distance h (JOURNEL; HUIJBREGTS, 1978). Kriging's application to precipitation gap-filling, however, requires a minimum network density that many tropical basins — including broad regions of the Atlantic Forest — do not meet, and assumes variogram stationarity (that the spatial correlation structure depends only on distance, not location), which is systematically violated by convective tropical precipitation whose spatial structure changes radically between stratiform and convective events (LY; CHARLES; DEGRÉ, 2011).

A third conceptual development relevant to the context of this thesis was *multiple imputation* (RUBIN, 1987), originating in biomedical statistics and gradually adopted in climatology. Unlike single imputation — which produces one point estimate as if it were observed — multiple imputation generates M complete plausible datasets, each with different values for missing data, thereby capturing the uncertainty of the imputation process and allowing its propagation into subsequent analyses. Its extension to nonparametric prediction, the **missForest** algorithm (STEKHOVEN; BÜHLMANN, 2012), applies Random Forest iteratively to impute each variable from all others, representing the direct conceptual bridge between classical gap-filling and the machine learning approaches developed in Section 2.4.

Structural limitations of classical methods. The preceding review reveals four dimensions along which all classical methods share fundamental inadequacies in the context of tropical convective precipitation. First, *inability to capture nonlinearities*: the relationship between precipitation at neighboring stations is highly nonstationary, conditioned on the synoptic pattern and the type of precipitating system.Second, *limited scalability*: when the network has dozens or hundreds of stations with different record periods, classical methods require case-by-case decisions that are difficult to automate at the scale needed for operational gap-filling. Third, *inability to integrate heterogeneous covariates*: satellite data, atmospheric reanalysis fields, and physiographic indices are difficult to incorporate formally without resorting to ad hoc formulations. Fourth, *absence of multi-objective selection criteria*: classical methods typically optimize a single metric — RMSE — which yields estimates that are accurate on average but poor for extreme events, the very events that matter most for modeling peak flows and canopy saturation dynamics (Section 2.3.3).

Conventional methods and their foundations

Gap-filling in pluviometric series has a long history in operational hydrometeorology. The most widely used methods, in increasing order of statistical sophistication, are (TEEGAVARAPU, 2012). The **arithmetic mean** (AM) is the simplest and most commonly adopted method in operational networks: the missing value is estimated as the mean of simultaneous values from N neighboring stations. It is robust and easy to implement, but completely ignores differential correlation between stations and assigns equal weight to all, regardless of distance or climatic similarity. The **inverse distance weighting** (IDW) method assigns weights inversely proportional to the distance d_i between the station with missing data and each neighboring station:

$$\hat{P} = \frac{\sum_{i=1}^{N} P_i/d_i^{\beta}}{\sum_{i=1}^{N} 1/d_i^{\beta}} \tag{2.11}$$

where the exponent β (typically $\beta = 2$) controls how rapidly the weight decreases with distance (TEEGAVARAPU, 2012). IDW represents an advance over the arithmetic mean by introducing geographic proximity as a proxy for climatic similarity, but it ignores topography, prevailing wind patterns, and orographic barriers that modulate the spatial variability of tropical precipitation. **Multiple linear regression** (OLS) fits a first-order model between the series with gaps (Y) and the series from neighboring stations ($X_1, \ldots, X_k$):

$$\hat{Y} = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon \tag{2.12}$$

The coefficients β_j are estimated by ordinary least squares (OLS) over periods of simultaneous data. The method captures linear correlations among stations, but implicitly assumes that the relationship is stationary (coefficients remain constant over time), which is rarely true for tropical convective precipitation. **SARIMA** (Seasonal AutoRegressive Integrated Moving Average) models the temporal structure of the individual series (seasonality and lag dependence) to predict missing values based on the series' own historical dynamics, without relying on neighboring stations. The complete formulation of a SARIMA$(p,d,q) \times (P,D,Q)_s$ model is:

$$\Phi_P(B^s)\phi_p(B)\nabla^d\nabla_s^D Y_t = \Theta_Q(B^s)\theta_q(B)\,\varepsilon_t \tag{2.13}$$

where B is the lag operator, ∇^d the difference of order d, ∇_s^D the seasonal difference of period s, and ε_t white noise (BOX; JENKINS, 1970). Although SARIMA captures the temporal structure of the series, its linearity assumption renders it inadequate for tropical convective precipitation.

Violations of assumptions and fundamental limitations

All of the above methods share statistical assumptions that are systematically violated by tropical convective precipitation (TEEGAVARAPU, 2012; KIM; PACHEPSKY, 2010). The assumption of **stationarity** implies that estimated parameters (regression coefficients, ARIMA parameters) remain valid over time; in tropical precipitation, seasonal variations in the South Atlantic Convergence Zone (SACZ), frontal systems, and sea surface temperature substantially alter the correlation structure among stations throughout the year, invalidating models fitted to complete annual series. The assumption of **Gaussianity** or **distributional symmetry** is violated by the asymmetric and heavy-tailed distribution of precipitation: least-squares-based methods minimize mean squared error, which is dominated by performance on median values; extreme values — rare but hydrologically critical events — receive insufficient weight in the estimation process and are systematically underestimated (KIM et al., 2024). The most serious consequence of this underestimation of extremes is practical: a series reconstructed by conventional methods will exhibit attenuated upper percentiles (P_{95}, P_{99}) relative to the original series, equivalent to a shrinkage of the distribution tail. A hydrological model trained on such a reconstructed series will learn that extreme events are less frequent than they actually are and will produce systematically conservative predictions during periods of greatest risk (TEEGAVARAPU, 2012).

Metrics for evaluating reconstruction quality

The evaluation of reconstructed series quality cannot be limited to the root mean square error (RMSE), which penalizes only absolute deviations without assessing whether the reconstruction preserves the distributive properties of the original series. The literature establishes a complementary set of metrics (GUPTA et al., 2009; KNOBEN; FREER; WOODS, 2019):

- **RMSE**: $\sqrt{n^{-1}\sum_t (Y_t - \hat{Y}_t)^2}$ — measures mean absolute error in variable units; sensitive to extreme events due to quadratic terms.

- **KGE** (developed in Section 2.6): decomposes error into correlation, bias, and variability, being more informative than RMSE in hydrological contexts.
- **Extreme preservation**: comparison of percentiles P_{95} and P_{99} between original and reconstructed series; a method that systematically underestimates extremes will produce ratios $P_{95}^{\text{rec}}/P_{95}^{\text{obs}} < 1$.
- **Kolmogorov–Smirnov test** (KS): non-parametric test of distributional equality between reconstructed and reference series; rejection of H_0 at 5 % indicates that reconstruction significantly altered the empirical distribution of the variable (KIM; PACHEPSKY, 2010).
- **Variability Maintenance Index** (VMI): $\text{VMI} = \sigma_{\text{rec}}/\sigma_{\text{obs}}$; a VMI value below 1.0 indicates excessive smoothing of natural variability.

2.3.4 Satellite and Reanalysis Products as Alternative Sources of Precipitation

Conceptual taxonomy: three physical principles of estimation

Near-global precipitation datasets are constructed from three distinct physical principles, and understanding them matters for how their errors should be interpreted in any local application.

The first principle is **atmospheric reanalysis**: a numerical weather prediction model is forced with assimilated historical observations — radiosonde profiles, surface pressure, satellite radiances, ship reports — via four-dimensional variational data assimilation (4D-Var) and run forward in time to produce a physically consistent four-dimensional estimate of the atmospheric state, from which precipitation is diagnosed as a model-derived output (DEE et al., 2011; HERSBACH et al., 2020). Reanalyses do not measure precipitation directly; they compute it as the product of simulated atmospheric dynamics, specifically the interaction between the model's dynamical core (which advects moisture) and its convection parameterization scheme (which determines when and how much condensation falls as rain). Reanalysis precipitation is therefore only as reliable as the model's representation of convective processes — parameterizations that are acknowledged to be a primary source of uncertainty in numerical weather prediction, particularly in humid tropical environments where deep convection is the dominant rainfall mechanism (DEE et al., 2011).

The second principle is **satellite remote sensing**, which retrieves precipitation indirectly from observed electromagnetic signals. Two approaches dominate: thermal infrared (TIR) retrieval uses geostationary satellite brightness temperatures at cloud tops as a proxy for convective intensity, offering high temporal resolution but relying on the uncertain, spatially variable relationship between cloud-top temperature and surface rainfall (HUFFMAN et al., 2020). Passive microwave (PMW) retrieval detects the scattering and emission signatures of hydrometeors in the atmospheric column from low-Earth-orbit sensors, providing physically more direct measurements but limited by coarser temporal sampling (KIDD; LEVIZZANI, 2011). In montane tropical environments, both approaches face a fundamental limitation: precipitation often originates from warm clouds (cloud-top temperatures above the -38 °C TIR threshold) driven by orographic lifting, a mechanism that neither TIR nor PMW algorithms reliably detect.

The third principle is **gauge-satellite merging**, which combines the spatial coverage of satellite or reanalysis fields with the point accuracy of surface rain gauge observations through statistical interpolation, regression, or machine-learning calibration (FUNK et al., 2015; VILA et al., 2009). The resulting product inherits the strengths and weaknesses of both components: spatial coverage from the remote sensing field, and local accuracy from the gauge network — but also the network's sparse and geographically uneven distribution, a particularly critical limitation in tropical countries where gauge density is low and biased toward lowland urban areas (BECK et al., 2019).

ERA5 and ERA5-Land: the ECMWF reanalysis family

ERA5 is the fifth-generation global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), covering the period from 1940 to the present at approximately 31 km horizontal resolution with 137 vertical levels and hourly output (HERSBACH et al., 2020). ERA5 represents a substantial advance over its predecessor ERA-Interim in terms of data assimilation methodology (4D-Var versus 3D-Var), horizontal resolution (31 versus 79 km), and the volume and diversity of assimilated observations. ERA5-Land is a companion product that replaces the land-surface component with a higher-resolution run ($\approx$9 km) of the ECMWF land-surface model forced by ERA5 atmospheric fields, improving the representation of soil moisture and its feedbacks on near-surface thermodynamics (MUÑOZ-SABATER et al., 2021).

The physical mechanism by which ERA5 generates precipitation is critical to understanding its error structure. The 4D-Var assimilation directly constrains wind, temperature, and specific humidity, but does not assimilate precipitation observations. Precipitation emerges entirely from the model's convective and microphysics parameterization, inheriting the biases of those schemes. In tropical environments, convective parameterization systematically underestimates the intensity of deep convective events because the parameterization represents sub-grid-scale convection as a smoothed statistical process, unable to resolve the spatial concentration of precipitation in the cores of mesoscale convective systems (AL-RAWAS et al., 2025; SALAZAR et al., 2025). In Brazil, ERA5 has been documented to systematically underestimate hourly precipitation above the 90th percentile (ZANDLER et al., 2019). Additionally, the 31 km grid smooths topographic gradients that drive orographic precipitation enhancement: in the Serra da Mantiqueira and Serra do Mar regions, this spatial averaging blends into a single pixel rainfall regimes that may differ by more than 500 mm year^{-1} between windward and leeward slopes (SUNILKUMAR et al., 2019).

Despite these limitations, ERA5 possesses a property of great practical value for precipitation reconstruction: *temporal phase fidelity*. Because ERA5 assimilates synoptic-scale atmospheric observations that determine the large-scale circulation patterns governing rainfall timing, it tends to correctly identify *when* rainfall events occur — even when underestimating their intensity — better than products based purely on satellite imagery. This property makes ERA5 a particularly valuable predictor for machine-learning-based reconstruction of event timing, even after intensity bias correction.

CHIRPS v2.0: gauge-corrected infrared precipitation

The Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) version 2.0 is a quasi-global daily precipitation product at approximately 5.5 km resolution covering 1981 to the present (FUNK et al., 2015). Its construction integrates three components: a monthly climatological precipitation surface (CHPclim) derived from gauge observations; a thermal infrared precipitation estimate from cold cloud duration (CCD) measured by geostationary satellites; and a network of quality-controlled daily rain gauge observations that bias-correct the satellite-IR estimates.

CHIRPS's primary physical limitation derives from the cold cloud duration algorithm. The CCD approach identifies precipitating systems by thresholding cloud-top temperatures below 235 K, based on the assumption that cold cloud tops co-occur with deep convection and surface precipitation. This assumption fails in two important ways in montane Atlantic Forest environments. First, warm rain processes — where precipitation forms and falls from clouds whose tops remain above the 235 K threshold — are entirely undetected by the algorithm. Orographic lifting in the Serra da Mantiqueira frequently produces precipitation from shallow convective or stratiform systems with relatively warm cloud tops, meaning CHIRPS systematically misses this orographic component (PAREDES-TREJO; BARBOSA; Lakshmi Kumar, 2017; ANDRADE et al., 2024). Second, the relationship between cloud-top temperature and surface precipitation intensity is highly variable: a deep convective cell can produce the same cold cloud-top signal whether it generates 5 mm or 50 mm of surface precipitation, depending on its microphysical efficiency, which TIR imagery cannot resolve.

The gauge correction step partially mitigates these biases at monthly time scales, but its effectiveness is limited by the density of the gauge network: in Brazil, gauge coverage is uneven, with greater density in major urban areas and lower density in forested and mountainous regions. Where gauge density is insufficient, the satellite-IR field dominates and its systematic limitations propagate into the corrected product. This generates the well-documented *drizzle effect*: by assigning small but non-zero precipitation amounts to days with weak positive IR signal, CHIRPS overestimates the frequency of low-intensity events ($<1\,\text{mm}\,\text{day}^{-1}$) while simultaneously underestimating peak event intensities (BAGILIKO et al., 2025). In distributional terms, CHIRPS tends to overestimate the lower quantiles and underestimate the upper quantiles of the precipitation distribution — a systematic distortion directly impacting extreme event analysis.

PERSIANN-CDR: neural network precipitation from geostationary infrared

The Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks – Climate Data Record (PERSIANN-CDR) is a daily global precipitation product at approximately 27 km resolution covering 1983 to the present (ASHOURI et al., 2015). It uses an artificial neural network trained on TRMM Precipitation Radar data to estimate precipitation from geostationary TIR brightness temperatures, subsequently bias-adjusted to GPCP monthly estimates. This monthly adjustment to GPCP gives PERSIANN-CDR good performance in reproducing seasonal and annual totals at large spatial scales — the primary purpose of its designation as a climate data record.

At the event and sub-daily scale, PERSIANN-CDR exhibits error characteristics that differ from CHIRPS in important ways relevant to hydrological applications. Because its neural network was trained to maximize precipitation prediction from TIR signals, it performs well when cold cloud-top temperatures reliably co-occur with heavy precipitation — the canonical signature of deep tropical convection. In these conditions, the network tends to *overestimate* precipitation intensity because it cannot distinguish between deep convective cells with high precipitation efficiency and those with low efficiency (MOKHTARI; SHARAFATI; RAZIEI, 2022; FARAMARZZADEH et al., 2023). Conversely, for stratiform or moderate-intensity events where cloud-top temperatures are insufficiently cold to trigger the network's high-precipitation response, PERSIANN-CDR tends to *underestimate*. The result is a product with high variance in the upper tail of the distribution — more extreme values than observed — which is problematic for machine learning applications because models may learn spurious precipitation patterns from training features, generating overconfident predictions during convective events.

Complementary bias structures: rationale for joint use

The physical characterization above reveals that the three products have bias structures that are distinct and in several respects *complementary*. ERA5 underestimates convective intensity but correctly captures event timing and synoptic-scale patterns. CHIRPS underestimates orographic and warm-rain precipitation and distributes total rainfall over too many low-intensity days. PERSIANN-CDR overestimates the upper tail for deep convection while underestimating moderate stratiform events. This complementarity implies that no single product is an adequate predictor for machine-learning reconstruction in montane Atlantic Forest environments: joint use of multiple products after individual bias correction allows the model to exploit complementary strengths while ensemble averaging partially cancels individual systematic errors — a bias-variance trade-off at the data level that parallels the ensemble methods applied at the model level (SHEN, 2018; WILLARD et al., 2022).

Bias correction methods: from scaling to distributional adjustment

The historical development of bias correction methods reflects the progressive recognition that systematic errors in precipitation products affect not only the mean but the entire distributional shape. The earliest operational approach was the *scaling factor* method (also called delta mapping), which multiplies model values by the ratio of observed to model mean precipitation (PANOFSKY; BRIER, 1958). Its critical limitation is that it corrects only the mean, leaving the distributional shape intact: in tropical convective precipitation, where the upper percentiles are the most distorted in reanalysis products, a mean-only correction is inadequate.

Quantile Mapping (QM) represents the conceptual advance that addresses this limitation: the cumulative distribution function (CDF) of the corrected product is forced to match the CDF of the reference observations by mapping quantile to quantile (GUDMUNDSSON et al., 2012). Formally:

$$P_{\text{corr}} = F_{\text{obs}}^{-1}[F_{\text{model}}(P_{\text{model}})] \tag{2.14}$$

where F_{obs}^{-1} is the quantile function (inverse CDF) of the observed distribution and F_{model} is the CDF of the product to be corrected (GUDMUNDSSON et al., 2012). The mapping may be parametric (assuming gamma or lognormal distributions) or non-parametric (direct interpolation in the empirical distribution). The principal advantage of QM over the scaling factor is that it corrects the entire distribution shape — including the tails — rather than only the mean. For tropical convective precipitation, whose extreme percentiles are the most distorted, this property is decisive (GUDMUNDSSON et al., 2012).

Modern variants of QM address the limitation that when the calibration period includes temporal trends — such as changes in extreme rainfall frequency associated with climate change — standard QM can distort those trends by mapping them toward the historical observed distribution. The Detrended Quantile Mapping (DQM) method (CANNON; SOBIE; MURDOCK, 2015) resolves this by removing the trend from the model field before applying the quantile mapping and reintroducing it afterward, preserving both the historical distribution shape and long-term trends.

An epistemological caveat is essential here to avoid over-interpretation of bias correction results: bias correction is not a downscaling method and cannot create information absent from the predictor (MARAUN, 2016). If ERA5 does not capture a physical precipitation pattern —

for example, the micro-scale orographic effect of the Serra da Mantiqueira on the Matinha site — no bias correction method will recover that signal. Corrected products are therefore tools for supporting series gap-filling during periods without surface observations, not substitutes for a surface pluviometric network at local scale. This distinction is conceptually essential for interpreting the results of Article 1 with appropriate methodological caution.

Box 2.3 – Main global precipitation products used in tropical hydrology and their characteristics.

Product	**Estimation principle**	**Spatial resolution**	**Temporal resolution**	**Coverage**
ERA5 Hersbach et al. (2020)	Global atmospheric reanalysis with data assimilation (4D-Var); precipitation is a model-derived output	≈ 31 km	1 hour	1940–present
CHIRPS Funk et al. (2015)	Rain gauge interpolation with TRMM satellite prior field; non-parametric regression on the grid	≈ 5.5 km	Daily / monthly	1981–present
PERSIANN-CDR Ashouri et al. (2015)	Neural network driven by infrared imagery from geostationary satellites; calibrated with GPCP	≈ 27 km	Daily	1983–present
MERGE Vila et al. (2009)	Fusion of GOES satellite data with INMET/ANA rain gauges for South America; interpolation by Kriging	≈ 4 km	Daily	Regional (SA)
GPM IMERG Huffman et al. (2020)	Fusion of passive microwave, infrared, and rain gauge data; reference product of TRMM/GPM	≈ 10 km	30 min	2000–present

Fonte: The author (2026).

Systematic biases and their physical causes

Despite their unprecedented spatial and temporal coverage, these products exhibit systematic biases that compromise their direct use — without correction — in local high-resolution applications (SUNILKUMAR et al., 2019; TIAN et al., 2007).

The **drizzle effect** is the most documented bias in products based on thermal infrared imagery: the algorithm, by identifying cold cloud tops as likely precipitation, tends to register light drizzle (<1 mm h^{-1}) too frequently, inflating the number of low-intensity events and, consequently, monthly totals in regions with persistent stratiform cloudiness (TIAN et al., 2007).

In terms of exceedance frequency, this manifests as overestimation in the lower percentiles of the distribution and underestimation in the upper percentiles.

The **underestimation of convective extremes** occurs for the opposite reason: very high-intensity convective events (exceeding 50 mm h^{-1}) are undercaptured by reanalysis products because convection parameterization in numerical models does not resolve the spatial scale of the most intense convective cores. In ERA5, this bias has been specifically documented for hourly precipitation in Brazil, with systematic underestimation of events above the 90th percentile (ZANDLER et al., 2019).

The problem of **insufficient spatial resolution** is particularly severe in regions of complex topography: no global product adequately captures the contrast between windward and leeward slopes of mountain ranges that characterizes the Atlantic Forest in the Serra da Mantiqueira and Serra do Mar regions. The 31 km grid of ERA5, for example, blends into a single pixel rainfall regimes that may differ by more than 500 mm $year^{-1}$ (SUNILKUMAR et al., 2019).

Fusion of satellite data with *in situ* observations

The integration of remote sensing products with surface rain gauges through statistical fusion or kriging with external drift (KED) substantially improves the representation of local variability in tropical precipitation. The MERGE product, developed by INPE for South America, applies this principle: the initial GOES field is corrected by Kriging interpolation using INMET and ANA rain gauges as conditioning values, producing a spatial field consistent with available surface observations (VILA et al., 2009). In independent validations for Brazil, MERGE outperformed raw ERA5 in representing daily precipitation variability in medium-sized Atlantic Forest catchments (TIAN et al., 2007).

Quantile Mapping and its statistical foundation

Quantile Mapping (QM) is the most rigorous and widely adopted statistical bias-correction method for adjusting precipitation products to local regimes (PANOFSKY; BRIER, 1958; GUDMUNDSSON et al., 2012). Its principle is straightforward: the cumulative distribution function (CDF) of the corrected product is forced to match the CDF of the reference observations by mapping quantile to quantile. Formally:

$$P_{\text{corr}} = F_{\text{obs}}^{-1}[F_{\text{modelo}}(P_{\text{modelo}})] \tag{2.15}$$

where F_{obs}^{-1} is the quantile function (inverse CDF) of the observed distribution and F_{modelo} is the CDF of the product to be corrected (GUDMUNDSSON et al., 2012). The mapping may be parametric (assuming gamma or lognormal distributions for F_{obs} and F_{modelo}, respectively) or non-parametric (direct interpolation in the empirical distribution). The main advantage of QM over simple scaling-factor corrections (bias ratio) is that it corrects the entire distribution shape — including the tails — and not only the mean or total. For tropical convective precipitation, whose extreme percentiles are the most distorted in reanalysis products, this property is decisive (GUDMUNDSSON et al., 2012).

2.4 Machine Learning in Hydrology: Foundations and Paradigms

2.4.1 The Data-Driven Hydrology Paradigm

Process-based hydrology has as its cornerstone the conviction that the behavior of a watershed can be deduced from conservation principles (mass, energy, momentum) and functional parameterizations that describe dominant processes. Models such as TOPMODEL, VIC, SWAT, and HBV are built upon this tradition: each parameter has, at least in principle, a physical interpretation, and the model respects certain properties that data alone do not guarantee (such as long-term water balance closure) (BEVEN, 2012). The fragility of this approach lies in the set of structural decisions that do not emerge from the data but are imposed *a priori* by the modeler: which routing function to use, which evapotranspiration parameterization to adopt, how to represent spatial heterogeneity of soil and vegetation cover. These decisions introduce structural errors that calibration cannot eliminate — only redistribute among parameters (BEVEN, 2012; SHEN, 2018).

Data-driven hydrology departs from the opposite pole: the model structure is not prescribed by the modeler but learned from data. There are no transport equations, no assumption of linearity, no predefined parametric loss function. The algorithm learns, from observed input–output pairs, a representation that maps predictor variables (precipitation, temperature, radiation, soil moisture) onto the response variable (runoff, effective precipitation, evapotranspiration) (SHEN, 2018). The epistemological foundations of this approach are **agnostic with respect to the internal mechanics** of the system: what matters is the fidelity of the input–output mapping, not the interpretability of each individual parameter.

This implies assumptions distinct from process-based models. Data-driven hydrology assumes that the system to be modeled exhibits a regularizable pattern that can be learned from the available data; that the training data are representative of the relevant state space; and that statistical stationarity is sufficiently valid during the training period to allow generalization to the validation period (SHEN, 2018). When any of these conditions fail — especially the representativeness of training data relative to more extreme future climatic conditions — the ML model loses predictive capacity outside its training domain, a problem known as the **extrapolation gap** (WILLARD et al., 2022).

The Physics-aware Machine Learning (PaML) paradigm

The most sophisticated response to the *extrapolation gap* is the Physics-aware Machine Learning (PaML) paradigm, whose methods Willard et al. (2022) organize into four non-mutually exclusive classes based on the mechanism by which physical knowledge is integrated into the model (Figure 2.1).

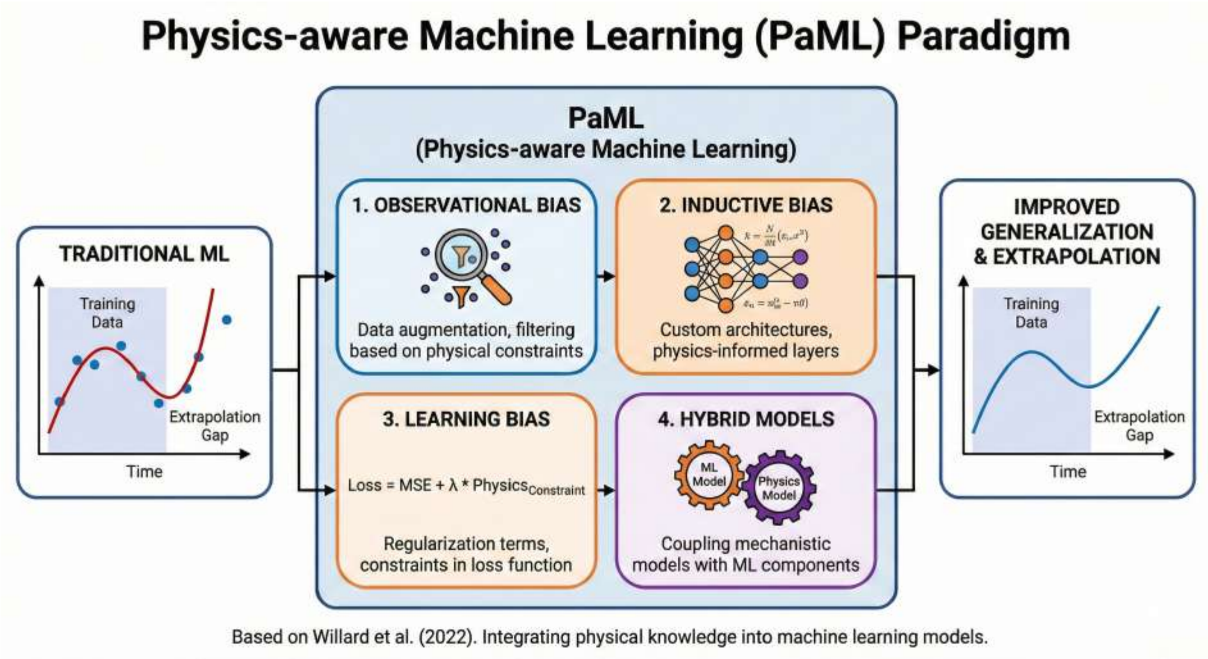
In **physics-guided initialization**, the model is pre-trained on synthetic data generated by process-based model simulations before being fine-tuned with real observations. Physical knowledge thus acts as an informed initial state that guides training convergence and reduces the requirement for observational data (WILLARD et al., 2022).

In **physics-guided loss functions**, training is regularized by additional terms in the objective function that penalize violations of known physical laws — such as mass conservation or energy balance. This constraint is considered *weak* because it steers, but does not guarantee, the physical consistency of predictions (WILLARD et al., 2022).

In **physics-guided architecture design**, certain physical properties are satisfied by construction within the model architecture — such as variable positivity, monotonicity relationships, or symmetry invariances — so that physically inconsistent predictions become structurally impossible (WILLARD et al., 2022).

In **hybrid physics-ML modeling**, a process-based model and an ML model operate in a coupled manner. Willard et al. (2022) identify four variants of this class: (i) residual modeling,

Figure 2.1 – The Physics-aware Machine Learning (PaML) paradigm. Methods are organized into four non-mutually exclusive classes based on the mechanism of integrating physical knowledge: (1) Observational Bias (Data), (2) Inductive Bias (Architecture), (3) Learning Bias (Loss Function), and (4) Hybrid Models. Adapted from Willard et al. (2022).



Fonte: The author (2026).

in which ML learns the systematic errors of the physical model; (ii) physical model output as ML input, in which process-model simulations feed the algorithms as covariates; (iii) replacement of physical model components by ML; and (iv) weighted combination of physical and data-driven predictions.

The most influential evidence that ML can converge toward physically coherent representations without explicit imposition came from Kratzert et al. (2019a), who demonstrated that an LSTM (*Long Short-Term Memory*) network trained simultaneously across hundreds of catchments using only climatic forcing data develops internal states interpretable as snow storage and soil moisture — even when these variables are not explicitly provided as model inputs. This result suggests that, given sufficient data and multi-catchment training, temporal memory structures can emerge spontaneously as implicitly physical representations.

Critiques of ML in hydrology and responses from the literature

The most robust critiques of ML in hydrology are organized around three central arguments (BEVEN, 2012; SHEN, 2018).

The first is the **lack of physical interpretability**: a well-calibrated ML model may work for the wrong reasons — learning spurious correlations that collapse outside the training window — with no guarantee that the learned patterns correspond to real causal mechanisms. The literature's response has been the adoption of explainability tools, particularly SHAP values (*SHapley Additive exPlanations*), which allow quantification of the contribution of each input variable to each individual prediction, rendering models physically auditable (elaborated in Section 5.3.5).

The second argument concerns **dependence on large data volumes**: algorithms such as LSTM typically require years of high-frequency series to learn reliable temporal patterns, potentially rendering them unsuitable for regions with short or fragmented records. The response lies in transfer learning strategies and multi-catchment training, which allow models trained on data-rich systems to be adapted to data-scarce contexts (KRATZERT et al., 2019b).

The third argument is the **inability to enforce water balance**: purely statistical models may violate mass conservation, producing physically impossible predictions — such as effective precipitation exceeding gross precipitation, or negative interception. The PaML paradigm addresses this directly: incorporating conservation laws as explicit constraints within the training process (WILLARD et al., 2022).

Positioning of this thesis within the PaML spectrum

The approach adopted in this thesis does not fall within a single PaML class, but mobilizes two classes in a complementary manner, one within each research axis.

Paper 1 (precipitation series reconstruction) is positioned within the **hybrid modeling subclass of physical model output as ML input** (WILLARD et al., 2022, Section 3.4.2): ERA5 and ERA5-Land products are outputs of general atmospheric circulation models — and therefore products of physical process models — used as input covariates in the ML algorithms. Quantile Mapping corrects the systematic bias between the distribution simulated by the physical model and the locally observed climatology, a step analogous to aligning physical model output prior to its ingestion by ML. In this configuration, the physical model is not replaced but rather coupled to ML as a provider of large-scale information that the local instrument alone could not capture.

Paper 2 (interception and effective precipitation modeling) combines two classes from Willard et al. (2022). On one hand, the physical constraint restricting the effective precipitation coefficient to the interval $0 < C < 1.5$ and the removal of physically inconsistent predictions prior to training correspond to the logic of **physics-guided loss functions** (3.1) — ensuring that the solution space is physically admissible from the outset. On the other hand, the monotonic constraints imposed on LightGBM — which structurally enforce that an increase in gross precipitation (P_e) implies an increase in effective precipitation (P_{eff}) — and the feature engineering entirely guided by eco-hydrological processes identified in the literature (event magnitude, Antecedent Precipitation Index, phenological seasonality, and spatial canopy heterogeneity) correspond to the spirit of **physics-guided architecture design** (3.3): domain knowledge embedded in the model structure to reduce the space of physically incoherent solutions, albeit applied to tree-based algorithms rather than neural networks.

In both papers, model selection based on multi-objective criteria that balance statistical precision (RMSE — *Root Mean Square Error*) and hydrological consistency (KGE — *Kling-Gupta Efficiency*) represents an additional layer of *post hoc* physical control: even if a model minimizes mean error, it is selected only if it preserves the statistical structure of the observed distribution — variability, bias, and correlation — providing an implicit check that reconstructed series and interception estimates respect the physics of the hydrological cycle.

Interpretability and critiques of ML in hydrology

The most robust critiques of ML in hydrology are organized around three main arguments (BEVEN, 2012; SHEN, 2018). The first is the **lack of physical interpretability**: a well-calibrated ML model may work for the wrong reasons — learning spurious correlations that collapse outside the training window. The literature's response has been the adoption of explainability tools such as SHAP values (developed in Section 5.3.5), which allow quantification of the contribution of each input variable to each individual prediction. The second argument concerns **dependence on abundant data**: algorithms such as LSTM typically require years of high-frequency series to learn reliable temporal patterns, potentially rendering them unsuitable for basins with short records. The response lies in transfer learning strategies and multi-basin training, which allow models trained on data-rich basins to be adapted to data-scarce basins (KRATZERT et al., 2019b). The third argument is the **inability to enforce water balance**: purely statistical models may violate mass conservation, producing physically impossible predictions. The PaML paradigm addresses this directly: incorporating conservation laws as explicit constraints within the training process (WILLARD et al., 2022).

2.4.2 Tree-Based Models: Random Forest and Gradient Boosting

Random Forest: bagging foundations and double randomization

Random Forest (RF), proposed by Breiman (2001), is an *ensemble learning* method built upon the principle that many weak and independent models together produce better predictions than any single strong model. Its mechanism is **bagging** (*bootstrap aggregating*): each of the B decision trees composing the ensemble is trained on a distinct bootstrap sample of the training set — with replacement, so that approximately 37 % of the original data are left out of each individual tree (*out-of-bag* data), which can be used for an unbiased estimate of generalization error (BREIMAN, 2001).

The central innovation of RF relative to simple bagging is the **additional randomization of attributes**: at each split node, only a random subset of m attributes (typically $m = \sqrt{p}$ for classification and $m = p/3$ for regression, where p is the total number of predictor variables) is considered when searching for the optimal split point. This double randomization — in samples and in attributes — produces more diverse and less correlated trees, amplifying the ensemble diversity effect and reducing the variance of the aggregated prediction by averaging (BREIMAN, 2001). The RF prediction for a new instance x is:

$$\hat{Y}_{\mathrm{RF}}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{2.16}$$

where $T_b(x)$ is the prediction of the b-th tree. For regression, the arithmetic mean across trees converges, through the diversity effect, to an estimator with lower variance than any individual tree, without increasing bias (BREIMAN, 2001). Table 2.4, at the end of this subsection, compares RF and gradient boosting algorithms in properties most relevant to hydrology.

Gradient Boosting: bias reduction through sequential correction

Gradient boosting operates in the opposite direction to bagging: instead of constructing independent trees and combining them by averaging, it builds trees sequentially, with each new tree correcting the residual errors of the previous ones (FRIEDMAN, 2001). The general algorithm, in the formulation of Friedman (2001), iteratively optimizes a loss function $\mathscr{L}(y, \hat{y})$ in function space:

$$F_m(x) \;=\; F_{m-1}(x) + \eta\, h_m(x) \tag{2.17}$$

where $F_m(x)$ is the prediction at iteration m, η is the learning rate, and $h_m(x)$ is a new tree fitted on the residuals (pseudo-gradients) of the previous model. While RF bagging reduces *variance* without changing *bias*, gradient boosting progressively reduces *bias* at the cost of increased variance, which is controlled through regularization mechanisms.

XGBoost

XGBoost (*eXtreme Gradient Boosting*), proposed by Chen e Guestrin (2016), is the most widely adopted implementation of gradient boosting in data science competitions and recent hydrological applications. Its central contribution is the introduction of a regularized objective function with second-order Taylor expansion. The model prediction $\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$ is obtained by minimizing:

$$\mathscr{L}(\phi) \;=\; \sum_i \ell(\hat{y}_i, y_i) \;+\; \sum_k \Omega(f_k) \tag{2.18}$$

where the regularization term $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ penalizes trees with many leaves (T) and large leaf weights (w), controlling model complexity. The second-order approximation of the loss function with respect to the previous model allows optimal leaf weights to be computed analytically:

$$w_j^* \;=\; -\frac{G_j}{H_j + \lambda} \tag{2.19}$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ are the sums of first- and second-order gradients for instances in leaf j (CHEN; GUESTRIN, 2016). XGBoost also implements native handling of missing values (learning optimal split direction for NaNs), which is particularly relevant for pluviometric series with gaps.

CatBoost and LightGBM

CatBoost (*Categorical Boosting*), developed by the Yandex team (PROKHORENKOVA et al., 2018), introduces the concept of *ordered boosting*: to avoid target leakage bias that occurs when residual statistics computed at iteration m already include information from the instance being predicted, CatBoost uses a temporal permutation of the data and computes leaf statistics only from prior instances in the ordering. For categorical variables, it employs target-based encoding with Bayesian smoothing. These properties make CatBoost particularly robust in small datasets and in datasets with mixed variable types (PROKHORENKOVA et al., 2018).

LightGBM (*Light Gradient Boosting Machine*) (KE et al., 2017) prioritizes computational efficiency through two innovative mechanisms: leaf-wise (best-first) growth, which splits the leaf with the highest gain instead of splitting all leaves at a given level (level-wise), resulting in deeper and more asymmetric trees that capture complex interactions with fewer iterations; and histogram-based binning, which discretizes continuous variables into at most 255 bins prior to training, drastically reducing split-search cost. A relevant property for hydrological applications is the ability to impose **monotonicity constraints** between predictors and the response, forcing the model to respect known physical relationships (KE et al., 2017).

Nonlinearity, interactions, and extrapolation limitations

Tree-based models are particularly suitable for capturing nonlinear relationships and high-order interactions among hydrological variables (see Quadro 2.4)) because each node split can create a stepwise response function in any direction of the feature space, without assuming a functional form (BREIMAN, 2001; FRIEDMAN, 2001). For effective precipitation modeling, this implies that the model can learn that the relationship between gross precipitation and interception changes abruptly when P_g exceeds S (canopy saturation threshold), without that threshold being specified *a priori*.

The fundamental limitation shared by all tree-based models is the inability to extrapolate beyond the training domain (*extrapolation gap*): decision trees are piecewise-constant models, and their prediction for values beyond the maximum observed during training is necessarily constant and equal to the highest observed value (WILLARD et al., 2022). In hydrology, this manifests as systematic underestimation of extreme precipitation events not observed during training. Regularization in XGBoost, ordered boosting in CatBoost, and leaf-wise growth in LightGBM reduce overfitting, but they do not solve the extrapolation problem, which is a structural limitation of tree-based approaches.

Box 2.4 – Comparison of main tree-based ensemble learning algorithms for hydrological applications.

Algorithm	**Core mechanism**	**Advantages in hydrology**	**Main limitations**
Random Forest Breiman (2001)	Bagging + double attribute randomization; averaging of independent trees	Robust to overfitting; OOB error estimate; good performance with limited data	No extrapolation; high memory usage; does not capture temporal trends
XGBoost Chen e Guestrin (2016)	Gradient boosting with L1/L2 regularization and second-order Taylor expansion	Native NaN handling; high accuracy; explicit regularization; fast convergence	Sensitive to hyperparameters; no extrapolation
CatBoost Prokhorenkova et al. (2018)	Ordered boosting with categorical encoding	Low bias in small datasets; robust with unordered series	Slower than LightGBM; less flexible for purely numerical variables
LightGBM Ke et al. (2017)	Leaf-wise growth + histogram binning	Very fast; monotonicity constraints; captures deep interactions	Prone to overfitting in small datasets; difficult interpretation of asymmetric trees

Fonte: The author (2026).

2.4.3 Recurrent Neural Networks and the Representation of Hydrological Memory

Hydrological memory and the need for stateful models

Hydrological processes have memory: the response of a catchment to a given precipitation event depends on the antecedent state of the system — soil moisture, shallow groundwater levels, canopy saturation state, air temperature, and vapor pressure deficit. This memory may be short, such as the residual canopy storage capacity (S_{res}) that is re-established within hours, or long, such as deep aquifer recharge that integrates precipitation over months or years (KRATZERT et al., 2019a; KRATZERT et al., 2019b). Static regression models, such as RF and XGBoost applied with a fixed feature window, represent memory indirectly by including antecedent-state variables as explicit predictors. Recurrent neural networks represent memory endogenously by maintaining a hidden state vector h_t that encapsulates all relevant information from previous time steps.

The vanishing gradient problem

Standard recurrent neural networks (RNNs) were proposed in the 1980s as an architecture for modeling temporal sequences. Their formulation is straightforward: the hidden state h_t at each time step is a nonlinear function of the previous state h_{t-1} and the current input x_t:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b) \tag{2.20}$$

The structural limitation of classical RNNs is the **vanishing gradient** problem: during training via backpropagation through time (*BPTT*), the gradient of the error with respect to activations many steps in the past is obtained by the product of many weight matrices. When the eigenvalues of the matrix W_h are smaller than 1 (as is typical in networks trained with sigmoidal saturation), this geometric product decays exponentially with the number of steps, rendering the gradient numerically negligible for long temporal dependencies (HOCHREITER; SCHMIDHUBER, 1997). In practice, this means that classical RNNs can only learn temporal dependencies over a few steps — insufficient to represent hydrological memory spanning weeks to months.

The LSTM gating mechanism

The LSTM (*Long Short-Term Memory*), introduced by Hochreiter e Schmidhuber (1997), resolves the vanishing gradient problem through a **gating** mechanism that controls information flow between memory cells across time steps. The core component is the **memory cell** c_t, which functions as an information "conveyor belt" across time: its gradient is preserved nearly intact because gradient flow through the cell involves only additions and multiplications by diagonal matrices (the gates), rather than products of dense matrices as in the classical RNN (HOCHREITER; SCHMIDHUBER, 1997).

The full LSTM equations at each time step t, given input x_t and previous states h_{t-1} and c_{t-1}, are:

$$f_t = \sigma\big(W_f[h_{t-1}, x_t] + b_f\big) \tag{2.21}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2.22}$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{2.23}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{2.24}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{2.25}$$

$$h_t = o_t \odot \tanh(c_t) \tag{2.26}$$

where $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent, and $\odot$ denotes the Hadamard (element-wise) product. The terms f_t, i_t, and o_t are the forget gate, input gate, and output gate, respectively — all taking values in $(0,1)$ and acting as analog valves that regulate information flow.

A physical analogy to hydrological memory

The elegance of the LSTM for hydrology lies in its empirical performance and in a deeper physical analogy carefully explored by Kratzert et al. (2019a). The **memory cell** c_t is analogous to the water storage of a reservoir (soil, shallow aquifer, snow): it receives inputs (precipitation, recharge) and loses water through dissipation processes. The **forget gate** f_t is analogous to reservoir losses — drainage, evapotranspiration, deep percolation: it determines what fraction of the previous storage is retained at the next time step. The **input gate** i_t regulates how much of the current input contributes to the stored state — analogous to the fraction of precipitation that effectively infiltrates and recharges the reservoir (versus that which becomes surface runoff). The **output gate** o_t determines what fraction of the stored state becomes observable as streamflow — analogous to the response function that transforms storage into measurable discharge at the catchment outlet (KRATZERT et al., 2019a).

Kratzert et al. (2019b) demonstrated, in an analysis of an LSTM trained on 531 basins from the CAMELS dataset (*Catchment Attributes and Meteorology for Large-sample Studies*), that individual cells in the LSTM layer developed internal representations statistically analogous to snow storage and soil moisture, even though these variables were not provided as inputs. This result suggests that the LSTM does not merely mimic the data — it implicitly learns the dynamic structure of the hydrological system.

Bidirectional LSTM and retrospective window

The **bidirectional LSTM** (*BiLSTM*) architecture processes the input sequence in two temporal directions simultaneously: one LSTM layer in the forward direction (past $\rightarrow$ future) and one in the reverse direction (future $\rightarrow$ past), whose hidden states are concatenated at each time step. In missing-value reconstruction tasks (where post-gap context is available at inference time), BiLSTM consistently outperforms the unidirectional LSTM by integrating both causal memory and future context into the estimate. For real-time forecasting, where future information is unavailable, the unidirectional LSTM is the appropriate design.

The choice of the **retrospective window length** (τ) — the number of past time steps provided to the model as context — is a critical modeling decision. A window of $\tau = 360$ h (15 days) captures canopy wetting and drying cycles, shallow soil moisture dynamics, and persistence patterns of synoptic systems responsible for tropical precipitation (KRATZERT et al., 2019b). Excessively long windows increase computational cost and introduce long-term noise that can degrade accuracy at the individual event scale.

2.4.4 Interpretability and Explainability in Hydrological ML Models

Mathematical foundations of SHAP values

The adoption of high-capacity ML models in hydrology faces a legitimate cultural obstacle: the hydrological community has a long tradition of models whose parameters are physically interpretable. An ML model that produces accurate predictions without explaining why is, within this tradition, a black box that generates no scientific learning and cannot be audited for physical consistency. **SHAP** values (*SHapley Additive exPlanations*), proposed by Lundberg e Lee (2017), address this issue by providing an additive and theoretically grounded decomposition of any model prediction into individual contributions from each predictor variable.

The mathematical foundation of SHAP values originates from Shapley's cooperative game theory. For a set of p predictor variables ("players") that collaborate to produce a prediction $\hat{y}$ (the "payout"), the Shapley value ϕ_j of variable j is the unique attribution that simultaneously satisfies four axioms: efficiency ($\sum_j \phi_j = \hat{y} - \mathbb{E}[\hat{y}]$), symmetry, dummy (no contribution for irrelevant variables), and linearity. Formally (LUNDBERG; LEE, 2017):

$$\phi_j = \sum_{S \subseteq \mathscr{F} \setminus \{j\}} \frac{|S|!\,(|\mathscr{F}| - |S| - 1)!}{|\mathscr{F}|!} \left[\hat{f}_{S \cup \{j\}}(x) - \hat{f}_S(x)\right] \tag{2.27}$$

where $\mathscr{F}$ is the full set of variables, S is a subset of $\mathscr{F} \setminus \{j\}$, and $\hat{f}_S(x)$ is the model prediction when only the variables in S are available. The SHAP value ϕ_j therefore represents the average marginal contribution of variable j when added to all possible coalitions of the remaining variables (LUNDBERG; LEE, 2017).

Global versus local importance

One of the most valuable properties of SHAP values is the distinction between **local** importance (for a specific prediction) and **global** importance (across the entire dataset). Global importance is estimated as the mean absolute SHAP value of each variable over all instances: $|\bar{\phi}_j| = n^{-1} \sum_i |\phi_{j,i}|$. Local importance, in contrast, explains why the model produced a specific prediction: for a particular precipitation event, what was the contribution of antecedent canopy moisture to the estimated interception? This decomposition is physically and scientifically informative.

In effective precipitation modeling, SHAP values allow one to address questions such as: is event intensity or antecedent canopy moisture the dominant driver of interception at this site? Does the answer vary between seasons? Are there intensity thresholds beyond which the contribution of certain variables changes sign? When SHAP values reveal patterns consistent with physical knowledge — greater positive influence of intensity on throughfall proportion above the canopy saturation threshold, greater importance of LAI during the wet season — this validates the model and quantifies physical relationships in an exploratory manner (LUNDBERG; LEE, 2017; SHEN, 2018).

2.5 Anomaly Detection in Hydrological Series: Conceptual Foundations

2.5.1 Concepts and Taxonomy of Anomalies in Hydrological Data

An **anomaly** (or *outlier*) in a hydrometeorological time series is, in formal terms, an observation that deviates significantly from the expected behavior given the context of the remaining observations (CHANDOLA; BANERJEE; KUMAR, 2009). The definitional challenge lies not in the word "significantly," but in the word "expected": what is expected depends on the adopted model of normal behavior — and in precipitation series, "normal" is intrinsically variable (seasons, frontal systems, extreme convective events) and markedly non-Gaussian. The literature distinguishes three fundamental types of anomalies in hydrometeorological time series (CHANDOLA; BANERJEE; KUMAR, 2009; CLIFTON et al., 2014).

Point anomalies: a single value that is inconsistent with its temporal neighbors and with simultaneous values of other variables. A record of 250 mm in 10 minutes during a completely dry day, without confirmation from neighboring stations, is the paradigmatic example. It typically corresponds to instrumental or transmission errors.

Contextual anomalies: a value that would be normal in another seasonal or climatic context but is anomalous given the current context. A record of 80 mm day^{-1} in July (the dry season in Lavras, where the historical daily maximum for that month rarely exceeds 30 mm) is a contextual anomaly: it would be entirely routine in January. Detection requires conditioning the analysis on seasonal context.

Collective anomalies: a contiguous subsequence of values that is individually plausible but collectively inconsistent. A series of 15 consecutive days with exactly constant values of 2 mm h^{-1} in a convective precipitation region is collectively anomalous even though each individual value is physically possible — the constancy pattern is incompatible with the natural variability of precipitation.

The specific challenge in distinguishing instrumental errors from real extreme events lies in the fact that both produce high values outside the historical experience range. The difference is contextual and relational: a real extreme event should be corroborated by other meteorological variables (wind, pressure, temperature) and by remote sensing information; it should occur within a temporal window consistent with the climatology of the generating systems; and it should exhibit an intensity profile compatible with the physics of convection, rather than a single isolated spike surrounded by zeros (CHANDOLA; BANERJEE; KUMAR, 2009; KIM; PACHEPSKY, 2010).

2.5.2 Unsupervised Algorithms for Anomaly Detection

Isolation Forest: isolation through recursive partitioning

Isolation Forest (IF), proposed by Liu, Ting e Zhou (2008), is an anomaly detection algorithm based on the principle that anomalous points are, by definition, few and different, and therefore easier to *isolate* through random partitioning of the feature space than normal points. The algorithm builds an ensemble of *isolation trees*: for each tree, a variable is randomly selected along with a split value between its observed minimum and maximum, partitioning the space; this process is recursively repeated until each point is isolated in its own leaf.

The average isolation depth of a point x — the mean number of partitions required to isolate it across all trees — is inversely proportional to its degree of anomaly. Anomalous points are isolated in few partitions (shallow trees) because they occupy low-density regions of the feature space; normal points require many partitions (deep trees) because they are surrounded by many similar observations. The anomaly score $s(x,n)$ is normalized to the interval $(0,1)$:

$$s(x, n) = 2^{-\frac{\mathbb{E}[h(x)]}{c(n)}} \tag{2.28}$$

where $h(x)$ is the path length of x in a tree, $\mathbb{E}[h(x)]$ its average over the ensemble, and $c(n)$ is the expected average path length for a set of n points in a binary search tree (BST) — a normalization factor (LIU; TING; ZHOU, 2008). Values of $s \approx 1$ indicate anomaly; $s \approx 0$ indicate normality.

The suitability of Isolation Forest for precipitation series is particularly strong compared to fixed-threshold methods, for a fundamental reason: by constructing partitions in the multivariate feature space (for example, hourly intensity $\times$ antecedent moisture $\times$ temperature $\times$ confirmation from neighboring stations), IF can identify as anomalous a combination of values that are individually plausible but jointly rare. A value of 60 mm h^{-1} during completely dry conditions according to all neighboring stations and the MERGE product is anomalous in multivariate space, even if 60 mm h^{-1} alone is historically plausible (LIU; TING; ZHOU, 2008).

Local Outlier Factor (LOF)

Local Outlier Factor (LOF), proposed by Breunig et al. (2000), is a density-based algorithm: a point is anomalous if its local density is significantly lower than the density of its nearest neighbors, quantified by the ratio between the local density of x and the average density of its k nearest neighbors. Formally:

$$\mathrm{LOF}_k(x) = \frac{\frac{1}{k}\sum_{o\in N_k(x)} \mathrm{lrd}_k(o)}{\mathrm{lrd}_k(x)} \tag{2.29}$$

where $\mathrm{lrd}_k(x)$ is the local reachability density of x and $N_k(x)$ is the set of its k nearest neighbors (BREUNIG et al., 2000). A $\mathrm{LOF}_k \approx 1$ indicates that x lies in a region of similar density to its neighbors (normal behavior); $\mathrm{LOF}_k \gg 1$ indicates a local anomaly.

The advantage of LOF over Isolation Forest is its sensitivity to **local anomalies**: a point may be anomalous relative to its *local context* even if its global distance from the median of the distribution is not extreme. In precipitation series, this is relevant for detecting contextual anomalies: a value of 20 mm h^{-1} is perfectly normal during an intense event, but anomalous if it occurs in isolation during a typically dry August day. LOF captures this contextual anomaly because the local density in that region of feature space (August, dry conditions, no neighboring confirmation) is low.

The limitation of LOF lies in its higher sensitivity to the parameter k and its greater computational cost ($\mathcal{O}(n^2)$ in the worst case) compared to Isolation Forest (BREUNIG et al., 2000). For high-frequency series with thousands to millions of records, Isolation Forest is computationally more scalable, whereas LOF is preferred when local heterogeneity in the feature space is pronounced and data volume is moderate.

The contamination parameter and the balance between errors

Both algorithms require specification of the **contamination** parameter f_c — the expected fraction of anomalies in the dataset. This choice is nontrivial and determines the balance between two types of errors with asymmetric consequences in hydrology (CHANDOLA; BANERJEE; KUMAR, 2009):

False positives (FP): real extreme events classified as anomalies and removed from the series. Each FP eliminates valuable information about system behavior under high-demand conditions, reducing the ability of ML models to learn these patterns.

False negatives (FN): instrumental errors not detected and retained in the series. Each FN contaminates the training set with false data, potentially distorting model parameters.

In high-frequency pluviometric series with prior basic quality control, contamination values between 0.5 % and 2 % are reported as appropriate in the literature (KIM; PACHEPSKY, 2010; LIU; TING; ZHOU, 2008). Very low f_c values increase FN; high values increase FP. A robust strategy is to conduct sensitivity analysis by varying f_c within the range 0.1 % to 5 % and evaluating the impact on the extreme-value distributions of the filtered series, selecting the value that best preserves the upper percentiles of the reference series while eliminating patterns inconsistent with the physical behavior of precipitation events.

2.6 Model Selection and Evaluation in Hydrology: Criteria and Metrics

2.6.1 Hydrological Performance Metrics

The choice of an evaluation metric is not a minor methodological detail: it is a statement of values. It implicitly defines what the model should optimize, which errors are penalized more heavily, and which properties of the series must be preserved. In hydrology, where errors have asymmetric physical consequences — underestimating a flood peak is more critical than overestimating it — metric selection requires careful analysis of the mathematical properties and biases of each index.

Limitations of NSE and RMSE

The **Nash–Sutcliffe Efficiency** (NSE), proposed by Nash e Sutcliffe (1970), has for decades been the reference metric in catchment hydrology:

$$\text{NSE} = 1 - \frac{\sum_t (Q_{\text{obs,t}} - Q_{\text{sim,t}})^2}{\sum_t (Q_{\text{obs,t}} - \bar{Q}_{\text{obs}})^2} \tag{2.30}$$

NSE compares the residual variance of the model with the total variance of the observed series: NSE = 1 indicates perfect fit; NSE = 0 indicates that the model performs no better than predicting the mean of the series; NSE < 0 indicates that the mean is a better predictor than the model. The fundamental limitation of NSE, demonstrated by Gupta et al. (2009), is that it can be decomposed into three components (correlation, bias, and variability) that compensate each other: a model may exhibit high NSE with substantial bias and compressed variability, provided that temporal correlation is high. Knoben, Freer e Woods (2019) showed that NSE = 0 corresponds, in practical performance terms, to a KGE of approximately -0.41 — indicating that the NSE = 0 threshold is overly optimistic as a discriminator between useful and useless models.

RMSE measures mean squared error in the units of the observed variable. Its principal limitation is that, being scalar and non-normalized, it does not decompose error sources, penalizes errors uniformly regardless of hydrological relevance, and is sensitive to the scale of the variable, complicating comparisons across basins or periods with different precipitation regimes.

The decomposition of KGE

The **Kling–Gupta Efficiency** (KGE), proposed by Gupta et al. (2009) and refined by Kling, Fuchs e Paulin (2012), represents a significant methodological advance by explicitly decomposing error into three independent diagnostic components:

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{2.31}$$

where:

$$r = \text{Cor}(Q_{\text{sim}}, Q_{\text{obs}}) \quad \text{(Pearson linear correlation — temporal synchrony)} \tag{2.32}$$

$$\alpha = \frac{\sigma_{\text{sim}}}{\sigma_{\text{obs}}} \quad \text{(standard deviation ratio — relative variability)} \tag{2.33}$$

$$\beta = \frac{\mu_{\text{sim}}}{\mu_{\text{obs}}} \quad \text{(mean ratio — systematic bias)} \tag{2.34}$$

The optimal KGE value is 1 (all components equal to 1). KGE was defined based on the recognition that models with good temporal correlation may still exhibit substantial systematic bias ($\beta \neq 1$) or systematically underestimated variability ($\alpha < 1$) — features that NSE merges into a single number without distinction. Knoben, Freer e Woods (2019) established that KGE ≥ -0.41 is functionally equivalent to NSE ≥ 0 as a criterion for a "useful" model, and that the empirically proposed reference value KGE = 0.5 by Kling, Fuchs e Paulin (2012) corresponds to models with substantially superior performance relative to the historical mean benchmark.

The RMSE–KGE dilemma in practice

A model with low RMSE but high KGE and a model with low RMSE but low KGE have radically different practical implications. A model that minimizes RMSE while exhibiting low KGE typically has $\alpha < 1$: it underestimates the variability of the simulated series (producing smoother predictions than observations), resulting in systematic underestimation of flow or interception peaks. This behavior is functionally equivalent to a model that "works on average" but fails when it matters most — during extreme events that dominate hydrological risk (GUPTA et al., 2009).

2.6.2 Multiobjective Optimization in Model Selection

Theoretical foundations: why single-metric selection fails in hydrology

Model selection based on a single performance criterion — a practice that dominated operational hydrology for decades — rests on a tacit assumption that is rarely examined: that there exists a single loss function capable of adequately capturing all hydrologically relevant aspects of model behavior. In practice, this assumption fails systematically. A model calibrated to minimize RMSE will learn to reproduce the central tendency of the distribution well, but will systematically underestimate peak precipitation or discharge values — which, being rare, contribute little to the quadratic loss even when they are hydrologically critical. As demonstrated by (GUPTA; SOROOSHIAN; YAPO, 1998), models selected by RMSE alone consistently underperform in representing peak flows, the events that matter most for flood risk assessment and interception dynamics in canopy saturation regimes.

The Nash-Sutcliffe Efficiency (NSE) (NASH; SUTCLIFFE, 1970), despite being the standard criterion in hydrological model calibration for over five decades, is mathematically equivalent to $1-\mathrm{RMSE}^2/\sigma^2_{\mathrm{obs}}$ and therefore inherits the same sensitivity to mean-performance values while penalizing extreme errors in a quadratic, non-robust manner. (LEGATES; MCCABE, 1999) documented that NSE values can be high even when a model systematically underestimates flood peaks or wet-season totals, because the criterion is dominated by performance during the more frequent moderate-intensity events.

The Kling-Gupta Efficiency (KGE), developed by (GUPTA et al., 2009) as a decomposition of the mean squared error into three physically interpretable components — linear correlation (r), bias ratio ($\beta = \mu_s/\mu_o$), and variability ratio ($\gamma = \mathrm{CV}_s/\mathrm{CV}_o$) — partially addresses these limitations by explicitly penalizing systematic bias and variance differences separately from random error. However, optimizing KGE and RMSE simultaneously often leads to a fundamental trade-off: models with high KGE tend to preserve distributional properties at the expense of event-scale precision, while models with low RMSE tend to smooth extreme values to minimize squared error. This tension is inherently a multiobjective problem that cannot be resolved by collapsing it into a single scalar criterion without introducing subjective weighting assumptions (EHRGOTT, 2005).

The Pareto frontier

Selecting a single “optimal” model from a set of candidates evaluated under multiple criteria is a multiobjective optimization problem without a trivial analytical solution. The central concept is the **Pareto frontier**: in the objective space $(\mathscr{O}_1, \mathscr{O}_2, \ldots, \mathscr{O}_k)$, a solution A **dominates** a solution B if A is at least as good as B in all objectives and strictly better in at least one. The Pareto frontier (or set) is the collection of all non-dominated solutions: for any solution on the frontier, it is impossible to improve one objective without worsening at least one other (BOZORG-HADDAD et al., 2016; SHEN; TOLSON; MAI, 2022).

In the context of this thesis, with two primary objectives — minimizing RMSE and maximizing KGE — the Pareto frontier consists of models for which no other model simultaneously presents lower RMSE and higher KGE. Choosing among models on the frontier involves a value-based decision regarding the trade-off between absolute error and preservation of distributional properties.

The Borda method as a compromise solution

The **Borda Method** (*Borda Count*), originally proposed in the eighteenth century by Jean-Charles de Borda as a voting system and later adapted for multiobjective model selection, provides an objective and traceable compromise solution (BOZORG-HADDAD et al., 2016; SANTOS et al., 2024). The procedure is:

1. For each objective k, rank all M candidate models from best to worst according to that objective, assigning ranks $r_k(m) \in \{1, \ldots, M\}$.
2. Compute the Borda score of each model as the sum of its ranks across all objectives:

$$\text{Borda}(m) = \sum_{k=1}^{K} r_k(m) \tag{2.35}$$

3. Select the model with the lowest Borda score (best aggregated position).

The most important property of the Borda Method in this context is its **resistance to single-metric dominance**: by aggregating ordinal positions rather than absolute metric values, it prevents a model with very low RMSE but very poor KGE from outperforming a model that is balanced across both criteria. The result is a compromise model that is simultaneously competitive in absolute accuracy and in preservation of variability and bias (SANTOS et al., 2024). The traceability of the criterion ensures reproducibility: any researcher replicating the ranking obtains the same selected model.

2.6.3 Temporal and Spatial Validation Strategies

The inadequacy of random K-fold for time series

Standard K-fold cross-validation — which randomly divides the dataset into K subsets and successively evaluates the model on each subset using the remaining data for training — is statistically rigorous for i.i.d. (*independently and identically distributed*) data. Hydrological time series fundamentally violate the independence assumption: observations close in time are highly correlated (positive autocorrelation at scales from hours to weeks). When past and future records of the same event are randomly assigned to training and testing sets, the model may "memorize" patterns from that event and exhibit artificially high performance on the test set — a phenomenon known as **temporal data leakage** (ROBERTS et al., 2017). The result is an overly optimistic estimate of the model's ability to generalize to unseen periods.

Time-Series Cross-Validation (Walk-Forward)

The appropriate temporal validation strategy for hydrological series is **Time-Series Cross-Validation** (TSCV), also known as *walk-forward* or *expanding window* validation (ROBERTS et al., 2017). The procedure is:

1. Define an initial training set (first t_0 years of the series).
2. Train the model on the initial set and evaluate its performance on the subsequent Δt years (test window).
3. Expand the training set to include the Δt years from the previous test.
4. Repeat until the entire series is consumed, obtaining K performance estimates over chronologically ordered and previously unseen periods.

TSCV ensures that the trained model never sees future data during training, faithfully reproducing operational application conditions — where the model must predict the future without knowledge of future states. The average performance across the K folds provides a conservative (and therefore reliable) estimate of the model's generalization capacity.

Leave-One-Point-Out (LOPO) for spatial generalization

Temporal validation assesses whether the model generalizes to new periods at the same sensor locations; **spatial** validation assesses whether the model generalizes to new locations within the same ecosystem. The *Leave-One-Point-Out* (LOPO) strategy is the spatial analogue of TSCV: in each iteration, the model is trained using data from all measurement points except one, and evaluated at the excluded point (ROBERTS et al., 2017). For the network of 32 understory rain gauges at Matinha (UFLA), LOPO evaluates whether the model can estimate effective precipitation at an uninstrumented forest location using climatic variables and measurements from neighboring instrumented points — a scenario directly relevant to operational application across the entire basin.

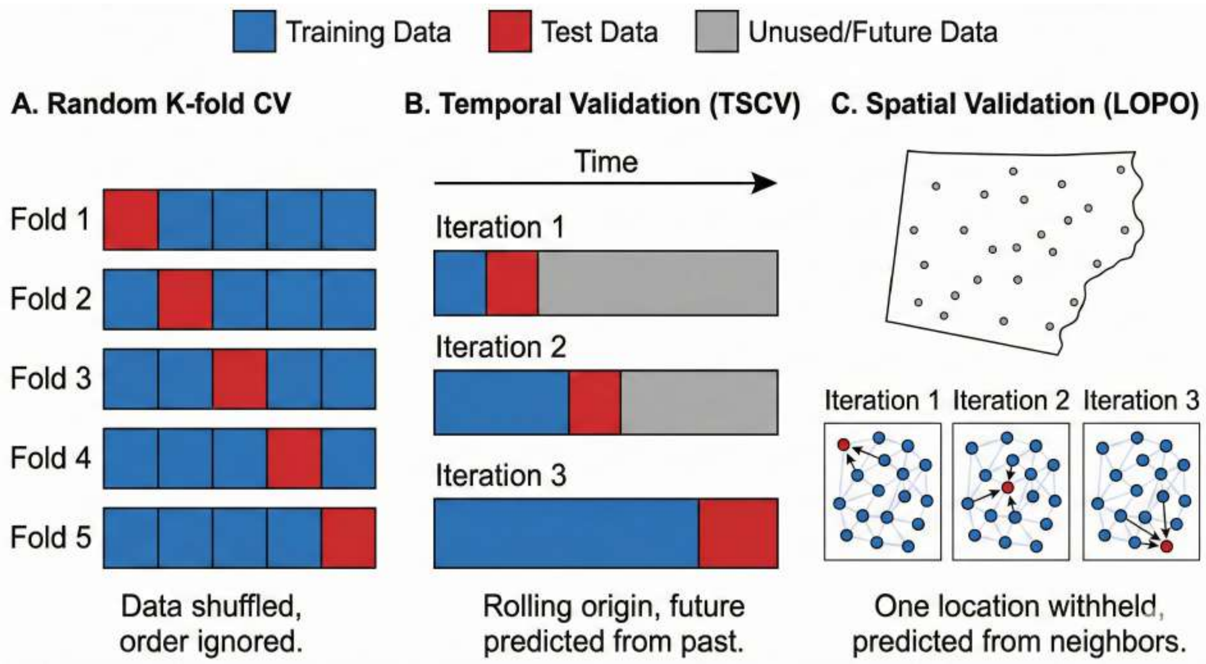
Figure 2.2 comparatively illustrates the random K-fold, TSCV, and LOPO validation schemes.

Multi-year sliding windows and detection of nonstationarity

A valuable extension of TSCV is the use of **sliding windows** of fixed size (instead of expanding windows), which evaluate model performance across different periods without allowing the training set to grow indefinitely. Comparing performance between consecutive temporal windows provides a sensitive indicator of **nonstationarity** and model degradation over time: if KGE systematically declines from window to window, this signals that the system has evolved in a way the model no longer captures — whether due to changes in precipitation regime, progressive phenological shifts in the forest, or degradation in sensor calibration.

This diagnostic capability is particularly relevant for long-term series such as those from Matinha, which span more than a decade and include exceptional drought periods (2014–2016) that may have permanently altered certain structural parameters of the ecosystem (RODRIGUES et al., 2021a).

Figure 2.2 – Comparative illustration of validation strategies. **A. Random K-fold Cross-Validation:** Data is randomly partitioned, ignoring spatiotemporal order. **B. Temporal Validation (TSCV):** Utilizes a rolling-origin approach where the model trains on past data (blue) to predict subsequent future periods (red), respecting chronological order. **C. Spatial Validation (Leave-One-Point-Out - LOPO):** The analogue of TSCV for space, illustrated for a sensor network like Matinha. In each iteration, a single sensor location is withheld as the testing target (red dot) to evaluate if the model trained on neighboring points (blue network) can generalize to new, uninstrumented locations (ROBERTS et al., 2017).



Fonte: The author (2026).

2.7 Synthesis of Gaps in the Literature and Research Positioning

The previous sections established four conceptual pillars — the ecohydrology of rainfall partitioning, the problem of data integrity, the foundations of machine learning, and model evaluation criteria — and each of them revealed structural fractures that the existing literature has not yet resolved. This section systematically gathers these gaps, organizes them by domain, and articulates the epistemological positioning of this thesis.

2.7.1 Gaps in Precipitation Reconstruction in Tropical Forests with Incomplete Monitoring Networks

The problem of reconstructing pluviometric series in tropical forest environments is superficially similar to the general problem of gap-filling in hydrometeorological time series, but it presents specific dimensions that the literature has not addressed systematically. The review by Teegavarapu (2012) catalogued dozens of gap-filling methods, and Kim e Pachepsky (2010) demonstrated the superiority of nonlinear methods (decision trees, neural networks) over linear regression and traditional models for daily precipitation in agricultural basins in the United States. However, these studies operate in a radically different context from that investigated here in at least three respects.

The first aspect is the **temporal scale**. The overwhelming majority of precipitation gap-filling studies use daily or monthly data. When shifting to hourly resolution — necessary to represent forest interception, which responds to events at the scale of minutes to hours — the problem changes in nature: the hourly precipitation distribution is much more right-skewed and contains a higher proportion of zeros than the daily distribution; spatial dependence between neighboring stations is weaker at the hourly scale because correlation among instantaneous convective records decays rapidly with distance; and linear regression methods lose further adequacy because inter-station relationships are strongly nonstationary at intradaily scales. None of the classic studies validated their methods specifically at hourly resolution in tropical convective environments, leaving a direct methodological gap.

The second aspect concerns the **preservation of distribution tails**. The gap-filling literature prioritizes average-error metrics (RMSE, MAE), which are dominated by performance at the median values of the distribution. The consequence, documented by Kim e Pachepsky (2010) but rarely addressed explicitly, is that models ranked best by RMSE systematically underestimate extreme precipitation percentiles (P_{95}, P_{99}). For interception modeling, this is particularly critical: large events — those that saturate the canopy and generate the greatest fraction of throughfall — that most strongly influence annual water balance. A reconstructed series with compressed tails will produce an interception model that systematically underestimates effective precipitation during the most hydrologically relevant events. To our knowledge, no study in the consulted literature proposes and validates a pluviometric gap-filling criterion specifically optimized for the simultaneous preservation of mean-error metrics and extreme values in hourly series from tropical forest environments.

The third aspect is the **integration of quality control, anomaly detection, and gap-filling** into a unified pipeline. In practice, these three processes are frequently treated sequentially and independently: first applying fixed-threshold quality control tests, then an anomaly detection algorithm, and finally a gap-filling method. However, as demonstrated in Section 5.3.1, the boundary between instrumental error and real extreme event is contextual and depends on multivariate information that none of the three processes can fully exploit in isolation. The proposal of an integrated pipeline that employs unsupervised anomaly detection algorithms (Isolation Forest, LOF), informed by contextual variables (spatial confirmations, atmospheric indices, forest phenological state), to distinguish errors from extremes and redirect data toward removal or conditional preservation appears to constitute a novel methodological contribution in tropical forest hydrology.

2.7.2 Gaps in Interception and Effective Precipitation Modeling in Semideciduous Atlantic Forest

The semideciduous Atlantic Forest is one of the most studied biomes in Brazilian environmental science, yet interception modeling in this specific ecosystem still presents gaps that go beyond the general limitations of the Gash and Rutter models discussed in Section 2.2.4. Four specific gaps emerge from the literature review.

The **first gap** is the absence of joint spatiotemporal validation. All reviewed interception studies in the Atlantic Forest — including those conducted near the study area of this thesis (Junqueira Junior et al., 2019; RODRIGUES et al., 2021a) — perform exclusively temporal validation: the model is calibrated and evaluated at one or a few understory rain gauges over time. This strategy does not assess the model's ability to estimate effective precipitation at *uninstrumented* forest locations, which is the relevant scenario for watershed management. The LOPO (*Leave-One-Point-Out*) strategy described in Section 2.6.3 is the appropriate validation design for this objective, yet it has not been systematically applied in any consulted Atlantic Forest interception study. The lack of spatial validation prevents determining whether models learn sensor-location-specific relationships (e.g., the canopy effect of a particular species) or genuinely generalizable relationships for the entire forest.

The **second gap** is the absence of models that explicitly and dynamically represent seasonal variation in interception parameters. The *Variable Storage Gash* model (MUZYLO et al., 2009) was proposed as an extension for variable canopy cover, but its application in semideciduous Atlantic Forest with complete multiannual series is virtually absent in the national literature. The few studies that calibrated the Gash model separately for wet and dry seasons — such as Rodrigues et al. (2021a) for the UFLA site — confirmed that parameters vary substantially between seasons but did not propose a dynamic function continuously updating S, $\bar{E}$, and p throughout the phenological cycle. Modeling this continuous temporal dynamics of parameters therefore remains an open problem.

The **third gap** is the systematic quantification of **canopy hydrological memory at the event scale**. Although the importance of antecedent canopy moisture (W_t) has been recognized since the original Gash formulation (BY; GASH, 1979), its influence over long hourly series has not been systematically quantified in the Atlantic Forest literature. Questions such as "what is the effective decorrelation time between events that allows a new event to be treated as conditioned on a dry canopy?" or "how do sequences of events affect monthly interception totals?" remain without quantitative answers for the UFLA site and, more broadly, for Brazilian semideciduous forests.

The **fourth and most directly relevant gap** is the absence, in the consulted literature, of operational models capable of predicting effective precipitation P_{eff} in real time — that is, using only climatic variables available from conventional automatic weather stations (gross precipitation, temperature, relative humidity, wind speed, barometric pressure) — without requiring high-cost canopy sensors (understory rain gauges, canopy water content sensors, dendrometers). Table 2.5 synthesizes these four gaps, the current state of the art, and the contribution of this thesis in relation to each.

Box 2.5 – Synthesis of specific gaps in interception and effective precipitation modeling in semideciduous Atlantic Forest, the current state of the art, and the contribution of this thesis.

Identified gap	**Current state of the art**	**Contribution of this thesis**
Joint spatiotemporal validation	Temporal validation only at 1–4 understory rain gauges; no systematic LOPO application in the Atlantic Forest	LOPO validation over 32 measurement points + multiannual walk-forward TSCV; simultaneous assessment of spatial and temporal generalization
Dynamically seasonal interception parameters	Separate seasonal calibrations; no continuous $S(t)$, $p(t)$ function for semideciduous forest; Variable Storage Gash not systematically applied in Brazil	LSTM architecture with 360 h retrospective window implicitly learns seasonal parameter dynamics without explicit recalibration
Event-scale canopy hydrological memory	Antecedent moisture effect acknowledged but not systematically quantified in long hourly Atlantic Forest series	SHAP-based analysis of the relative contribution of estimated antecedent canopy state as a function of season and inter-event interval
Operational P_{eff} model without canopy sensors	All published models require direct throughfall measurement for calibration or validation	Model trained and validated using only conventional automatic station variables as input; throughfall used exclusively for calibration/validation phase

Sources: Gash e Morton (1978), Muzylo et al. (2009), Junqueira Junior et al. (2019), Rodrigues et al. (2021a), Teegavarapu (2012), Kim e Pachepsky (2010).
Fonte: The author (2026).

2.7.3 The Absence of Operational Models as a Central Scientific Problem

The distinction between a scientific model and an operational model is frequently overlooked in academic hydrology. A scientific model may require input variables that demand high-cost instrumentation, continuous calibration, and specialized technical expertise for maintenance; its value lies in demonstrating that a process can be represented with precision when all necessary data are available. An operational model, in contrast, must function with the data that actually exist within conventional monitoring systems, not with the data we would ideally like to have. This distinction matters because most interception models published in the literature belong to the first category.

The Gash and Rutter models, in their original formulations, require field-measured throughfall series as calibration variables; without observed throughfall, there is no way to estimate S, p, and $\bar{E}$. Machine learning algorithms presented in the forest hydrology literature — such as studies cited by Shen (2018) in streamflow forecasting contexts — are also rarely trained specifically to estimate forest effective precipitation, and when they are, they typically use input variables that include understory measurements. The practical consequence is that the approximately 5,000 Brazilian municipalities with remaining forest cover in their water supply basins do not have validated models to estimate how much of the rainfall falling over forest effectively contributes to groundwater recharge and surface runoff.

This absence is not merely a scientific problem: it is a bottleneck for water resources management. Water use charges in Brazilian river basins are regulated by the National Water Resources Policy (Law 9.433/1997) and require reliable estimates of water availability. In basins with significant forest cover, effective precipitation — not the gross precipitation recorded at conventional stations — is the variable that feeds surface runoff and aquifer recharge. Using gross precipitation as a proxy for effective precipitation introduces a systematic positive bias (overestimation of water availability) that may result in water use permits exceeding the real capacity of the system (BRUIJNZEEL, 2004).

This thesis directly addresses this problem by proposing and validating models that estimate P_{eff} from variables available at conventional automatic weather stations of the Vaisala or Davis Instruments type — temperature, relative humidity, wind speed, barometric pressure, and gross precipitation — without the need for any additional canopy sensors. Article I of the thesis focuses on the development and validation of multiple algorithms for this task; Article

II explores the physical interpretability of the models and their generalization to periods with climatic conditions distinct from those observed during training.

2.7.4 Positioning of the Thesis in Relation to the State of the Art

In terms of conceptual and methodological contributions, this thesis proposes three articulated advances. The **first** is the demonstration that machine learning algorithms trained with readily obtainable climatic variables — without canopy sensors — can estimate effective precipitation at hourly scale with performance (KGE and RMSE) comparable to or superior to the Gash model calibrated with complete field data. This not only validates the ML approach, but implies that the model implicitly learned the physical structure of the interception process from climatic data, without it being explicitly encoded.

The **second advance** is methodological: the integration of a preprocessing pipeline based on unsupervised anomaly detection (Isolation Forest or LOF) with ML model training, such that input series quality is evaluated and corrected consistently within the same feature space used by the predictive model. To our knowledge, this integration is unprecedented in the specific domain of tropical forest hydrology.

The **third advance** is the use of SHAP values to extract quantitative physical interpretations from the trained models, addressing questions about the relative importance of antecedent moisture, event intensity, and time of year (phenological proxy) in determining interception efficiency throughout the annual cycle. This interpretability exercise transforms the model from a predictive tool into an instrument of scientific discovery regarding the ecohydrological processes of the studied forest.

2.7.5 Experimental Infrastructure: The Matinha Dataset in Context

The existence of a long-term, continuously monitored forest site such as Matinha is central to this thesis. A detailed assessment of how Matinha compares to other major monitoring infrastructures—both internationally and within Brazil—is presented in Chapter 3 (Section 3.2.5). There, we show that while networks such as FLUXNET (>500 sites globally) and the newly released CAMELS-BR dataset (897 Brazilian catchments) provide important regional coverage, they lack the specific combination of features that make Matinha uniquely suited for this research:

simultaneous hourly-scale precipitation monitoring at 32 spatial points, species-specific stemflow data for 15 tree species, and continuous soil moisture observations across five depths over more than a decade.

In brief, this infrastructure enables the two-part research strategy outlined above: first, to develop robust methods for reconstructing high-frequency precipitation series in data-scarce tropical contexts (Article 1), and second, to build operational models of rainfall partitioning that capture real canopy heterogeneity without relying on expensive canopy sensors (Article 2).

The combination of hourly resolution, spatial density of 32 understory rain gauges distributed beneath identified forest species, more than a decade of continuous records, and complementary time series of soil moisture, phenology, and complete meteorological variables is a dataset without parallel for semideciduous Atlantic Forest. It is this singularity that justifies the geographical focus of the research and that enables development of the proposed models with sufficient statistical power to detect subtle differences among algorithms and to support SHAP-based interpretability analysis with statistical robustness.

Box 2.6 – Comparison between the Matinha (UFLA) dataset and similar experimental infrastructures in aspects relevant to this thesis.

Characteristic	**Matinha (UFLA)**	**FLUXNET-BR**	**CAMELS-BR**	**LTER-Brasil**
Forest typology	Montane Semideciduous Seasonal Forest — Atlantic Forest (FERREIRA et al., 2018; BOTREL; CARVALHO, 2004)	Variable (HAUGHTON et al., 2018)	Variable (croplands, mosaics, forests, shrublands) (CHAGAS et al., 2020)	Variable (Amazon, savanna, coastal, marine) (BERGALLO et al., 2023; CORDEIRO et al., 2022)
Temporal resolution	Not specified in sources	30 min (fluxes) (HAUGHTON et al., 2018)	Daily (CHAGAS et al., 2020)	Variable (weekly, monthly, annual) (CORDEIRO et al., 2022)
Throughfall measurement	Yes (throughfall and stemflow equipment installed) (Junqueira Junior et al., 2019)	Not specified in sources	No (CHAGAS et al., 2020)	Variable / Rarely
Record duration	Not specified in sources	Variable	1980–2018 (CHAGAS et al., 2020)	Variable (up to 29 years) (CORDEIRO et al., 2022)
Canopy variables	Not specified in sources	CO_2/H_2O/energy fluxes (HAUGHTON et al., 2018; VUICHARD; PAPALE, 2015)	No (static land cover fractions) (CHAGAS et al., 2020)	Variable (BERGALLO et al., 2023)
Soil coverage	Not specified in sources	Variable	Yes (static characteristics at 30 cm depth) (CHAGAS et al., 2020)	Variable

Fonte: The author (2026).

3 THESIS OUTLINE

3.1 Overview and Structure of the Thesis

This thesis sits at the intersection of *data-driven hydrology* and forest ecosystem science. Rather than replacing physical understanding with machine learning, the approach uses ML algorithms where they are needed most: in territories where field data are noisy, sparse, or highly variable—conditions that defeat classical linear methods but are typical of tropical forest hydrology (SHEN, 2018; BAşAğAOğLU et al., 2022).

The fundamental problem is straightforward in concept but difficult in practice. We lack a validated framework for two tightly linked tasks: (1) reconstructing continuous, high-fidelity precipitation time series from fragmentary field observations, and (2) using those series to estimate effective precipitation—the water that actually reaches the soil—at the scale of individual storms. Without this capability, reliable water balance modeling and climate-change impact assessment remain out of reach.

To address this problem, the thesis is structured in two complementary parts, each presented as an original research article:

Article 1 tackles **precipitation series reconstruction**. The 11-year historical dataset from Matinha (September 2012 to December 2023) contains unavoidable gaps: sensors fail, power is lost, maintenance disrupts collection. Traditional gap-filling methods use linear interpolation or historical averaging. These approaches are problematic because they smooth away the sharp, convective character of tropical rainfall. Article 1 proposes an alternative: a hybrid framework combining unsupervised outlier detection, bias correction using global reanalysis data (ERA5), and multi-objective model selection (Pareto-Borda method). The result is a reconstructed precipitation series that preserves both statistical accuracy and physical realism.

Article 2 addresses **rainfall partitioning and interception modeling**. Once precipitation is reliably reconstructed, the next challenge is predicting how much of it reaches the soil. Traditional models assume constant interception—that the canopy always retains the same fraction of incoming rain. This assumption breaks down in semideciduous forests where trees lose their leaves seasonally. Article 2 uses machine learning ensembles (Random Forest, XGBoost, CatBoost, LightGBM) to learn interception dynamics directly from data. By incorporating variables like rainfall intensity, antecedent moisture, and season, the model captures the event-by-event variability that static parameters cannot.

The two articles are linked not only thematically but procedurally. The clean, quality-controlled precipitation series from Article 1 becomes the input for computing hydrological memory indices—such as the Antecedent Precipitation Index (API)—that the models in Article 2 need. In other words, data quality is not a separate problem; it directly shapes the credibility of the ecohydrological findings.

The sections that follow describe the physical and human context in which this work takes place: the geography and climate of the study area, the soils and vegetation, and the monitoring infrastructure that makes both articles possible.

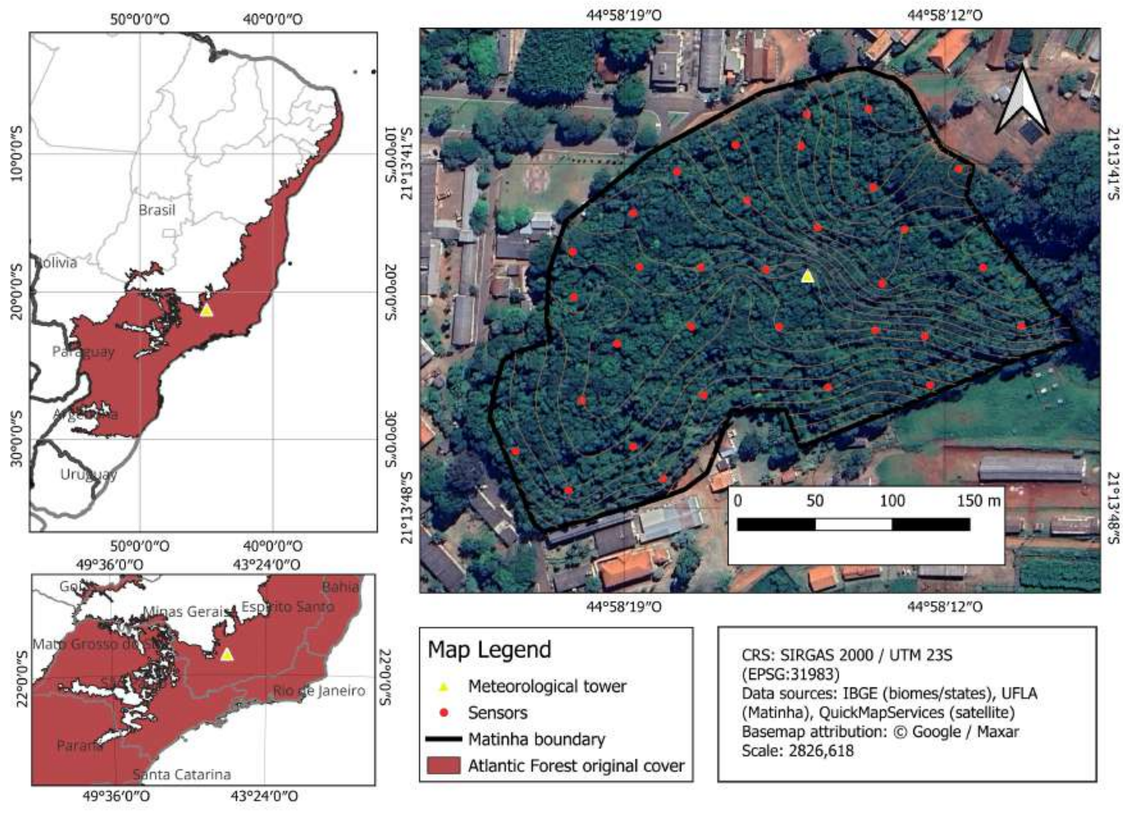
3.2 Study Area

3.2.1 Location and Geographic Context

The research is conducted in a Montane Semideciduous Atlantic Forest remnant locally known as Matinha, officially designated as the Forest Reserve of Montane Semideciduous Seasonal Forest (RF-FESM). The experimental site is located on the campus of the Federal University of Lavras (UFLA), in the municipality of Lavras, southern Minas Gerais, at coordinates 21°13’40”S and 44°57’50”W, see Figure 3.1, with a mean elevation of 925 m a.s.l. (Junqueira Junior et al., 2019). The forested area covers 6.35 ha of native vegetation in an advanced stage of ecological succession and holds one of the longest continuous records of intensive hydrological monitoring among Atlantic Forest fragments in Brazil (GUAUQUE-MELLADO et al., 2022; GUAUQUE-MELLADO et al., 2026).

Southern Minas Gerais lies within the influence zone of the Mantiqueira Range. The undulating relief, with slopes of 5–15% across the experimental area, drives local orographic convection and amplifies precipitation variability at sub-kilometric scales (Junqueira Junior et al., 2019; RODRIGUES et al., 2021a). This makes Matinha representative of conditions found throughout the montane Atlantic Forest — a biome that supplies water services to more than 125 million people and retains less than 15% of its original cover.

Figure 3.1 – Topographic characterization of the Matinha study area (RF-FESM) at the Federal University of Lavras (UFLA). The map highlights the undulating relief and elevation (925 m a.s.l.) which influences local atmospheric convection and moisture capture. Map data follows the SIRGAS 2000 / UTM 23S coordinate system.



Fonte: The author (2026).

3.2.2 Climate and Regional Atmospheric Dynamics

Lavras sits in a subtropical climate zone (Köppen classification: Cwa) characterized by rainy summers and dry winters. This seasonality is not merely a climatological fact; it is the primary forcing that shapes how the forest grows, loses leaves, and partitions water. Mean annual precipitation is approximately 1,511 millimeters, but the distribution is highly skewed: about 85% falls between October and March, often in intense convective storms. The complementary dry season (April to September) imposes water stress that causes up to 50% of tree species to shed their leaves. This phenological response directly alters canopy storage capacity and how much water can be intercepted and evaporated.

Understanding the variability of rainfall is as important as understanding its average amount. The regional rainfall regime is shaped by the interplay of four main atmospheric systems—the South Atlantic Convergence Zone (SACZ), frontal systems (cold fronts), the South American Monsoon System (SAMS), and local orographic convection triggered by the Mantiqueira Range. These systems interact in complex ways: the SACZ may bring persistent rainfall over days, while a local convective cell can dump intense rain on a small area in minutes. For hydrological modeling, these distinctions matter enormously. A model calibrated on average monthly rainfall cannot capture event-scale physics. This is why detailed characterization of these atmospheric drivers, presented in Table 3.1, is essential background for the precipitation reconstruction methods in Article 1.

3.2.3 Physiographic and Pedological Characterization

The predominant soil in the experimental area is a Red Dystroferric Latosol (*Oxisol* according to the U.S. Soil Taxonomy), with very clayey texture—clay content greater than 60%— and a granular microaggregated structure resulting from intense tropical weathering (Junqueira Junior et al., 2017). This pedological configuration confers specific physical–hydrological properties that are essential for understanding water dynamics within the soil profile and, by extension, for interpreting the results of both articles. Table 3.1 systematizes the main properties and their hydrological relevance.

This soil composition—60% clay yet with high macroporosity—seems paradoxical. Intuitively, a clay-rich soil should resist water movement. But the granular microaggregated

Box 3.1 – Main atmospheric systems governing the rainfall regime of Southern Minas Gerais and their implications for hydrological modeling.

Atmospheric System	Period (Months)	Main Mechanism	Contribution to Precipitation	Reference
South Atlantic Convergence Zone (SACZ)	Oct–Mar (Summer)	Persistent cloud band oriented northwest-southeast. It is driven by moisture convergence from the Amazon basin towards the South Atlantic Ocean.	Accounts for a substantial volume of summer rainfall; contributes up to 56% in March and 41% in January across the Southeast.	(VERDAN, 2023; REBOITA et al., 2015)
Frontal Systems (Cold Fronts)	Year-round (Peak: Apr–Sep)	Cold air masses (polar advection) advancing over the continent. They interact with Amazonian moisture in summer and break high-pressure blocking systems during winter.	Modulates 56.8% of winter, 39.2% of spring, 34.3% of summer, and 28.1% of autumn rainfall. Provides nearly 60% of winter rainfall in areas like Maria da Fé.	(REBOITA et al., 2015)
South American Monsoon System (SAMS)	Oct–Mar (Summer)	Seasonal reversal of winds due to continental surface heating. The Low-Level Jet (LLJ) transports moisture from the Amazon to southeastern Brazil.	Defines the rainy season. Over 50% of the total annual precipitation occurs during this period in monsoon-affected regions.	(REBOITA et al., 2015)
Local Convection (Orographic Effect)	Oct–Mar (Spring/Summer)	Surface heating and convective activity enhanced by the rugged mountainous relief (e.g., Mantiqueira Range), which forces the ascent of warm, moist air.	Triggers short-duration, high-intensity rainfall. Combined with the SACZ and SAMS, summer accounts for up to 84% of total annual rainfall (Dec–Jan alone contribute 33%).	(Junqueira Junior et al., 2019; REBOITA et al., 2015)

Fonte: The author (2026).

Table 3.1 – Physical–hydrological properties of the Red Dystroferric Latosol in the RF-FESM and their relevance for ecohydrological modeling.

Property	Value / description	Hydrological relevance
Textural class	Very clayey (clay > 60%)	High water-holding capacity; high potential for subsurface recharge
Soil structure	Granular / microaggregated	High macroporosity; preferential infiltration even in clayey texture
Saturated hydraulic conductivity (K_s)	High (70–150 mm h^{-1} in A horizons)	Promotes rapid infiltration; reduces surface runoff
Temporal stability of SWC	High (temporal CV < 15%)	Dampening of climatic anomalies; maintenance of baseflow during droughts
Monitoring depth	5 layers (0–10, 10–20, 20–40, 40–60, 60–100 cm)	Characterizes vertical redistribution; underpins the water balance in Article 2
Diagnostic horizon	Latosolic B horizon (Bw)	Low-activity clays; high subsurface permeability; advanced weathering

Sources: (Junqueira Junior et al., 2019); (GUAUQUE-MELLADO et al., 2022).
Fonte: The author (2026).

structure creates abundant large voids through which water moves quickly, even in torrential rain. The result is a soil that infiltrates rapidly but retains water deeply.

This property is crucial for the water balance. When rainfall reaches the ground as throughfall or stemflow, it does not pond at the surface. Instead, it penetrates rapidly into the soil profile, where the clay matrix stores it like a subsurface reservoir. This subsurface buffer sustains streamflow and tree growth through the dry season—a resilience mechanism documented by Junior (2017), who found that soil water content remains remarkably stable (coefficient of variation less than 15%) despite strong seasonal cycles. For Article 2, this soil stability is relevant because it means that the amount of water effectively available in the root zone is less variable than one might expect, a pattern that our interception models must capture.

3.2.4 Phenology, and Ecohydrological Relevance

The vegetation at Matinha belongs to a specific category: Montane Semideciduous Seasonal Forest. "Semideciduous" means that between 20% and 50% of tree species shed their leaves seasonally—not because it is winter (this is the tropics), but because water becomes scarce.

When the dry season arrives (April to September), the soil water declines, air evaporative demand increases, and many trees respond by dropping leaves to conserve water. This is an adaptive strategy, but it has profound consequences for rainfall partitioning.

A full canopy intercepts more water than a sparse one. During the rainy season, when trees are in full leaf, the leaf area index (LAI—a measure of total leaf surface relative to ground area) reaches 4.5 square meters per square meter. By the end of the dry season, it drops to 3.6 square meters per square meter. This may seem like a small change in numbers, but in hydrological terms, it is substantial: a 20% reduction in canopy leaf area can alter how much water is intercepted and evaporated back to the atmosphere, changing the effective precipitation that reaches the soil.

Here lies a critical problem for traditional hydrological models. The Gash model and the Rutter model, workhorses of forest hydrology for decades, assume that canopy storage capacity is a constant parameter that does not change through the year. They cannot represent seasonally varying leaf cover. When applied to a semideciduous forest like Matinha, these models misrepresent the coupling between phenology and water partitioning. This failure is exactly what Article 2 addresses: by using machine learning algorithms with explicit phenological features, we can learn—rather than assume—how interception changes as the forest greens and browns with the seasons.

3.2.5 Uniqueness of the Experimental Infrastructure: A Comparative Perspective

One of the central arguments of this thesis rests on the existence of an experimental infrastructure that, by virtue of its combination of features, has no direct parallel in the Brazilian literature on tropical forest hydrology and is rare in the international context. To make this claim credible, it is essential to move beyond general statements and show, through concrete comparison, what actually distinguishes Matinha from other sites where forest hydrologists work.

How does Matinha compare internationally?

The most comprehensive global monitoring network is FLUXNET, which currently brings together more than 500 sites measuring carbon, water, and energy exchanges across all continents

(PASTORELLO et al., 2020). However, tropical ecosystems are markedly underrepresented: most sites are concentrated in temperate regions of the Northern Hemisphere, and those tropical sites that do exist typically measure ecosystem energy fluxes (via eddy covariance), not the fine-scale rainfall partitioning that this thesis addresses. Importantly, FLUXNET sites in the Brazilian Atlantic Forest are scarce and do not include systematic event-scale measurements of how much precipitation actually reaches the soil after passing through the canopy.

A second global resource is the CAMELS dataset (*Catchment Attributes and Meteorology for Large-sample Studies*). The original CAMELS databases covered only the contiguous United States and Chile, capturing watersheds in temperate and arid climates. More recently, the CAMELS-BR dataset was released, extending coverage to 897 Brazilian catchments and finally including tropical data. Yet even CAMELS-BR operates at a daily temporal resolution. For the study of rainfall partitioning and interception—processes that can change dramatically within a single storm event—daily data are too coarse.

In Europe, the ICP Forests network has collected detailed precipitation, throughfall, and deposition chemistry data in temperate forests for over three decades (POVEDA; MESA, 2000). This is an excellent model, but it is confined to a different climate type and does not address the specific challenges posed by semideciduous tropical phenology.

How does Matinha compare in Brazil?

Within Brazil, several long-term forest monitoring initiatives exist. The LTER-Brasil network (also known as PELD—Pesquisas Ecológicas de Longa Duração) includes sites across multiple biomes. INPA maintains large research projects in central Amazonia (such as the LBA experiment), and USP's Hydrology Laboratory has operated experimental forested basins in São Paulo state. Within the Atlantic Forest specifically, researchers at UFRRJ have conducted long-term work in the Fluminense region, and ICMBio manages the Serra dos Órgãos Biological Reserve with hydrological monitoring.

However, when we examine what each of these initiatives actually measures, a pattern emerges: none of them simultaneously provides all five of the following characteristics that Matinha possesses:

1. **Spatial density of throughfall measurement:** Matinha has 32 fixed-point rain gauges strategically placed throughout the understory to capture how the canopy's spatial structure

creates rainfall variability. Most other sites use between 1 and 10 gauges, which misses the within-canopy heterogeneity that matters for modeling.

2. **Species-specific stemflow monitoring:** At Matinha, stemflow is measured individually on 15 dominant tree species. This allows the model to learn how different tree architectures partition water differently. Most other sites either do not measure stemflow at all, or measure it in an undifferentiated way that pools all species together.

3. **Sub-hourly temporal resolution:** Matinha's data are collected at 10-minute intervals. This captures the fine structure of tropical convective storms, which often last only 1–2 hours and can generate sharp changes in rainfall intensity. Other Brazilian sites typically observe daily or at best hourly resolution, smoothing away the dynamical detail needed to model event-scale physics.

4. **Soil moisture monitoring across vertical depth:** Matinha has sensors at five distinct depths (0–10, 10–20, 20–40, 40–60, 60–100 cm), enabling tracking of water redistribution through the soil column. Many other sites lack soil moisture data entirely, or measure at only one or two depths.

5. **Multiannual continuity:** Matinha's monitoring spans from September 2012 to December 2023—11 years of unbroken records. This duration is necessary to capture both seasonal and interannual variability. While some Brazilian sites have long histories, few combine this duration with the spatial density and high temporal resolution described above.

What makes Matinha unique?

The **Matinha** forest fragment (6.35 hectares) sits on the campus of the Federal University of Lavras in the municipality of Lavras, Minas Gerais. Its coordinates (21°13'40"S, 44°57'50"W) and elevation (925 meters above sea level) place it in the heart of the montane Atlantic Forest zone, making it representative of a landscape that supplies water to more than 125 million people but retains less than 15% of its original vegetation cover.

The monitoring infrastructure at Matinha is physically simple but strategically comprehensive. A 22-meter micrometeorological tower continuously records precipitation, air temperature, humidity, and wind speed at 10-minute intervals. Below the canopy, 32 *Ville de Paris* rain gauges are positioned throughout the study area to measure throughfall—the water that successfully reaches the forest floor. Additionally, stemflow is collected from individual trees of 15 species: *Blepharocalyx salicifolius*, *Copaifera langsdorffii*, *Miconia willdenowii*, *Xylopia brasiliensis*, and others. Underground, volumetric soil water content is measured continuously at 32 locations and five depths, providing a three-dimensional picture of how water moves through the soil after each rainfall event.

The result is a dataset that spans 11 years (2012–2023) of continuous, high-frequency observations. This is long enough to encompass multiple annual cycles, short-term droughts, and extreme storms. It is detailed enough to resolve individual rainstorms at the scale of minutes, not days. And it is spatially explicit enough to capture how forest heterogeneity—differences in species composition, canopy structure, and soil properties across a small area—influences the partitioning of water.

This combination has no parallel for semideciduous Atlantic Forest. It is this singularity that justifies the focus of the thesis on this specific location and that provides sufficient statistical power to develop and validate models capable of distinguishing subtle physical differences in how rainfall becomes effective precipitation.

The characterization and comparison of ecohydrological monitoring infrastructure across different research sites is essential for assessing the representativeness and scope of the data produced, as well as for positioning a study site within the international state of the art. Table 3.2 systematically compiles the main methodological features of monitoring sites in tropical and subtropical forests worldwide, ranging from automated regional networks to local-scale plot experiments. The attributes selected for comparison—monitoring duration, number of throughfall (Tf) measurement points, species-level stemflow (Sf) measurements, temporal resolution of observations, and availability of soil-moisture data—were chosen because they represent the most critical determinants of the robustness and applicability of the resulting datasets (Junqueira Junior et al., 2019; RODRIGUES et al., 2021a).

The comparative analysis shows that the vast majority of studies reported in the literature exhibit limitations in at least one of these attributes, whether due to short monitoring periods, a reduced number of sampling points, absence of species-specific Sf differentiation, or lack of integration with soil-moisture observations. In this context, the RF-FESM (Matinha–UFLA) experimental site stands out by uniquely combining, among the evaluated sites, long temporal continuity (11 years), high sampling-point density, sub-hourly resolution, stemflow monitoring across 15 distinct species, and simultaneous soil-moisture measurements at five depths (Junqueira Junior et al., 2017; Junqueira Junior et al., 2019; GUAUQUE-MELLADO et al., 2022). Consequently, it can be considered one of the most comprehensive reference sites for the study of ecohydrology in montane seasonal semideciduous Atlantic Forest ecosystems.

Table 3.2 – Comparison of ecohydrological monitoring infrastructure at the RF-FESM (Matinha–UFLA) and national and international reference sites in tropical and subtropical forests. Tf = throughfall sampling points; Sf = stemflow; Temporal Res. = measurement temporal resolution; SM = soil moisture.

Biome / Location	Duration	Monitoring Points (Tf)	Sf	Temporal Res.	SM	Ref.
Montane Semideciduous Atlantic Forest (Matinha – UFLA, MG, Brazil)	**11 yr (2012–2023)**	**32 fixed *Ville de Paris* rain gauges**	**Yes (15 spp.)**	**Hourly / Daily / Event**	**Yes (32 pts, 5 depths)**	**(Junqueira Junior et al., 2017; GUAUQUE-MELLADO et al., 2026)**
Atlantic Forest – Brazil						
Semi-deciduous Atlantic Forest (SP, Brazil)	7 mo (2009–2010)	10 fixed plastic collectors	Yes (1.0%)	Weekly	No	(TONELLO; FILHO, 2014)
Dense Atlantic Rainforest (SP, Brazil)	1 yr (2007–2008)	16 fixed gauges; 38 instrumented trees	Yes (38 trees, 0.2%)	Weekly	No	(ARCOVA; CICCO; ROCHA, 2003)
Upper Cloud Rainforest (MG, Brazil)	2 yr (2009–2011)	23 fixed *Ville de Paris* gauges	NR	Daily	No	(ÁVILA et al., 2014)

(Continued on next page)

(Table 3.2 continued)

Biome / Location	Duration	Monitoring Points (Tf)	Sf	Temporal Res.	SM	Ref.
Seasonal Deciduous Atlantic Forest & *Eucalyptus* plantation (Brazil)	6 yr (2009–2015)	3 plots with 20–40 roving Tf collectors each	Yes (3 trees/plot)	Event-based	No	(SARI; PAIVA; PAIVA, 2016)
Caatinga / Seasonally Dry Tropical Forest – Brazil						
Seasonally dry tropical forest / Regenerating Caatinga (NE Brazil)	8 yr (2010–2017)	10 Tf gauges at 1 site	Yes (6 trees)	5 min & event	No	(BRASIL et al., 2018)
Northeast Atlantic Forest / Caatinga (NE Brazil)	1 yr (2017)	1 plot (20×10 m) with 10 gauges	Yes (6 trees)	5 min & event	No	(FILHO et al., 2019)
Successional tropical dry forest (Costa Rica & Brazil)	1 rainy season	9 plots (CR) and 9 plots (BR)	Yes (6–12/stage)	Daily / Event	No	(CALVO-ALVARADO et al., 2018)
Amazon Tropical Rainforest – Brazil						
Lowland Amazon Tropical Rainforest (AM, Brazil)	1 yr (2004)	25 fixed plastic collectors	Yes (1.7%)	Weekly	No	(OLIVEIRA et al., 2008)

(Continued on next page)

(Table 3.2 continued)

Biome / Location	**Duration**	**Monitoring Points (Tf)**	**Sf**	**Temporal Res.**	**SM**	**Ref.**
Open Tropical Rainforest with Palms (AM, Brazil)	5 wk (2005)	20 collectors in 1.37 ha catchment	Yes (48 trees/palms)	5 min & event	No	(GERMER; ELSENBEER; MORAES, 2006)
Plantations and Forest Restoration						
Rocky mountainous area – *Pinus tabulaeformis* plantation (China)	3 yr (2009–2011)	1 plot (25×30 m) with 30 fixed containers	Yes (6 trees)	Event-based	No	(WEI et al., 2017)
Tropical species for forest restoration (Brazil)	3 mo (2006)	12 gauges per stand	Yes (0.7–2.5%)	Daily	No	(CARLYLE-MOSES; LISHMAN; MCKEE, 2014)
Tropical Montane and Cloud Forests – Caribbean, Mexico & Central America						
Lowland Tropical Rainforest (Luquillo, Puerto Rico)	3 mo (1996)	20 non-roving collectors	Yes (2.3%)	Weekly	No	(Junqueira Junior et al., 2019)
Mature & Secondary Montane Cloud Forest (Veracruz, Mexico)	∼1 yr (2006–2008)	7 troughs; 10 instrumented trees	Yes (1.0–4.3%)	Not specified	No	(Junqueira Junior et al., 2019)

(Continued on next page)

(Table 3.2 continued)

Biome / Location	**Duration**	**Monitoring Points (Tf)**	**Sf**	**Temporal Res.**	**SM**	**Ref.**
Tropical Montane Cloud Forest (Sierra de las Minas, Guatemala)	11 yr (1995–2006)	58 & 36 fixed rain gauges	NR	Weekly	No	(Junqueira Junior et al., 2019)
Tropical Forests – Panama						
Secondary Lower Montane Forest (Panama)	14 mo (2007–2008)	30 plastic funnels; 14 trees	Yes (1.3%)	Weekly	No	(Junqueira Junior et al., 2019)
Seasonal tropical forests – Agua Salud (Panama)	Ongoing since 2008	13 instrumented experimental catchments	NR	15 min	No	(REGINA et al., 2021)
Regional Monitoring Networks – Brazil						
Multiple biomes across Brazil – automated gauge network	7 yr (2014–2021)	∼3500 automated rain gauges	No	10–60 min	No	(LEMOS et al., 2023)
West Africa						
Moist semi-deciduous & evergreen forest (Ghana)	2 yr (1997–1999)	Trees across 3 topographic positions	NR	Daily (rain); 30 min (SM)	Yes (matric pot., 15 cm)	(BAKER; BURSLEM; SWAINE, 2003)

West African ecoclimatic gradient – Sudanian & Sahelian zones	>25 yr (since 1990)	4 mesoscale sites; 290 stations total	NR	Min. to daily	Yes (up to 1 m depth)	(GALLE et al., 2018)

Notes: NR = Not Reported; Tf = Throughfall; Sf = Stemflow; SM = Soil Moisture; spp. = tree species monitored individually for stemflow; mo = months; yr = year(s); wk = weeks. Stemflow values in parentheses indicate the percentage of gross rainfall as reported in the original study. The Matinha–UFLA row (shaded) is the reference site of the present work; its full monitoring period spans 2012–2023, with earlier campaigns published in (Junqueira Junior et al., 2017) and (Junqueira Junior et al., 2019).

Beyond the duration and density of monitoring, the UFLA experimental area is distinguished by explicitly documenting exceptional climatic anomalies within the 2012–2023 series, including the prolonged droughts of 2014–2015 and 2019–2020. These episodes substantially altered rainfall partitioning dynamics and constitute a natural laboratory for testing machine learning models under conditions of extreme atmospheric stress (GUAUQUE-MELLADO et al., 2022; RODRIGUES et al., 2021a).

3.3 Database and Monitoring Infrastructure

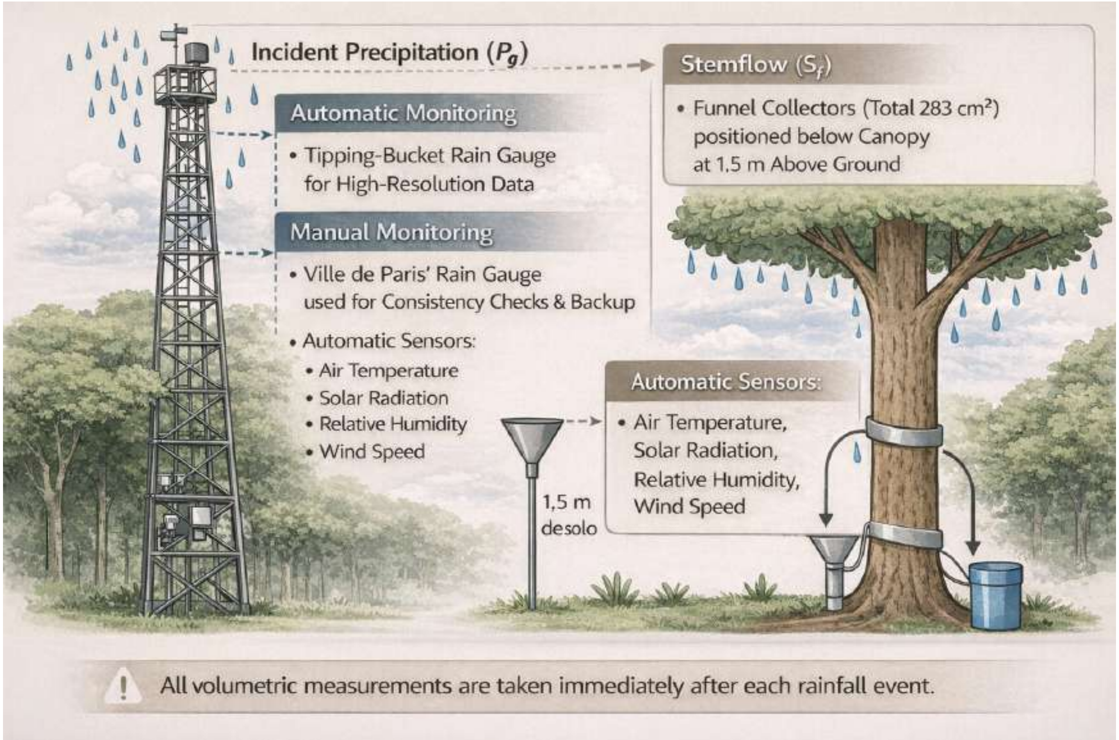
The database supporting both articles consists of high-resolution time series collected between September 2012 and the present period, spanning multiple hydrological cycles with substantial climatic variability. Table 3.2 summarizes the monitoring infrastructure and all external products employed, while Figure 3.2 illustrates the integrated field measurement system. The monitoring network quantifies the complete canopy water balance, including incident precipitation, throughfall, stemflow, soil water content, and micrometeorological drivers. The following subsections describe each component separately.

3.3.1 Monitoring of Incident Precipitation

Incident precipitation (P_g), also referred to as gross rainfall or above-canopy precipitation, is continuously monitored at the top of a 22 m Meteorological Observation Tower (MOT), installed above the forest canopy to avoid vegetation interference (Junqueira Junior et al., 2019). The tower hosts the primary precipitation instrumentation and provides the reference input to the canopy water balance (Figure 3.2).

Automatic monitoring is performed with a tipping-bucket rain gauge with a resolution of 0.20 mm per tip, recording data at 10-minute intervals and allowing derivation of hourly intensities and daily totals.

Figure 3.2 – Monitoring infrastructure of the forest hydrological observatory. Incident precipitation (P_g) is measured above the canopy at a 22 m meteorological observation tower using an automatic tipping-bucket rain gauge and a manual "Ville de Paris" rain gauge for consistency verification. Throughfall (T_f) is collected below the canopy using distributed collectors, and stemflow (S_f) is measured with collar collectors installed around representative trees. All volumetric measurements are performed immediately after rainfall events.



Fonte: The author (2026).

Box 3.2 – Monitoring infrastructure and auxiliary data sources used in this study.

Variable	Measurement device / Source	Temporal resolution	Spatial representation
Incident precipitation (P_g)	Tipping-bucket rain gauge (Campbell CR10X, 0.20 mm)	10 min / Hourly / Daily	22 m meteorological tower, above canopy
Incident precipitation (P_g) – control	"Ville de Paris" rain gauge	Daily / Event	Tower platform manual verification
Throughfall (T_f)	32 fixed rain gauges (Ville de Paris type)	Event / Daily	Systematic grid, 6.35 ha experimental plot
Stemflow (S_f)	Collar collectors and storage containers	Event / Daily	Representative trees across DBH classes
Meteorological variables	HOBO and CR10X dataloggers	10 min / Hourly	Tower micrometeorological measurements
Soil water content	PR2/6 capacitance probe	Event / Monthly	0–100 cm soil profile (5 depths, 32 points)
ERA5 / ERA5-Land	ECMWF reanalysis	Hourly / Daily	0.25° and 0.1° grid cells
CHIRPS v2.0	Satellite–gauge blended product	Daily	0.05° spatial resolution
PERSIANN-CDR	Satellite neural network product	Daily	0.25° spatial resolution
INMET (Lavras 83023)	Conventional weather station	6 h / Daily	Reference station for bias correction

Fonte: The author (2026).

This temporal resolution is required for Paper 1, which addresses hourly rainfall reconstruction, and for Paper 2, where event intensity acts as a central predictor in interception modeling.

Manual monitoring is conducted using a “Ville de Paris” rain gauge, read after each significant rainfall event. These manual measurements serve as an independent control for the automatic records and allow quality verification of accumulated totals during datalogger failures (RODRIGUES et al., 2021a).

The historical record spans from September 2012 to December 2023, providing more than a decade of continuous observations. Exploratory data analysis identified a high-continuity instrumental period between January 2018 and September 2022, which was selected as the training window for calibration of the rainfall imputation models in Paper 1. This period contains the lowest fraction of missing data and no instrumental failures longer than 30 consecutive days, ensuring seasonal representativeness and allowing the models to learn the patterns of convective rainfall variability.

3.3.2 Rainfall Partitioning: Throughfall and Stemflow

To quantify the ecohydrological interaction between rainfall and the forest canopy, Matinha maintains an event-scale rainfall-partitioning monitoring network that simultaneously measures throughfall (P_i) and stemflow (S_f). The governing mass balance is:

$$P_{\text{eff}} = P_i + S_f \tag{3.1}$$

$$I = P_g - P_{\text{eff}} \tag{3.2}$$

where P_{eff} is effective precipitation, I is forest interception, and P_g is incident precipitation.

Throughfall is measured by a network of 32 fixed rain gauges distributed systematically within the understory, with placement defined by criteria that ensure statistical representation of canopy spatial heterogeneity. The adequacy of sampling density was validated by Junqueira Junior et al. (2019), who found a relative standard error below 5% for event totals, confirming the robustness of mean estimates. This set of 32 points enables not only computation of mean throughfall for the fragment, but also exploration of intra-event spatial variability—a central aspect of Article 2, which uses point-level data as the independent unit of analysis.

Stemflow is collected from 32 representative trees across 15 species occurring in the area, selected to cover different diameter-at-breast-height (DBH) classes. Each tree is equipped with sealed closed-cell foam collars connected to collection tanks measured volumetrically immediately after each rainfall event (Junqueira Junior et al., 2019; RODRIGUES et al., 2021a). Stemflow, although on average a small fraction of total rainfall—typically between 0.5% and 3% in tropical forests (WU et al., 2024)—is highly localized at the base of each tree and plays a relevant role in near-root-zone subsurface recharge, being important for rigorous closure of the fragment water balance.

After all quality-control procedures—including application of physical constraints to the effective precipitation coefficient ($0 < C < 1.5$) and anomaly detection via Isolation Forest with a 2% contamination threshold—the Paper 2 database comprised 21,057 valid event–point observations, forming a dataset with seasonal and spatial representativeness adequate for training and validating machine learning models.

3.3.3 External Products: Reanalysis and Satellite Estimates

For precipitation-series reconstruction at daily and hourly scales (Article 1), local information from the meteorological tower is complemented by four global-coverage external products, which act as exogenous predictors in the machine learning models. These products provide large-scale synoptic context—including atmospheric humidity fields, water-vapor fluxes, and thermodynamic profiles—that is not captured by a point *in situ* measurement but is essential to reconstruct episodes during which the local sensor failed.

The ERA5 reanalysis, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), is the highest temporal-resolution product (hourly) used in this study, with spatial resolution of approximately 31 km (HERSBACH et al., 2020). ERA5-Land, a land-surface-focused version of the same system with approximately 9 km resolution, complements spatial coverage for daily-scale applications (MUÑOZ-SABATER et al., 2021). The CHIRPS v2.0 product (*Climate Hazards Group InfraRed Precipitation with Station data*) provides the highest spatial resolution among the products used ($\sim$5 km) and is produced by fusing infrared-based estimates with surface-station data (FUNK et al., 2015). PERSIANN-CDR (*Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks – Climate Data Record*) is based on neural networks trained on geostationary infrared data and calibrated with GPCP, with 0.25° coverage at the daily scale (ASHOURI et al., 2015).

All these products exhibit systematic biases relative to local observations —including the so-called *drizzle effect* in satellite products, which tends to underestimate extreme events and overestimate days with very low precipitation—which are corrected in the Article 1 workflow through quantile mapping (*Quantile Mapping*). In addition, data from the INMET conventional meteorological station in Lavras (Code 3023) are used as an auxiliary reference for bias-correction calibration, due to their greater climatic homogeneity and because they represent surface conditions of the same region (VIANA et al., 2023).

3.4 Link Between the Articles: Precipitation as a Continuous Process

The architecture of this thesis reflects a core methodological conviction: reconstructing precipitation and modeling its transformation into effective precipitation are not independent steps that could be addressed by separate studies. Rather, they are inseparable components of the same physical process, whose adequate representation requires that data integrity at the system input be treated with the same analytical rigor applied to modeling hydrological fluxes.

The dependence between the two articles is both operational and epistemic. Operationally, the hydrological memory features used in Article 2—in particular the Antecedent Precipitation Index (API) and rolling accumulations over multiple time windows—are computed from the hourly and daily precipitation series reconstructed in Article 1. A series with uncorrected instrumental errors or gaps filled by interpolations that underestimate convective variance would directly compromise representation of canopy moisture state, distorting interception models. Epistemically, Article 1 shows that machine learning algorithms, when trained on quality-controlled data and enriched with large-scale physical predictors, faithfully reproduce the convective variability intrinsic to tropical climates—exactly the variability that Article 2 must represent in order to capture the nonlinear relationships among rainfall intensity, canopy state, and interception efficiency.

Taken together, the two articles establish what this thesis terms an *Integrated Methodological Framework* for hydrology in seasonal forest ecosystems: a coherent workflow that begins with automated quality control and reconstruction of the raw rainfall series, proceeds through bias correction of exogenous products and multi-objective selection of the imputation model, and culminates in a physically interpretable estimation of effective precipitation at event scale and across the canopy spatial domain. This framework is not limited to the Matinha site: the modularity of components and the emphasis on reproducibility—with all scripts made available in a public version-controlled repository— make the framework transferable to other Atlantic Forest fragments and, with appropriate adaptations, to seasonal forest ecosystems in other tropical regions where data scarcity and precipitation nonlinearity pose challenges analogous to those addressed here.

REFERENCES

ABOUZIED, G. A. E. N. A. et al. Completion of the Central Italy daily precipitation instrumental data series from 1951 to 2019. *Geoscience Data Journal*, v. 12, n. 1, p. 1–18, 2025. ISSN 20496060. Doi: 10.1002/gdj3.267.

Agência Nacional de Águas e Saneamento Básico. *Conjuntura dos Recursos Hídricos no Brasil 2021*. Brasília: [*s.n.*], 2021.

AL-RAWAS, G. et al. Integrated quantile mapping and spatial clustering for robust bias correction of satellite precipitation in data-sparse regions. *Sustainability*, MDPI, v. 17, n. 18, p. 8321, 2025. Doi: 10.3390/su17188321.

ANDRADE, J. M. et al. Efficiency of global precipitation datasets in tropical and subtropical catchments revealed by large sampling hydrological modelling. *Journal of Hydrology*, 2024. Doi: 10.1016/j.jhydrol.2024.131016.

ANDRADE, J. M. de et al. A comprehensive assessment of precipitation products: Temporal and spatial analyses over terrestrial biomes in Northeastern Brazil. *Remote Sensing Applications: Society and Environment*, 2022. Doi: 10.1016/j.rsase.2022.100842.

ARCOVA, F. C. S.; CICCO, V. d.; ROCHA, P. A. B. Precipitação efetiva e interceptação das chuvas por floresta de Mata Atlântica em microbacia experimental do Núcleo Cunha, São Paulo. *Revista Árvore*, v. 27, n. 2, p. 257–262, 2003.

ASCENSO, G. et al. Downscaling, bias correction, and spatial adjustment of extreme tropical cyclone rainfall in ERA5 using deep learning. *Weather and Climate Extremes*, 2024. Doi: 10.1016/j.wace.2024.100724.

ASHOURI, H. et al. PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bulletin of the American Meteorological Society*, v. 96, n. 1, p. 69–83, 2015. ISSN 00030007. Doi: 10.1175/BAMS-D-13-00068.1.

BAGILIKO, J. et al. Bias correction of satellite and reanalysis products for daily rainfall occurrence and intensity. 2025. Preprint. arXiv:2510.27456v1.

BAKER, T. R.; BURSLEM, D. F. R. P.; SWAINE, M. D. Associations between tree growth, soil fertility and water availability at local and regional scales in ghanaian tropical rain forest. *Journal of Tropical Ecology*, Cambridge University Press, v. 19, n. 2, p. 109–125, 2003.

BAşAğAOğLU, H. et al. A Review on Interpretable and Explainable Artificial Intelligence in Hydroclimatic Applications. *Water (Switzerland)*, v. 14, n. 8, 2022. ISSN 20734441. Doi: 10.3390/w14081230.

BECK, H. E. et al. MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, v. 100, n. 3, p. 473–500, 2019. Doi: 10.1175/BAMS-D-17-0138.1.

BERGALLO, H. G. et al. Long-term ecological research: Chasing fashions or being prepared for fashion changes? *Anais da Academia Brasileira de Ciências*, SciELO Brasil, v. 95, n. 3, p. e20230051, 2023.

BEVEN, K. J. *Rainfall-Runoff Modelling: The Primer*. 2. ed. [*S.l.*]: Wiley-Blackwell, 2012.

BLEIDORN, M. T. et al. Investigation of using missing data imputation methodologies effect on the SARIMA model performance: application to average monthly flows. *Revista Brasileira de Recursos Hidricos*, v. 29, p. 1–18, 2024. ISSN 23180331. Doi: 10.1590/2318-0331.292420230131.

BOTREL, M. C. G.; CARVALHO, D. d. Variabilidade isoenzimática em populações naturais de jacarandá paulista (machaerium villosum vog.). *Revista Brasileira de Botânica*, SciELO Brasil, v. 27, n. 4, p. 621–627, 2004.

BOX, G. E. P.; JENKINS, G. M. *Time Series Analysis: Forecasting and Control*. [*S.l.*]: Holden-Day, 1970.

BOZORG-HADDAD, O. et al. Development of a Comparative Multiple Criteria Framework for Ranking Pareto Optimal Solutions of a Multiobjective Reservoir Operation Problem. *Journal of Irrigation and Drainage Engineering*, 2016. Doi: 10.1061/(ASCE)IR.1943-4774.0001028.

BRASIL, J. B. et al. Canopy Effects on Rainfall Partition and Throughfall Drop Size Distribution in a Tropical Dry Forest. *Atmosphere*, 2022. Doi: 10.3390/atmos13071126.

BRASIL, J. B. et al. Characteristics of precipitation and the process of interception in a seasonally dry tropical forest. *Journal of Hydrology: Regional Studies*, 2018. Doi: 10.1016/j.ejrh.2018.10.006.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Doi: 10.1023/A:1010933404324.

BREUNIG, M. M. et al. LOF: Identifying Density-Based Local Outliers. *SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, n. June, p. 93–104, 2000. Doi: 10.1145/342009.335388.

BRUIJNZEEL, L. A. Hydrological functions of tropical forests: not seeing the soil for the trees? *Agriculture, Ecosystems & Environment*, v. 104, p. 185–228, 2004.

BY, J.; GASH, C. An analytical model of rainfall interception by forests. p. 43–55, 1979. Doi: 10.1002/qj.49710544304.

CALVO-ALVARADO, J. C. et al. Interception of rainfall in successional tropical dry forests in brazil and costa rica. *Geosciences*, v. 8, n. 12, p. 486, 2018. Doi: 10.3390/geosciences8120486.

CANNON, A. J.; SOBIE, S. R.; MURDOCK, T. Q. Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? *Journal of Climate*, v. 28, n. 17, p. 6938–6959, 2015. Doi: 10.1175/JCLI-D-14-00754.1.

CARLYLE-MOSES, D. E.; LISHMAN, C. E.; MCKEE, A. J. A preliminary evaluation of throughfall sampling techniques in a mature coniferous plantation. *Journal of Forest Research*, v. 19, p. 271–282, 2014.

CARVALHO, L. M. V.; JONES, C.; LIEBMANN, B. The South Atlantic Convergence Zone: Intensity, form, persistence, and relationships with intraseasonal to interannual activity and extreme rainfall. *Journal of Climate*, v. 17, p. 88–108, 2004.

CAU, A. C. et al. Water Balance in an Atlantic Forest Remnant: Focus on Representative Tree Species. *Forests*, v. 16, n. 5, p. 1–18, 2025. ISSN 19994907. Doi: 10.3390/f16050812.

CHAGAS, V. B. et al. Camels-br: hydrometeorological time series and landscape attributes for 897 catchments in brazil. *Earth System Science Data*, Copernicus Publications, v. 12, p. 2075–2096, 2020.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM Computing Surveys*, v. 41, n. 3, p. 1–58, 2009.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [*S.l.*]: ACM, 2016. p. 785–794. Doi: 10.1145/2939672.2939785.

CLIFTON, D. A. et al. A review of novelty detection. *Signal Processing*, v. 99, p. 215–249, 2014.

CORDEIRO, C. A. et al. Long-term monitoring projects of brazilian marine and coastal ecosystems. *PeerJ*, PeerJ Inc., v. 10, p. e14313, 2022.

DEE, D. et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, v. 137, n. 656, p. 553–597, 2011. Doi: 10.1002/qj.828.

DEPIN, H. et al. Incertezas no Prenchimento De Falhas De Chuvas Horárias Com Redes Neurais Artificiais. p. 48–57, 2014. Doi: https://doi.org/10.7867/1983-1501.2013v15n2p48-57.

DIJK, A. I. J. M. V. et al. Rainfall interception and the coupled surface water and energy balance. *Agricultural and Forest Meteorology*, v. 214, p. 402–415, 2015. Doi: 10.1016/j.agrformet.2015.09.006.

DURRE, I. et al. Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, v. 49, n. 8, p. 1615–1633, 2010. Doi: 10.1175/2010JAMC2375.1.

EHRGOTT, M. *Multicriteria Optimization*. 2. ed. [*S.l.*]: Springer, 2005.

EISCHEID, J. K. et al. Creating a serially complete, national daily time series of temperature and precipitation for the Western United States. *Journal of Applied Meteorology*, v. 34, n. 12, p. 2927–2948, 1995. Doi: 10.1175/1520-0450(1995)034<2927:CASCND>2.0.CO;2.

FARAMARZZADEH, M. et al. Application of Machine Learning and Remote Sensing for Gap-filling Daily Precipitation Data of a Sparsely Gauged Basin in East Africa. *Environmental Processes*, v. 10, p. 8, 2023. Doi: 10.1007/s40710-023-00625-y.

FERREIRA, L. A. S. et al. Atributos da restauração florestal avaliados através de indicadores ecológicos em uma floresta estacional semidecidual. *Treedimensional*, 2018. Doi: 10.18677/Tree_Dimensional2018B1.

FILHO, J. C. R. et al. Partitioning of rainfall into throughfall , stemflow , and interception loss in the Brazilian Northeastern Atlantic Forest de interceptação vegetal no semiárido nordestino. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 23, n. 1807-1929, p. 21–26, 2019. Doi: http://dx.doi.org/10.1590/1807-1929/agriambi.v23n1p21-26 Partitioning.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, v. 29, p. 1189–1232, 2001.

FUNK, C. et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, v. 2, p. 150066, 2015. Doi: 10.1038/sdata.2015.66.

GALLE, S. et al. Amma-catch, a critical zone observatory in west africa monitoring a region in transition. *Vadose Zone Journal*, 2018. Doi: 10.2136/vzj2018.03.0062.

GASH, J. H. C.; LLOYD, C. R.; LACHAUD, G. Estimating sparse forest rainfall interception with an analytical model. *Journal of Hydrology*, v. 170, p. 79–86, 1995.

GASH, J. H. C.; MORTON, A. J. AN APPLICATION OF THE RUTTER MODEL TO THE ESTIMATION OF THE INTERCEPTION LOSS FROM THETFORD FOREST. *Journal of Hydrology*, v. 38, p. 49–58, 1978. Doi: https://doi.org/10.1016/0022-1694(78)90131-2.

GERMER, S.; ELSENBEER, H.; MORAES, J. M. Throughfall and temporal trends of rainfall redistribution in an open tropical rainforest, south-western amazonia (rondônia, brazil). *Hydrology and Earth System Sciences*, v. 10, n. 3, p. 383–393, 2006. Doi: 10.5194/hess-10-383-2006.

GUAUQUE-MELLADO, D. et al. Long-term water balance dynamics in a Brazilian neotropical forest: Insights from seven years of continuous observation. *Journal of Environmental Management*, v. 398, 2026. Doi: 10.1016/j.jenvman.2025.128475.

GUAUQUE-MELLADO, D. et al. Evapotranspiration under Drought Conditions: The Case Study of a Seasonally Dry Atlantic Forest. *Atmosphere*, 2022. Doi: 10.3390/atmos13060871.

GUDMUNDSSON, L. et al. Technical note: Downscaling RCM precipitation to the station scale using statistical transformations — a comparison of methods. *Hydrology and Earth System Sciences*, v. 16, p. 3383–3390, 2012.

GUPTA, H. V. et al. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, v. 377, n. 1–2, p. 80–91, 2009. Doi: 10.1016/j.jhydrol.2009.08.003.

GUPTA, H. V.; SOROOSHIAN, S.; YAPO, P. O. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, v. 34, n. 4, p. 751–763, 1998. Doi: 10.1029/97WR03495.

HAMID, H.; SAMANTARAY, S. Discharge Patterns and Trends in the Jhelum River Basin: A Statistical and Spatial Analysis. *Journal of The Institution of Engineers (India): Series A*, 2025. Doi: 10.1007/s40030-025-00913-w.

HAUGHTON, N. et al. Does predictability of fluxes vary between fluxnet sites? *Biogeosciences*, Copernicus Publications, v. 15, n. 14, p. 4495–4513, 2018.

HAWKINS, D. M. *Identification of Outliers*. London: Chapman and Hall, 1980.

HERRERA, R. V.; FOWLER, H. J. The creation and climatology of a large independent rainfall event database for Great Britain. n. October 2022, p. 6020–6037, 2023. Doi: 10.1002/joc.8187.

HERSBACH, H. et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, v. 146, n. 730, p. 1999–2049, 2020. Doi: 10.1002/qj.3803.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997. ISSN 08997667. Doi: 10.1162/neco.1997.9.8.1735.

HUFFMAN, G. J. et al. Integrated multi-satellite retrievals for the global precipitation measurement (GPM IMERG). *Satellite Precipitation Measurement*, 2020.

IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. [*S.l.*]: Cambridge University Press, 2021. Doi: 10.1017/9781009157896.

JONES, C.; CARVALHO, L. M. V. The South Atlantic convergence zone: observations and future changes in a warming climate. *Climate Dynamics*, 2019.

JOURNEL, A. G.; HUIJBREGTS, C. J. *Mining Geostatistics*. London: Academic Press, 1978.

Junqueira Junior, J. A. et al. Time-stability of soil water content (SWC) in an Atlantic Forest - Latosol site. *Geoderma*, 2017. Doi: 10.1016/j.geoderma.2016.10.034.

Junqueira Junior, J. A. et al. Rainfall partitioning measurement and rainfall interception modelling in a tropical semi-deciduous Atlantic forest remnant. *Agricultural and Forest Meteorology*, Elsevier, v. 275, n. May, p. 170–183, 2019. ISSN 0168-1923. Doi: 10.1016/j.agrformet.2019.05.016. Disponível em: https://doi.org/10.1016/j.agrformet.2019.05.016.

KE, G. et al. LightGBM : A Highly Efficient Gradient Boosting Decision Tree. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. [*s.n.*], 2017. p. 1–9. ISBN 9781510860964. Doi: 10.5555/3294996.3295074. Disponível em: https://dl.acm.org/doi/10.5555/3294996.3295074.

KEIM, R. F.; SKAUGSET, A. E.; WEILER, M. Storage of water on vegetation under simulated rainfall of varying intensity. *Advances in Water Resources*, v. 29, p. 974–986, 2006.

KIDD, C.; LEVIZZANI, V. Status of satellite precipitation retrievals. *Hydrology and Earth System Sciences*, v. 15, n. 4, p. 1109–1116, 2011. Doi: 10.5194/hess-15-1109-2011.

KIM, J.-W.; PACHEPSKY, Y. A. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology*, v. 394, p. 305–314, 2010.

KIM, Y. et al. An Effective Algorithm of Outlier Correction in Space-Time Radar Rainfall Data Based on the Iterative Localized Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 62, p. 1–16, 2024. ISSN 15580644. Doi: 10.1109/TGRS.2024.3366400.

KLING, H.; FUCHS, M.; PAULIN, M. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, v. 424-425, p. 264–277, 2012.

KNOBEN, W. J. M.; FREER, J. E.; WOODS, R. A. Technical note: Inherent benchmark or not? comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, v. 23, p. 4323–4331, 2019.

KRATZERT, F. et al. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, v. 55, p. 11344–11354, 2019.

KRATZERT, F. et al. Towards learning universal, regional, and local hydrological behaviors via machine-learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, v. 23, p. 5089–5110, 2019.

LEGATES, D. R.; MCCABE, G. J. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, v. 35, n. 1, p. 233–241, 1999. Doi: 10.1029/1998WR900018.

LEMOS, F. C. et al. Spatiotemporal distribution of precipitation and its characteristics under tropical atmospheric systems of Brazil: Insights from a large sub-hourly database. *Hydrological Processes*, v. 37, n. 11, p. 1–19, 2023. ISSN 10991085. Doi: 10.1002/hyp.15017.

LEVIA, D. F.; CARLYLE-MOSES, D. E.; TANAKA, T. Forest hydrology and biogeochemistry: Synthesis of past research and future directions. *Ecological Studies*, v. 216, 2011.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: IEEE. *2008 Eighth IEEE International Conference on Data Mining (ICDM)*. [*S.l.*], 2008. p. 413–422. Doi: 10.1109/ICDM.2008.17.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *NeurIPS 2017*. [*S.l.: s.n.*], 2017. p. 4765–4774.

LY, S.; CHARLES, C.; DEGRÉ, A. Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium. *Hydrology and Earth System Sciences*, v. 15, n. 7, p. 2259–2274, 2011. Doi: 10.5194/hess-15-2259-2011.

MARAUN, D. Bias correcting climate change simulations – a critical review. *Current Climate Change Reports*, v. 2, n. 4, p. 211–220, 2016. Doi: 10.1007/s40641-016-0050-x.

MATHERON, G. Principles of geostatistics. *Economic Geology*, v. 58, n. 8, p. 1246–1266, 1963. Doi: 10.2113/gsecongeo.58.8.1246.

MELLO, C. R. D. et al. Deciphering global patterns of forest canopy rainfall interception (FCRI): A synthesis of geographical , forest species , and methodological influences. v. 358, n. March, 2024.

MENNE, M. J. et al. An overview of the global historical climatology network. *Journal of Atmospheric and Oceanic Technology*, v. 29, p. 897–910, 2012.

MICHILES, A. A. S.; GIELOW, R. Above-ground thermal energy storage rates, trunk heat fluxes and surface energy balance in a central amazonian rainforest. *Agricultural and Forest Meteorology*, v. 148, p. 917–930, 2008. Doi: 10.1016/j.agrformet.2008.01.001.

MIRALLES, D. G. et al. Global canopy interception from satellite observations. *Journal of Geophysical Research Atmospheres*, v. 115, n. 16, p. 1–8, 2010. ISSN 01480227. Doi: 10.1029/2009JD013530.

MOHAMMED, J.; MENGISTE, Y.; GEBREMICHAEL, M. Enhancing monthly precipitation forecasting by integrating multi - source data with machine learning models : a study in the Upper Blue Nile Basin. *Modeling Earth Systems and Environment*, Springer International Publishing, v. 11, n. 1, p. 1–16, 2025. ISSN 2363-6211. Doi: 10.1007/s40808-024-02223-9. Disponível em: https://doi.org/10.1007/s40808-024-02223-9.

MOKHTARI, S.; SHARAFATI, A.; RAZIEI, T. Satellite-based streamflow simulation using CHIRPS satellite precipitation product in Shah Bahram Basin, Iran. *Acta Geophysica*, 2022. Doi: 10.1007/s11600-021-00724-0.

MONEGO, V.; ANOCHI, J.; VELHO, H. C. Machine Learning for Uncertainty Quantification in Climate Precipitation Prediction Over South America. *Coordination of Applied Research and Technological Development*, p. 0–29, 2023. Doi: https://doi.org/10.21203/rs.3.rs-3310465/v1. Disponível em: https://www.researchsquare.com/article/rs-3310465/v1.

MONEGO, V. S.; ANOCHI, J. A.; de Campos Velho, H. F. South America Seasonal Precipitation Prediction by Gradient-Boosting Machine-Learning Approach. *Atmosphere*, v. 13, n. 2, 2022. ISSN 20734433. Doi: 10.3390/atmos13020243.

MUÑOZ-SABATER, J. et al. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 2021. Doi: 10.5194/essd-13-4349-2021.

MUZYLO, A. et al. A review of rainfall interception modelling. *Journal of Hydrology*, Elsevier B.V., v. 370, n. 1-4, p. 191–206, 2009. ISSN 0022-1694. Doi: 10.1016/j.jhydrol.2009.02.058. Disponível em: http://dx.doi.org/10.1016/j.jhydrol.2009.02.058.

NASH, J. E.; SUTCLIFFE, J. V. River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, v. 10, n. 3, p. 282–290, 1970. ISSN 00221694. Doi: 10.1016/0022-1694(70)90255-6.

NATIONS, U. *Sustainable Development Goals*. 2015. Acessado em janeiro de 2026. Disponível em: https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/.

NIELSEN, D. M. et al. Mechanisms governing the south atlantic convergence zone over subtropical South America and its relationship with precipitation. *Climate Dynamics*, v. 53, p. 2953–2970, 2019.

OLIVEIRA, L. L. d. et al. Partição das chuvas em uma floresta estacional semidecidual no Sudeste do Brasil. *Revista Árvore*, v. 32, p. 577–584, 2008. TODO: verify complete bibliographic data.

OLIVEIRA, V. A. de et al. Spatiotemporal modelling of soil moisture in an Atlantic forest through machine learning algorithms. *European Journal of Soil Science*, v. 72, n. 5, p. 1969–1987, 2021. ISSN 13652389. Doi: 10.1111/ejss.13123.

PAN, X. et al. Bias corrections of precipitation measurements across experimental sites in different ecoclimatic regions of western Canada. p. 2347–2360, 2016. Doi: 10.5194/tc-10-2347-2016.

PANOFSKY, H. A.; BRIER, G. W. Some applications of statistics to meteorology. Pennsylvania State University Press, 1958.

PAREDES-TREJO, F. J.; BARBOSA, H. A.; Lakshmi Kumar, T. V. Validating CHIRPS-based satellite precipitation estimates in Northeast Brazil. *Journal of Arid Environments*, 2017. Doi: 10.1016/j.jaridenv.2016.12.009.

PASTORELLO, G. et al. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, v. 7, p. 225, 2020. Doi: 10.1038/s41597-020-0534-3.

PAULHUS, J.; KOHLER, M. Interpolation of missing precipitation records. *Monthly Weather Review*, v. 80, n. 8, p. 129–133, 1952. Doi: 10.1175/1520-0493(1952)080<0129:IOMPR>2.0.CO;2.

PINTHONG, S. et al. Optimized Rainfall Imputation Using ERA5-Land and Tree-Based Machine Learning: A Scalable Framework for Data-Sparse Regions. *Earth Systems and Environment*, 2025. Doi: 10.1007/s41748-025-00787-9.

POVEDA, G.; MESA, O. J. Effects of deforestation on the hydrological cycle in humid tropical regions: A study with a regional climate model. *Advances in Water Resources*, v. 23, p. 775–789, 2000. TODO: verify – cited as Poveda2005 in text re: ICP Forests monitoring network; check correct reference.

PROKHORENKOVA, L. et al. CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*. [*S.l.*]: Curran Associates, Inc., 2018. v. 31. Doi: 10.48550/arXiv.1706.09516.

QUADRO, M. F. L. d. Estudo de episódios de Zona de Convergência do Atlântico Sul (ZCAS) sobre a América do Sul. 1994.

REBOITA, M. S. et al. Aspectos climáticos do estado de minas gerais. *Revista Brasileira de Climatologia*, 2015. Este archivo corresponde a anexos o copias del estudio de 128.pdf.

REGINA, J. A. et al. Agua salud project experimental catchments hydrometric data, panama. *Hydrological Processes*, 2021. Doi: 10.1002/hyp.14359.

ROBERTS, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, v. 40, n. 8, p. 913–929, 2017.

ROCHA, C.; CERQUEIRA, P. et al. Forest coverage and streamflow of watersheds in the tropical Atlantic rainforest. *Revista Árvore*, v. 42, n. 2, 2018.

RODRIGUES, A. F. et al. Modeling canopy interception under drought conditions : The relevance of evaporation and extra sources of energy. *Journal of Environmental Management*, v. 292, n. April, 2021. Doi: 10.1016/j.jenvman.2021.112710.

RODRIGUES, A. F. et al. Soil water content and net precipitation spatial variability in an Atlantic forest remnant. *Acta Scientiarum - Agronomy*, v. 42, p. 1–13, 2020. ISSN 18078621. Doi: 10.4025/actasciagron.v42i1.43518.

RUBIN, D. B. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987. Doi: 10.1002/9780470316696.

RUTTER, A. J.; MORTON, A. J.; ROBINS, P. C. A predictive model of rainfall interception in forests. II. generalization of the model and comparison with observations in some coniferous and hardwood stands. *Journal of Applied Ecology*, v. 12, p. 367–380, 1975.

SALAZAR, Á. et al. Spatio-Temporal Evaluation of MSWEP, CHIRPS and ERA5-Land Reveals Regional-Specific Responses Across Complex Topography in Bolivia. *Atmosphere*, 2025. Doi: 10.3390/atmos16111281.

SANTOS, P. S. D. et al. OPSBC : A method to sort Pareto-optimal sets of solutions in multi-objective problems. *Expert Systems With Applications*, Elsevier Ltd, v. 250, n. February, p. 123803, 2024. ISSN 0957-4174. Doi: 10.1016/j.eswa.2024.123803. Disponível em: https://doi.org/10.1016/j.eswa.2024.123803.

SANTOS, V.; DURÁN-QUESADA, A. M.; SÁNCHEZ-MURILLO, R. Day-night differences in δ 18O and d-excess of convective rainfall in the inland tropics of Brazil. *Research Square*, p. 1–12, 2023. Doi: doi.org/10.21203/rs.3.rs-2409508/v1 License:.

SARI, V.; PAIVA, E. M. C. D. de; PAIVA, J. B. D. de. Interceptação da chuva em diferentes formações florestais na região sul do Brasil. *Revista Brasileira de Recursos Hídricos*, v. 21, p. 65–79, 2016. Doi: 10.21168/rbrh.v21n1.p65-79.

SCHMIDT, L. et al. Challenges in Applying Machine Learning Models for Hydrological Inference : A Case Study for Flooding Events Across Germany. 2020. Doi: 10.1029/2019WR025924.

SEARCY, J. K.; HARDISON, C. H. *Double-mass curves*. Washington, D.C., 1960.

SEGOVIA-CARDOZO, D. A.; BERNAL-BASURCO, C.; RODRÍGUEZ-SINOBAS, L. Tipping bucket rain gauges in hydrological research: Summary on measurement uncertainties, calibration, and error reduction strategies. *Sensors*, 2023. Doi: 10.3390/s23125385.

SHEN, C. A trans-disciplinary review of deep learning research for water resources scientists. *Water Resources Research*, v. 54, n. 11, p. 8558–8593, 2018. Doi: 10.1029/2018WR022643.

SHEN, H.; TOLSON, B. A.; MAI, J. Time to Update the Split-Sample Approach in Hydrological Model Calibration. *Water Resources Research*, 2022. Doi: 10.1029/2021WR031523.

SILVA, J. C. d. M. e.; COSTA, D. d. A.; CHRISTOFFERSEN, M. L. Favorable Impacts of a Fragment of Atlantic Forest on Water Quality in an Urban River in Northeastern Brazil. *Environmental Smoke*, v. 2, n. 2, p. 23–35, 2019. Doi: 10.32435/envsmoke.20192223-35.

SILVA, V. O.; MELLO, C. R. Meteorological droughts in part of southeastern brazil: Understanding the last 100 years. *Anais da Academia Brasileira de Ciências*, SciELO Brasil, v. 93, n. suppl 4, p. e20201130, 2021. Doi: 10.1590/0001-3765202120201130.

SINGH, V. et al. Data assimilation with machine learning for constructing gridded rainfall time series data to assess long-term rainfall changes in the northeastern regions in India. v. 15, n. 6, p. 2687–2713, 2024. Doi: 10.2166/wcc.2024.644.

SOS Mata Atlântica; INPE. *Atlas dos Remanescentes Florestais da Mata Atlântica*. 2021.

STEKHOVEN, D. J.; BÜHLMANN, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, v. 28, n. 1, p. 112–118, 2012. Doi: 10.1093/bioinformatics/btr597.

SUNILKUMAR, K. et al. Evaluation of TRMM multi-satellite precipitation analysis (TMPA) in four distinct climate zones. *Climate Dynamics*, 2019.

TEEGAVARAPU, R. S. V. Spatial interpolation using artificial neural networks. *Advances in Water Resources*, v. 57, p. 13–25, 2012.

TIAN, Y. et al. Systematic anomalies over inland water bodies in satellite-based precipitation estimates. *Geophysical Research Letters*, v. 34, p. L14403, 2007.

TONELLO, K. C.; FILHO, J. T. Interception and stemflow in native forest and reforestation areas in the state of São Paulo, Brazil. *Scientia Forestalis*, v. 42, n. 102, p. 259–270, 2014. TODO: verify complete bibliographic data.

VALENTE, F.; DAVID, S. Modelling interception loss for two sparse eucalypt and pine forests in central Portugal using reformulated Rutter and Gash analytical models. v. 190, p. 141–162, 1997.

VÉLIZ-CHÁVEZ, C. et al. Canopy Storage Implications on Interception Loss Modeling. *American Journal of Plant Sciences*, n. September, p. 3032–3048, 2014. Doi: dx.doi.org/10.4236/ajps.2014.520320 Canopy.

VERDAN, I. ZONA DE CONVERGÊNCIA DO ATLÂNTICO SUL : UMA REVISÃO SISTEMÁTICA E DE DISCURSO South Atlantic Convergence Zone : a systematic and discourse review Zona de Convergencia del Atlántico Sur : una revisión sistemática y del discurso. p. 313–340, 2023.

VIANA, L. B. et al. Análise comparativa entre dados de precipitação gerados pelo "Climate Prediction Center – CPC" versus dados observados para diferentes biomas no Brasil. *Ciência e Natura*, v. 45, n. esp. 2, p. e81776, 2023. ISSN 0100-8307. Doi: 10.5902/2179460x81776.

VILA, D. A. et al. MERGE: A methodology for merging verified satellite and surface observations. *Meteorological Applications*, 2009.

VUICHARD, N.; PAPALE, D. Filling the gaps in meteorological continuous data measured at fluxnet sites with era-interim reanalysis. *Earth System Science Data*, Copernicus Publications, v. 7, n. 2, p. 157–171, 2015.

WANG, F.; TIAN, D.; CARROLL, M. Customized deep learning for precipitation bias correction and downscaling. *Geoscientific Model Development*, 2023. Doi: 10.5194/gmd-16-535-2023.

WAQAS, M. MethodsX A critical review of RNN and LSTM variants in hydrological time series predictions. *MethodsX*, Elsevier B.V., v. 13, n. September, p. 102946, 2024. ISSN 2215-0161. Doi: 10.1016/j.mex.2024.102946. Disponível em: https://doi.org/10.1016/j.mex.2024.102946.

WEI, X. et al. Rainfall partitioning in a typical plantation of northern China: effects of stand age and rainfall characteristics. *Ecohydrology*, v. 10, p. e1886, 2017. TODO: verify – cited for Pinus tabulaeformis plantation (China).

WILLARD, J. D. et al. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, v. 55, n. 4, p. 1–35, 2022.

World Meteorological Organization. *Guide to Hydrological Practices: Volume I — Hydrology From Measurement to Hydrological Information*. 2008.

WU, Q. et al. Responses of rainfall partitioning to water conditions in Chinese forests. *Journal of Hydrology*, v. 637, n. May, 2024. Doi: 10.1016/j.jhydrol.2024.131410.

ZANDLER, H. et al. Evaluation needs and temporal performance differences of gridded precipitation products in peripheral mountain regions. *Scientific Reports*, 2019.

ZARGAR, M. et al. Comparison and hydrological evaluation of different precipitation data for a large tropical region: the Blue Nile Basin in Ethiopia. *Frontiers in Water*, 2025. Doi: 10.3389/frwa.2025.1536881.

ZILLI, M. T.; HART, N. C. G. Synoptic climatology and changes in precipitation associated with the South Atlantic Convergence Zone utilising a cloud band identification technique. *International Journal of Climatology*, 2019.

ÁVILA, L. F. et al. Spatio-temporal variability of rainfall erosivity in Brazil. *Scientia Agricola*, v. 71, n. 5, p. 367–379, 2014. TODO: verify – cited for Upper Cloud Rainforest (MG) interception study; check correct reference.

4 RECOVERING THE ATMOSPHERIC WATER INPUT TO A TROPICAL ATLANTIC FOREST: MACHINE LEARNING RECONSTRUCTION OF PRECIPITATION RECORDS WITH PHYSICALLY CONSISTENT REANALYSIS PREDICTORS AND PROCESS-ORIENTED MULTI-OBJECTIVE EVALUATION

Abstract

Precipitation is the primary atmospheric water input controlling hydrological processes in tropical ecosystems—soil moisture dynamics, streamflow generation, groundwater recharge, and forest water balance—yet monitoring networks in these regions suffer from systematic data gaps caused by sensor failures. Extended gaps in rainfall records prevent the closure of water balances, distort estimates of antecedent moisture conditions, and obscure the temporal structure of wet/dry cycles that govern eco-hydrological functioning. This study presents a reproducible multi-stage machine learning framework for daily and hourly rainfall reconstruction, validated at an Atlantic Forest research site in southeastern Brazil (2013–2024, 25.6% missing data). The pipeline integrates three components rarely combined in the tropical hydrology literature: unsupervised outlier detection (Isolation Forest and Local Outlier Factor), bias adjustment of global reanalysis and satellite products (ERA5, CHIRPS, PERSIANN-CDR) via Quantile Mapping, and multi-objective model selection based on Pareto–Borda ranking that simultaneously optimizes statistical precision and hydrological consistency. More than 150 configurations spanning statistical models (SARIMA, Prophet) and machine learning algorithms (XGBoost, Random Forest, LSTM) were evaluated. Isolation Forest combined with bias-corrected ERA5 predictors consistently dominated the Pareto frontier. The optimal configuration (XGBoost + ERA5 + Isolation Forest) reduced RMSE by 74.6% relative to the SARIMA benchmark (8.76 to 2.22 mm d^{-1}) and nearly doubled the Kling–Gupta Efficiency (KGE: 0.26 to 0.48), preserving the variance structure and event frequency needed for hydrological process modeling. At the hourly scale, a Bidirectional LSTM ($r = 0.87$) substantially outperformed Prophet (KGE $= -0.26$), confirming that recurrent architectures better capture sub-daily convective dynamics. ERA5 reanalysis, not satellite products, emerges as the dominant predictor—a finding with direct implications for understanding how large-scale atmospheric dynamics drive water input to tropical ecosystems and for reconstruction efforts across the data-scarce tropical Global South. The complete codebase is openly available (DOI: https://doi.org/10.5281/zenodo.18096940).

Keywords: rainfall reconstruction; machine learning; ERA5 reanalysis; Isolation Forest; multi-objective optimization; Quantile Mapping; LSTM; Atlantic Forest.

4.1 Introduction

In tropical ecosystems, the water cycle begins and ends with rainfall. Every soil moisture pulse, every streamflow event, every episode of groundwater recharge traces its origin to a precipitation record—and in the Atlantic Forest of southeastern Brazil, 25.6% of that record is missing. Precipitation governs soil moisture dynamics, partitions rainfall into canopy interception, surface runoff, and deep percolation, drives baseflow generation and groundwater recharge, and sets the atmospheric boundary condition for the entire water cycle in tropical forests (Junqueira Junior et al., 2017; OLIVEIRA et al., 2021; CAU et al., 2025). Understanding these processes—and building the models that represent them—requires continuous, high-quality rainfall records (ANDRADE et al., 2024; CHAKRI et al., 2025). When those records have gaps, every downstream estimate of evapotranspiration, soil moisture, or streamflow inherits an uncertainty that no modeling refinement can resolve.

The challenge is particularly acute in the Brazilian Atlantic Forest, a biome with steep topographic gradients (BORMA et al., 2022; MARQUES; GRELLE, 2021), intense convective seasonality, and sparse monitoring infrastructure (ANDRADE et al., 2024). One of the world's most threatened biodiversity hotspots, the Atlantic Forest retains less than 12% of its original cover, and the remaining forest is distributed in thousands of small, fragmented remnants (CARNEIRO et al., 2023; MARQUES; GRELLE, 2021; MYERS et al., 2000; RIBEIRO et al., 2009). In such landscapes dominated by agriculture and urban land use, small protected fragments are essential for maintaining regional biodiversity and ecosystem services (BORMA et al., 2022; CARNEIRO et al., 2023). Consequently, they concentrate the remaining opportunities to observe and model near-natural forest water cycles in the region (CAU et al., 2025; OLIVEIRA et al., 2021; Junqueira Junior et al., 2017). The *Reserva Florestal da Mata Atlântica* ("Matinha") is one of these remnants—a montane semideciduous forest fragment embedded in a highly transformed matrix—where continuous rainfall forcing is a prerequisite for any water balance study (OLIVEIRA et al., 2021; Junqueira Junior et al., 2017; MELLO et al., 2019; RODRIGUES et al., 2020; RODRIGUES et al., 2021a). Tropical precipitation is inherently difficult to measure and characterize: a high proportion of daily records show no rainfall, while isolated convective events deliver most of the annual total (BAO et al., 2024; LEMOS et al., 2023). These heavy-tailed, zero-inflated distributions mean that missing records are not random: they cluster around sensor overloads, datalogger failures during peak events, and prolonged wet seasons, precisely the periods most informative for hydrological process understanding (PUGAS, 2023; TEEGAVARAPU, 2014). At the global scale, recent assessments of the precipitation observing network show that station density and record completeness are particularly deficient in monsoon-affected regions and in much of the tropical Global South, exacerbating data scarcity in hydroclimatically complex areas such as the Brazilian highlands (BORMA et al., 2022; SURI; AZAD, 2024; SU; ZHANG et al., 2026). In this context, each missing day of rainfall is a missing piece of information about how atmospheric moisture interacts with the land surface—whether a storm recharged the soil profile, triggered shallow lateral flow, or drove baseflow in the following weeks (CAU et al., 2025; MELLO et al., 2019; RODRIGUES et al., 2021a).

Recovering this information requires methods that go beyond simple interpolation and can reconstruct the physical structure of precipitation events—their timing, duration, and variability—not just their average values (TEEGAVARAPU, 2014). Classical reconstruction approaches—multiple imputation, linear interpolation, and historical averaging—assume stationarity, linearity, and normality that tropical precipitation actively violates (ADDI et al., 2022; KUMAR; DWARAKISH, 2025; SULLIVAN; KELLY, 2024; BANJARA et al., 2025). Classical methods tend to smooth wet–dry variability and underrepresent high-intensity events (CINKUS et al., 2023; GUPTA et al., 2009), precisely the features that control soil saturation, runoff generation, and groundwater recharge in Atlantic Forest headwaters (OLIVEIRA et al., 2021; Junqueira Junior et al., 2017).

Global reanalysis and satellite products—ERA5, ERA5-Land, CHIRPS, PERSIANN-CDR—provide spatially continuous atmospheric information that can help fill these gaps (BEKIĆ; LESKOVAR, 2025; COUTINHO et al., 2025; FUNK et al., 2015; HERSBACH et al., 2020; PADULANO et al., 2025). ERA5, in particular, assimilates a three-dimensional atmospheric state vector and resolves synoptic and mesoscale circulation at ~31 km and hourly resolution, allowing the large-scale convective environment that drives precipitation to be characterized consistently in space and time (HERSBACH et al., 2020; MUÑOZ-SABATER et al., 2021; TASZAREK et al., 2021). However, these products also carry systematic biases and scale mismatches that must be corrected before they are useful at the local scale, especially in steep, heterogeneous terrain (AL-RAWAS et al., 2025; ANSARI; GROSSI, 2022; BEKIĆ; LESKOVAR, 2025; JIAO et al., 2021; MOKHTARI; SHARAFATI; RAZIEI, 2022). Quantile-based bias-correction methods are increasingly used to align reanalysis and satellite climatologies with local gauge observations while preserving key aspects of the climate signal, including extremes (ABOUZIED et al., 2025; CANNON; SOBIE; MURDOCK, 2015; LI et al., 2023; PADULANO et al., 2025).

Machine learning (ML) offers a more flexible path for exploiting these multi-source datasets (CHAKRI et al., 2025; SILVA et al., 2025; ISHIDA et al., 2024; KUMAR; DWARAKISH, 2025; MOREIRA et al., 2023; ZOUNEMAT-KERMANI et al., 2021). Yet the existing tropical rainfall reconstruction literature has two recurrent weaknesses that limit its hydrological usefulness. First, ML models are often trained on uncleaned gauge records and uncorrected satellite or reanalysis inputs, causing them to learn instrumental noise as physical signal and propagate it into the reconstructed series (ALSALEHY; BAILEY, 2025; AWE; VANCE, 2025; CHANDOLA; BANERJEE; KUMAR, 2009; CLIFTON et al., 2014; KIM; KIM; KOK, 2025). Second, model selection typically relies on a single error metric, which rewards predictions close to the mean and suppresses variance, producing statistically “safe” but hydrologically fragile reconstructions that attenuate wet–dry variability and fail to reproduce tail events (ARIAS-MUÑOZ et al., 2023; CINKUS et al., 2023; DUBEY et al., 2025; GUPTA et al., 2009;

HODSON, 2022; SIMOLO et al., 2010). These weaknesses are particularly problematic in data-scarce tropical watersheds, where reconstructed rainfall often serves as the primary forcing for water balance, erosion, and ecohydrological models (MELLO et al., 2019; MIRALLES et al., 2010; PAREDES-TREJO; BARBOSA; Lakshmi Kumar, 2017; PUGAS, 2023; RODRIGUES et al., 2021a).

This study addresses those weaknesses by developing a multi-stage framework that reconstructs not just data values but the physical behaviour of the precipitation process. The methodology combines three components that are rarely deployed together in tropical hydrology: (1) unsupervised quality control using Isolation Forest (iForest) and Local Outlier Factor (LOF) to remove instrumental artefacts without truncating legitimate convective events (ALSALEHY; BAILEY, 2025; BĂRBULESCU; DUMITRIU; DRAGOMIR, 2021; BREUNIG et al., 2000; KIM; KIM; KOK, 2025; LIU; TING; ZHOU, 2008; SHAHRESTANI; SANISLAV, 2025); (2) multi-source predictor fusion, where ERA5, ERA5-Land, CHIRPS, and PERSIANN-CDR are bias-corrected via Quantile Mapping to anchor global atmospheric signals to local climatology (AL-RAWAS et al., 2025; CANNON; SOBIE; MURDOCK, 2015; FUNK et al., 2015; HERSBACH et al., 2020; PADULANO et al., 2025; SALAZAR et al., 2025; VILA et al., 2009); and (3) multi-objective model selection via Pareto–Borda ranking, which jointly optimises statistical precision (RMSE) and hydrological consistency (KGE), ensuring that the reconstructed series preserves the variance structure and event frequency essential for process-based modelling (BOZORG-HADDAD et al., 2016; SANTOS et al., 2024; EHRGOTT, 2005; GUPTA et al., 2009; KLING; FUCHS; PAULIN, 2012; KNOBEN; FREER; WOODS, 2019).

The framework is evaluated over more than 150 configurations spanning classical statistical models, kernel methods, tree-based ensembles, and recurrent neural networks at the Matinha forest reserve, where continuous rainfall forcing is a prerequisite for forest water balance analysis (Junqueira Junior et al., 2017; MELLO et al., 2019; RODRIGUES et al., 2021a).

The training window (January 2018–September 2022) is the only fully continuous segment of the record. Formal distributional tests (Kolmogorov–Smirnov, Anderson–Darling, Mann–Whitney U) show that it is statistically indistinguishable from the full 2013–2024 climatology across the 10th–90th percentile range, confirming that this period captures the regional rainfall regime that drives soil moisture, evapotranspiration, and streamflow in the surrounding headwaters (GUAUQUE-MELLADO et al., 2022; GUAUQUE-MELLADO et al., 2026; Junqueira Junior et al., 2017; MELLO et al., 2019; SILVA; MELLO, 2021). Long-term, high-quality records from a single well-instrumented fragment thus provide one of the few opportunities to observe a relatively undisturbed Atlantic Forest water cycle and to quantify how local canopy and soil conditions mediate regional hydroclimatic forcing (BORMA et al., 2022; RIBEIRO et al., 2009; CARNEIRO et al., 2023).

Matinha is therefore treated as an ecohydrological sentinel site whose local rainfall statistics and water balance responses are representative of the southern Minas Gerais highlands, and whose data can serve as a blueprint for other small forest remnants in data-sparse tropical regions (CAU et al., 2025;

OLIVEIRA et al., 2021; Junqueira Junior et al., 2017; PAREDES-TREJO; BARBOSA; Lakshmi Kumar, 2017; PUGAS, 2023).

4.1.1 Research question and contributions

Beyond providing a validated reconstruction pipeline for Matinha, the study addresses a question with international relevance: how do (i) the quality of training data, (ii) the choice of atmospheric predictors, and (iii) the evaluation framework each contribute to the hydrological fidelity of ML-reconstructed precipitation records in tropical, data-scarce regions? Three contributions follow.

First, data quality as a precursor to precision: unsupervised outlier detection tailored to zero-inflated, heavy-tailed rainfall (Isolation Forest) is shown to be a prerequisite for predictive performance, and the choice of filter interacts non-trivially with the downstream algorithm.

Second, physically consistent predictors over statistical convenience: bias-corrected ERA5 reanalysis, which encodes three-dimensional atmospheric state, systematically outperforms satellite-only products as a source of exogenous information.

Third, hydrologically meaningful model selection: a Pareto–Borda ranking jointly optimises statistical precision (RMSE) and hydrological consistency (KGE), trading a small loss in point-wise error for gains in variance preservation, event frequency, and structural realism (GUPTA et al., 2009; MIZUKAMI et al., 2019; CINKUS et al., 2023; DASGUPTA et al., 2023; YESTE et al., 2023).

Answering this question at a well-instrumented Atlantic Forest sentinel site provides a reproducible blueprint for rainfall reconstruction in other headwater catchments and forest remnants across the Atlantic Forest and the tropical Global South, where monitoring networks are sparse and hydrological process characterisation is otherwise underdetermined (ANDRADE et al., 2024; BORMA et al., 2022; PAREDES-TREJO; BARBOSA; Lakshmi Kumar, 2017; PINTHONG et al., 2025; PUGAS, 2023; SURI; AZAD, 2024; SU; ZHANG et al., 2026).

4.2 Materials and Methods

The proposed methodology is grounded in principles of reproducibility and robust hydrological validation (ARSENAULT; BRISSETTE; MARTEL, 2018; SHEN; TOLSON; MAI, 2022; CINKUS et al., 2023). Unlike simple imputation approaches based on linear interpolation or historical averaging (ADDI et al., 2022; BLEIDORN et al., 2024), this study implements an adaptive workflow designed to learn the covariance structure between in-situ observations and a diverse set of exogenous predictors (KUMAR; DWARAKISH, 2025; PIADEH et al., 2023; COUTINHO et al., 2018). The process follows six hierarchical phases: (i) data acquisition and quality control (QC) (ABOUZIED et al., 2025; KIM; KIM;

KOK, 2025); (ii) advanced feature engineering (GHODRATI; DARIANE, 2024); (iii) bias correction of satellite products via Quantile Mapping (CANNON; SOBIE; MURDOCK, 2015; PADULANO et al., 2025); (iv) multi-scale model training (daily and hourly); (v) multi-objective model selection via Pareto optimization (FOWLER et al., 2018); and (vi) reconstruction of the complete series. Figure 4.1 illustrates the complete system architecture.

4.2.1 Data acquisition

The study site is the *Reserva Florestal da Mata Atlântica* ("Matinha", Figure 4.2), a montane semideciduous forest fragment within the Federal University of Lavras (UFLA) campus in Minas Gerais, Brazil ($21^\circ 13' 40''$ S, $44^\circ 57' 50''$ W) (Junqueira Junior et al., 2017). The site lies at 920–940 m a.s.l. and represents a critical remnant of the threatened Atlantic Forest biome (MAGALHÃES et al., 2019; RODRIGUES et al., 2020).

The regional climate is Cwa (Köppen), characterized by a wet season driven by the South Atlantic Convergence Zone (SACZ) from October to March and a dry season from April to September (OLIVEIRA et al., 2021). Mean annual precipitation is approximately 1380 mm, with more than 80% falling during the wet season (Junqueira Junior et al., 2017; LEMOS et al., 2023). This seasonality imposes a dual challenge for imputation algorithms: modeling consecutive zero values during droughts and reproducing the variance of convective storms during the wet season (BACK; GALATTO; SOUZA, 2024; FERREIRA; MENDES, 2022).

In-situ observations and exogenous predictors

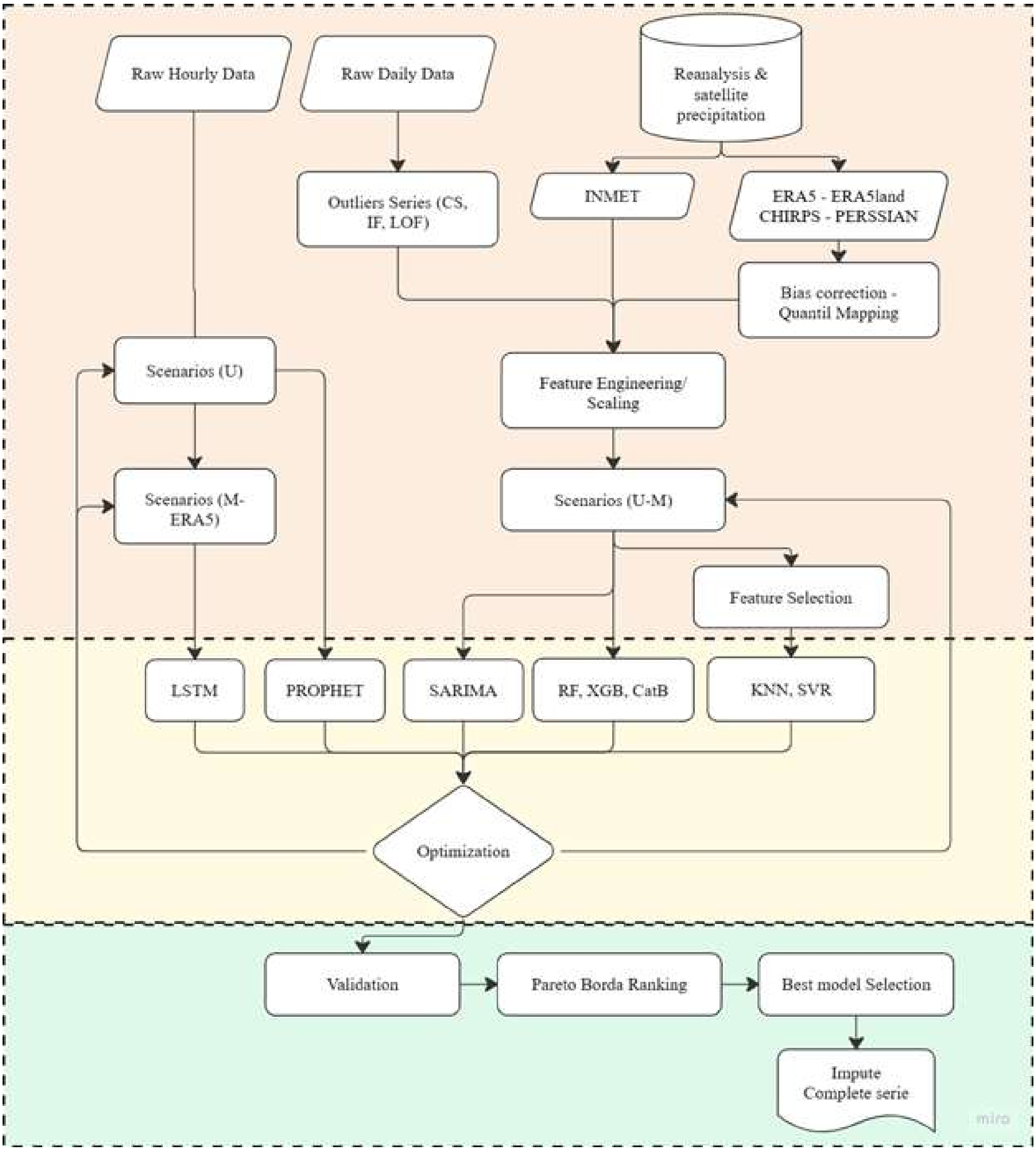
The target variables are the hourly and daily precipitation measured at the Meteorological Tower of Matinha (Junqueira Junior et al., 2017; OLIVEIRA et al., 2021). The record (2013–2024) contains gaps from sensor failures. The interval January 2018 to September 2022, which is fully continuous and failure-free, was designated as the **Training Window** for model calibration. Data were acquired at hourly and daily resolutions. Additionally, a conventional INMET station (code 3023) serves as an auxiliary reference for bias correction (VIANA et al., 2023).

Multiple global precipitation products were integrated as exogenous covariates (Table 4.1).

4.2.2 Raw series characterization

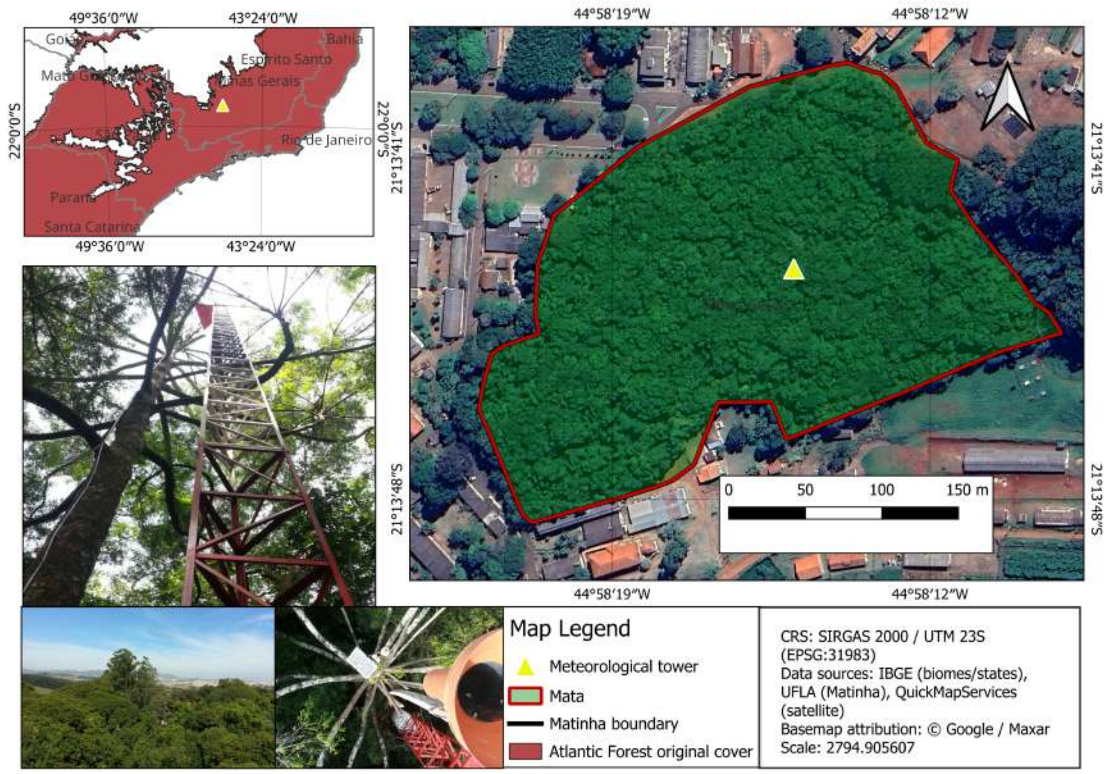
Before defining the preprocessing pipeline, the raw record was systematically characterized to expose the nature of its deficiencies and justify the methodological choices that follow.

Figure 4.1 – Conceptual workflow of the proposed imputation framework, from raw data ingestion and unsupervised anomaly detection to multi-objective Pareto–Borda model selection.



Fonte: The author (2026).

Figure 4.2 – Study area: Brazil (Atlantic Forest biome) and Matinha forest reserve (UFLA). Coordinate Reference System: SIRGAS 2000/UTM 23S.



Fonte: The author (2026).

Missing data structure

Figure 4.3 reveals two structurally distinct failure modes in the 2013–2024 record. Extended block failures (2013–2016 and 2022–2024) correspond to prolonged sensor outages spanning multiple wet seasons, for which simple interpolation is physically inadmissible. Sporadic short-duration failures, scattered throughout the record, reflect transient malfunctions during datalogger resets or high-intensity convective events. Jointly, these failures account for 1,097 missing days out of 4,382 total (25.6%), as summarized in Table 4.1, establishing the need for a supervised learning approach anchored in reliable atmospheric covariates. The sole continuous, failure-free segment—January 2018 to September 2022—was therefore designated as the Training Window.

Statistical diagnostics and spectral structure

Descriptive analysis confirms a strongly zero-inflated daily distribution, with a median near zero and daily maxima ranging from 30.9 mm (2013) to 93.7 mm (2016) (CHIVERS et al., 2018;

Box 4.1 – Summary of meteorological and remote-sensing datasets used in this study.

Variable	Source	Spatial Res.	Temporal Res.	Role / Reference
Precipitation (Target)	Matinha Tower (MT)	In-situ	Hourly/Daily	Dependent variable (Junqueira Junior et al., 2017; OLIVEIRA et al., 2021)
Precipitation (Ref.)	INMET (3023)	In-situ	Daily	Bias reference (SILVA; MELLO, 2021; VIANA et al., 2023)
Total precipitation	ERA5	0.25° (∼31 km)	Hourly/Daily	Exogenous predictor (HERSBACH et al., 2020)
Total precipitation	ERA5-Land	0.1° (∼9 km)	Daily	Exogenous predictor (MUÑOZ-SABATER et al., 2021)
Precipitation	CHIRPS v2.0	0.05° (∼5 km)	Daily	Exogenous predictor (FUNK et al., 2015)
Precipitation	PERSIANN-CDR	0.25° (∼25 km)	Daily	Exogenous predictor (ASHOURI et al., 2015)

Fonte: The author (2026).

FARAMARZZADEH et al., 2023). The Mann–Kendall test detected no significant monotonic trend ($\tau = -0.019$, $p = 0.144$) (HODSON, 2022), and the Augmented Dickey–Fuller test confirmed strict stationarity ($p < 10^{-12}$) (ANSARI; GROSSI, 2022). Together, these results support the transferability of models trained on the 2018–2022 window to earlier and later periods. Spectral analysis identified dominant periodicities at 340–425 days (annual hydrological cycle) and secondary peaks near 20 days (mesoscale convective organization), directly motivating the inclusion of harmonic Day-of-Year features (DOY_{sin}, DOY_{cos}) and multi-scale rolling accumulations (3, 7, and 14 days) in the feature engineering stage (Section 5.2.5).

Representativeness of the Training Window

A formal distributional comparison between the 2018–2022 window and the full 2013–2024 record showed near-identical distributions across the 10th–90th percentile range. Non-parametric tests (Kolmogorov–Smirnov, Anderson–Darling, and Mann–Whitney U) all yielded $p > 0.90$, confirming that the selected window is statistically representative of the local rainfall climatology. This justifies using it as the exclusive training dataset, preventing information leakage and ensuring that all learned relationships derive from uninterrupted, unimputed observations (SHEN; TOLSON; MAI, 2022; ARAYA et al., 2023).

4.2.3 Advanced pre-processing framework

Unsupervised anomaly detection

Precipitation time series often contain spurious values that reflect instrumental errors rather than real meteorological phenomena. To filter these without removing legitimate extreme events (BĂRBULESCU; DUMITRIU; DRAGOMIR, 2021), three versions of the target variable were generated:

1. **Raw series**: retains all original data.

2. **Isolation Forest (iForest)**: an unsupervised algorithm that isolates anomalies via recursive partitioning, effective in high-dimensional spaces and appropriate for non-normal distributions such as precipitation (LIU; TING; ZHOU, 2008; MASTOUR; FADUL, 2023; KIM; KIM; KOK, 2025). The scikit-learn implementation with automatic contamination estimation was used; observations flagged as anomalies were masked as NaN for subsequent imputation.

3. **Local Outlier Factor (LOF)**: a density-based method that identifies values anomalous within their local temporal context, even when not extreme in absolute magnitude (BREUNIG et al., 2000; SHAHRESTANI; SANISLAV, 2025). This is useful for detecting sensor drift.

Generating these three filtered series allows evaluating not only the models' ability to fill gaps, but also their sensitivity to training data quality.

Characterization of removed records

Both methods were applied exclusively to the daily series. Each algorithm flagged 3.9% of valid observations in the complete record (2013–2024), corresponding to approximately 128 daily values per method out of 3,283 valid days.

Figure 4.3 reveals a clear physical divergence between the two detection strategies. Isolation Forest flags predominantly high-magnitude peaks (>20–30 mm day^{-1}), distributed throughout the record and coinciding with the most intense convective events of each wet season. LOF, by contrast, identifies a larger proportion of low-to-moderate anomalous values that are locally inconsistent with their immediate temporal neighbors—a pattern consistent with sensor drift, partial blockage of the rain gauge funnel, or datalogger transmission errors rather than physically extreme rainfall.

This contrast has direct consequences for the reconstructed series. Isolation Forest removes records that are extreme in absolute magnitude but uses a recursive partitioning criterion—rather than local density—that avoids removing values simply because they are rare, which is appropriate for a zero-inflated, heavy-tailed distribution (LIU; TING; ZHOU, 2008; KIM; KIM; KOK, 2025). LOF tends to penalize high-intensity events that are anomalous relative to their immediate neighborhood, even when physically plausible, thereby suppressing the upper tail of the training distribution.

The study site instrument is an 8-inch tipping-bucket rain gauge (0.20 mm per tip, Campbell Scientific CR10X) (Junqueira Junior et al., 2017; OLIVEIRA et al., 2021). A documented correction factor of $P_{\text{manual}} = 1.047 \times P_{\text{bucket}}$ ($r^2 = 0.95$) and aerodynamic undercatch affecting 4.9% of events at wind speeds above 5.0 m s^{-1} (Junqueira Junior et al., 2017) imply that values flagged by Isolation Forest at the high end of the distribution may reflect genuine extreme events recorded with moderate underestimation rather than sensor malfunctions. Consequently, the filtered series is most reliable for water balance studies and ecohydrological modeling, and should be used with caution in analyses focused exclusively on high-intensity extremes (>50 mm day^{-1}).

Bias correction via Quantile Mapping

To correct systematic biases in satellite products (e.g., the drizzle effect), Quantile Mapping (QM) was applied (CANNON; SOBIE; MURDOCK, 2015; PADULANO et al., 2025). The transformation for a modeled value x_t is:

$$\hat{x}_t = F_O^{-1}(F_P(x_t)) \tag{4.1}$$

where F_P is the empirical CDF of the predictor and F_O^{-1} is the inverse CDF of the INMET reference station. This generates a dual feature set (raw and QM-corrected) for input into the ML models (LI et al., 2023).

Feature engineering

A feature matrix **X** was constructed to capture temporal dynamics (AMINI; DOLATSHAHI; KERACHIAN, 2024):

- **Lagged variables**: from $t-1$ to $t-14$ days, capturing system memory (SILVA et al., 2025; AWE; VANCE, 2025).
- **Rolling accumulations**: 3-, 7-, and 14-day sums, acting as proxies for antecedent soil moisture (LIU; LI; DUAN, 2023; AMJAD; SHAHZAD; UL, 2025).
- **Cyclical seasonality**: harmonic transformations of the Day of Year (DOY) (DIEZ-SIERRA; JESUS, 2020; COSTA; CAMPOS; SOARES, 2023):

$$\text{DOY}_{\text{sin}} = \sin\left(\frac{2\pi \cdot \text{DOY}}{365}\right), \quad \text{DOY}_{\text{cos}} = \cos\left(\frac{2\pi \cdot \text{DOY}}{365}\right) \tag{4.2}$$

- **Hydrological indicators**: binary flags for the wet season and continuous indices for the hydrological year phase (ARAYA et al., 2023).
- **Exogenous features**: contemporary (t) and lagged ($t-1, t-2$) values of all satellite and reanalysis products, in both raw and QM-corrected versions.

To prevent dimensionality issues and multicollinearity in distance-sensitive models (KNN and SVR), a correlation filter ($\rho > 0.90$ threshold) and Variance Inflation Factor (VIF) analysis were applied.

4.2.4 Computational modeling strategy

Two experimental scenarios were defined: Univariate (U), using only endogenous tower memory, and Multivariate (M), integrating exogenous predictors (WANG; TIAN; CARROLL, 2023).

Daily imputation models

Five machine learning algorithms and one statistical benchmark were selected to cover distinct learning paradigms—decision tree ensembles, instance-based methods, and kernel methods (Table 4.2).

Hyperparameter optimization. Hyperparameters were tuned using a hybrid approach: Grid Search for simpler model spaces and Optuna (Bayesian Optimization) for complex ensembles, maximizing the KGE metric on the validation set (MONEGO; ANOCHI; de Campos Velho, 2022).

Hourly imputation models

For high-frequency reconstruction, two models were implemented:

1. **Prophet** (TAYLOR; LETHAM, 2018): a generalized additive model with custom seasonalities for wet/dry periods:
$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{4.3}$$
where $g(t)$ is the trend component, $s(t)$ seasonality, and $h(t)$ holiday effects (not used here). Distinct daily cycles were defined for wet and dry seasons, and hourly external regressors (radiation, humidity, wind) were included. Hyperparameter tuning used Grid Search over `changepoint_prior_scale` and `seasonality_prior_scale`.

2. **Bidirectional LSTM** (HOCHREITER; SCHMIDHUBER, 1997; ISHIDA et al., 2024): a stacked architecture where the first layer (64 units) processes the sequence in both temporal directions,

Box 4.2 – Description and configuration of the imputation algorithms evaluated for daily precipitation reconstruction.

Model	Type	Description and rationale	Key reference
SARIMA	Statistical linear	Captures explicit seasonality $(p,d,q)(P,D,Q)_s$; used as statistical benchmark.	Bleidorn et al. (2024), Dabral e Murry (2017)
KNN	Instance-based	Imputes via weighted average of k nearest neighbors in feature space.	Anaraki et al. (2023), Cover e Hart (1967)
SVR	Kernel method	Maps inputs to high-dimensional space using RBF kernel for non-linear regression.	Sriwahyuni et al. (2025), Smola e Schölkopf (2004)
Random Forest	Ensemble (Bagging)	Builds independent trees; robust to noise and multicollinearity.	Breiman (2001), Başağaoğlu et al. (2022)
XGBoost	Gradient Boosting	Sequential tree building with L1/L2 regularization to prevent overfitting.	Chen e Guestrin (2016), Okolie et al. (2024)
CatBoost	Gradient Boosting	Uses Ordered Boosting to prevent target leakage; handles categorical features.	Prokhorenkova et al. (2018)

Fonte: The author (2026).

followed by a Dropout layer (rate 0.2) and a second layer (32 units) before the dense output. A lookback window of **360 hours** (15 days) was used to capture antecedent moisture conditions identified in the spectral analysis. The Adam optimizer with Mean Squared Error loss and Early Stopping (patience 10) was used for training.

4.2.5 Experimental design and robust validation

Validation schemes

Three data partitioning strategies were implemented to evaluate model robustness: (1) K-Fold cross-validation (5 folds): evaluates global generalization capability; (2) Time-Series Split (TSS): walk-forward validation where the training set grows progressively and evaluation is always performed on the subsequent temporal block, simulating real-time operational conditions; (3) Hydrological Year Split: validation blocks are defined by hydrological years (October–September) to avoid intra-seasonal autocorrelation leakage (ARAYA et al., 2023).

All partitions respect strict temporal ordering to prevent data leakage (SHEN; TOLSON; MAI, 2022).

Multi-objective model selection: the Pareto–Borda framework

Single-metric evaluations relying solely on RMSE minimization frequently lead to variance underestimation and flattening of hydrological peaks (GUPTA et al., 2009; MIZUKAMI et al., 2019). To overcome this, a multi-objective selection framework was adopted.

Models were evaluated simultaneously using RMSE (precision) and the Kling–Gupta Efficiency (KGE, hydrological consistency):

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{4.4}$$

where r is the Pearson correlation, $\alpha = \sigma_{\mathrm{sim}}/\sigma_{\mathrm{obs}}$ is the variability ratio, and $\beta = \mu_{\mathrm{sim}}/\mu_{\mathrm{obs}}$ is the bias ratio.

The hybrid Pareto–Borda system operates in two phases:

Phase 1: Pareto front identification. Non-dominated models were identified in the bi-objective space. A model A dominates model B if:

$$\mathrm{RMSE}_A < \mathrm{RMSE}_B \quad \text{and} \quad \mathrm{KGE}_A > \mathrm{KGE}_B.$$

All non-dominated models constitute the Pareto-optimal front.

Phase 2: Borda count ranking. To select a single best-compromise solution from the Pareto-optimal set, the Borda voting method was applied (BOZORG-HADDAD et al., 2016). Models were ranked independently by each metric and a combined score was computed as:

$$\mathrm{BordaScore}(m) = \mathrm{Rank}_{\mathrm{RMSE}}(m) + \mathrm{Rank}_{\mathrm{KGE}}(m) \tag{4.5}$$

where $\mathrm{Rank}_j(m)$ denotes the rank of model m for metric j (Rank 1 = best). The model with the lowest cumulative BordaScore was selected for final imputation. This approach provides an explicit, reproducible framework that balances error magnitude with the physical realism required for ecohydrological applications in data-scarce regions (SANTOS et al., 2024).

4.3 Results

Results are organized along two temporal scales. Section 4.3.1 addresses the daily-scale reconstruction pipeline: data quality assessment, multi-objective model selection, structural fidelity, and model interpretability. Section 4.3.2 presents the hourly-scale analysis using Prophet and LSTM. Section 4.3.3 integrates both scales in the final reconstructed series (2013–2024).

4.3.1 Daily-scale results

Data quality assessment

The rainfall record for 2013–2024 shows substantial heterogeneity in data completeness, with extended discontinuities between 2013–2016 and additional gaps during 2022–2024 (Fig. 4.3). The absence of any seasonal pattern in the distribution of missing values suggests that these interruptions are attributable to instrumental failures or maintenance periods rather than to climate processes. The interval 2018–2022 is the only fully continuous, failure-free segment (0% missing data), making it the most suitable window for training and validation.

Descriptive analysis confirmed the highly intermittent character of rainfall, with daily medians close to zero and isolated high-intensity convective events. Daily maxima ranged from 30.9 mm (2013) to 93.7 mm (2016), reinforcing the need for imputation methods that can preserve both the frequency of dry days and the variability of the upper tail.

Isolation Forest (IF) and Local Outlier Factor (LOF) both showed the same nominal removal rate (3.9% of the 3,283 valid days), but filtering affected less than 1.4% of valid data, confirming that preprocessing enhanced data integrity without removing hydrologically meaningful signals (Table 4.1).

Table 4.1 – Descriptive statistics of the rainfall time series and percentage of data removed by Isolation Forest and LOF.

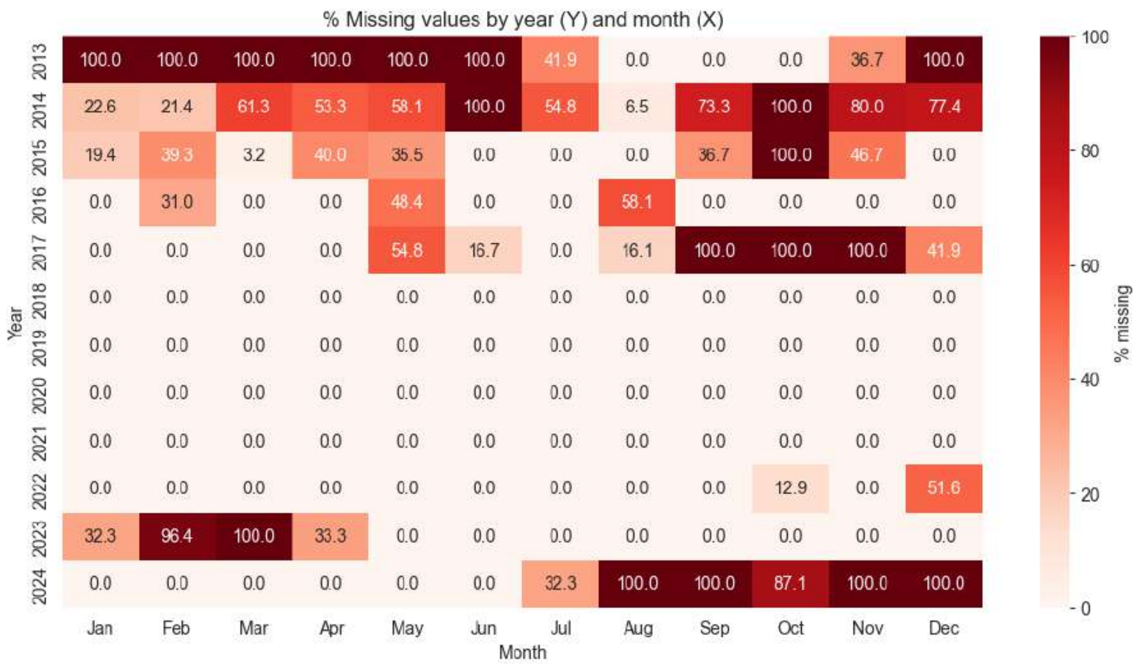
Statistic	**Full Series**
Start date	2013-01-01
End date	2024-10-01
Total days	4382
Missing days	1097
Valid days	3283
% removed (IF)	1.25%
% removed (LOF)	1.32%

Fonte: The author (2026).

Exploratory data analysis

Daily rainfall is strongly zero-inflated: the majority of values fall below 5 mm day^{-1}, events above the 90th percentile account for most temporal variability, annual seasonality governs series structure, and missing values occur in blocks consistent with the history of instrumental failures.

Figure 4.3 – Temporal structure of missing rainfall data across the full observational period (2013–2024).



Fonte: The author (2026).

Justification for the 2018–2022 training window

Although the dataset spans more than a decade, only 2018–2022 provides full continuity without instrumental failures. A comparative analysis between this window and the full 2013–2024 record showed near-identical distributions for the 10th–90th percentiles. Non-parametric tests (Kolmogorov–Smirnov, Anderson–Darling, Mann–Whitney U) yielded $p > 0.90$ in all cases, confirming that the selected window accurately reflects the extended local climate. This choice prevents information leakage, ensures learning from uninterrupted observations, preserves the distributional shape of rainfall, and provides a robust climatology for bias correction.

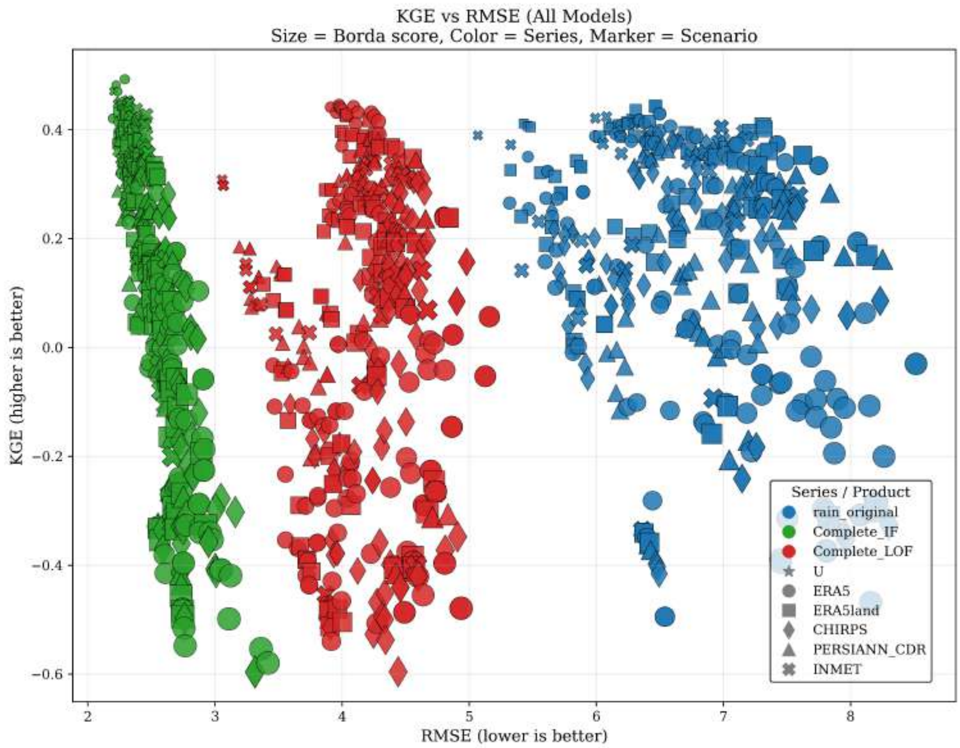
Predictive performance and model selection

Daily rainfall reconstruction involved a rigorous multi-objective competition of more than 150 model configurations, focused on identifying the optimal trade-off between statistical precision (RMSE) and hydrological consistency (KGE).

The performance distribution (Figure 4.4) reveals a clear hierarchy governed by data quality and predictor richness. Models trained on the Isolation Forest-filtered series (green markers) consistently occupy the region of minimal error and maximal efficiency, dominating the Pareto frontier. Models based on the raw series exhibit significant dispersion and higher residual errors, confirming that uncorrected instrumental noise interferes with the convergence of gradient-boosting algorithms.

As quantified in Table 4.2, the optimal model—an XGBoost regressor driven by ERA5 reanalysis and trained on the IF-filtered series—reduced RMSE by **74.6%** relative to SARIMA (from 8.76 to 2.22 mm d^{-1}) and nearly doubled the KGE (from 0.26 to 0.48).

Figure 4.4 – Multi-objective model selection. Distribution of all trained models in the RMSE–KGE performance space. Marker size is proportional to the Borda score. Green markers (`Complete_IF`) cluster in the optimal top-left quadrant, demonstrating superior consistency over LOF-filtered (red) and unfiltered (blue) series.



Fonte: The author (2026).

Table 4.2 – Performance comparison between the statistical baseline (SARIMA) and the top-ranked machine learning models for each target series.

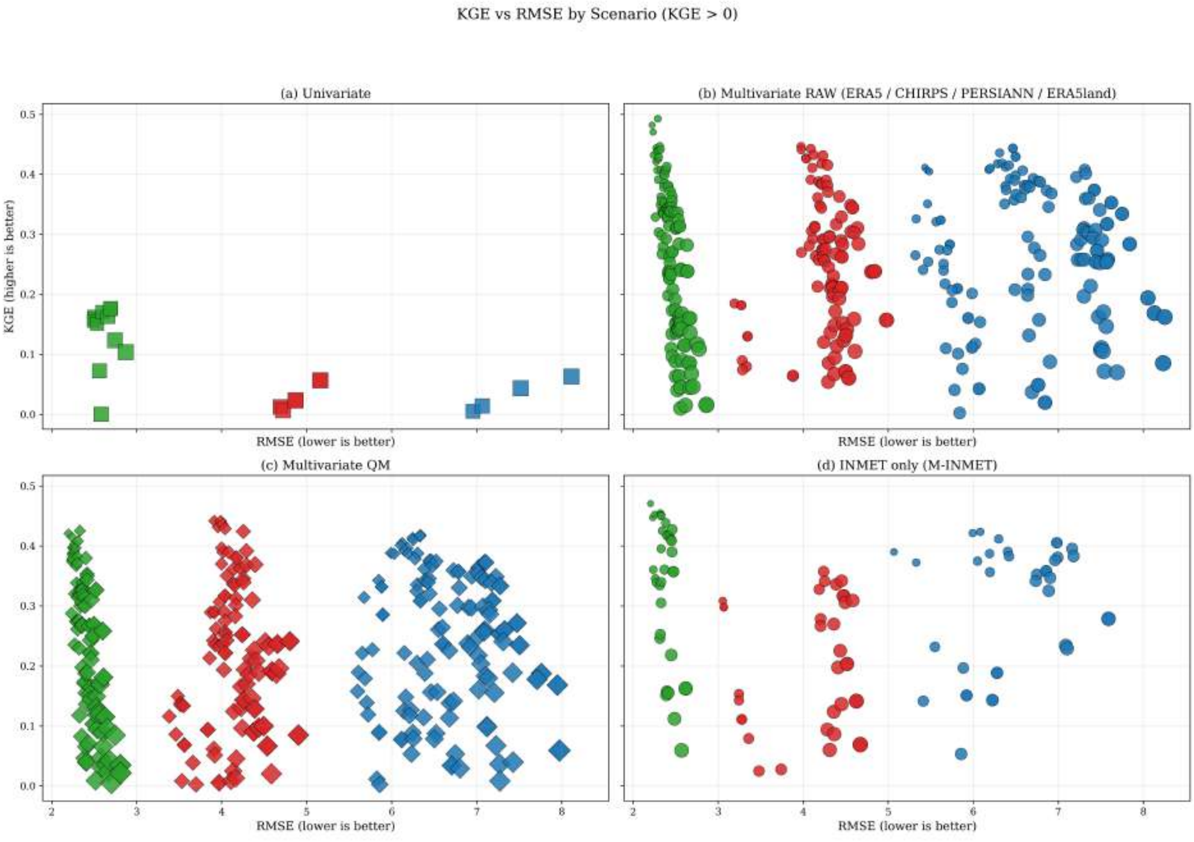
Target series	Scenario	Model	RMSE (mm d^{-1})	MAE (mm d^{-1})	NSE	KGE
Benchmark	Univariate	SARIMA	8.76	4.70	0.04	0.26
Complete_IF	**M-ERA5**	**XGB (Grid)**	**2.22**	**0.89**	**0.38**	**0.48**
Complete_LOF	M-INMET	RF (Default)	3.06	1.21	0.30	0.31
rain_original	M-ERA5land	RF (Grid)	5.43	2.45	0.29	0.41

Fonte: The author (2026).

Detailed performance drivers: scenarios, sources, and algorithms

Impact of data fusion and bias correction. Performance stratified by scenario (Figure 4.5) confirms that multi-source data fusion is critical for reconstruction in tropical regions. Univariate models are confined to a region of poor performance (KGE < 0.2). Raw multivariate predictors shift performance vertically, improving KGE considerably while maintaining moderate RMSE values. The most pronounced improvement occurs with Quantile Mapping, which produces the tightest clustering in the high-efficiency zone (KGE > 0.4, RMSE < 3.0 mm d^{-1}), particularly for the IF-filtered series.

Figure 4.5 – Performance disaggregation by modeling scenario. Scatter plots of KGE vs. RMSE for (a) univariate models, (b) multivariate models with raw inputs, (c) multivariate models with bias-corrected (QM) inputs, and (d) models using INMET station data.

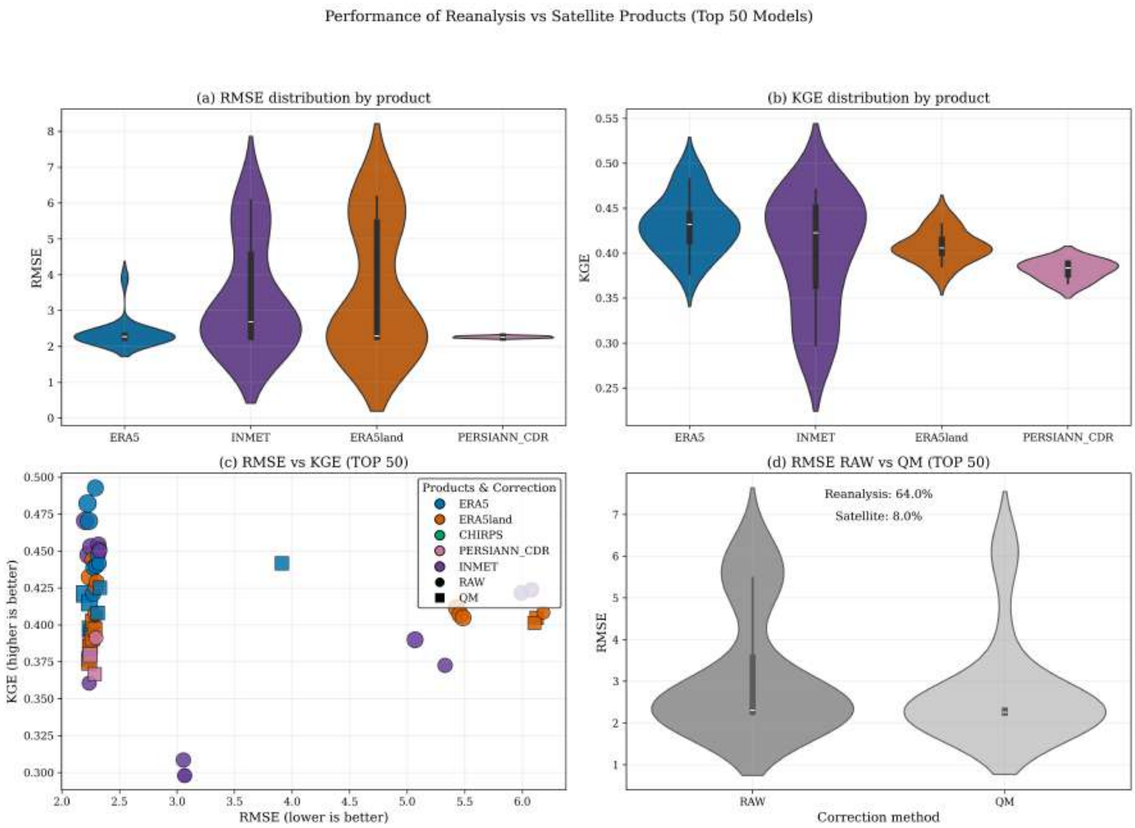


Fonte: The author (2026).

Hierarchy of exogenous predictors: reanalysis vs. satellite. Analysis of the top-50 best-performing models reveals a clear dominance of reanalysis over satellite products (Figure 4.6). ERA5 and ERA5-Land consistently achieve lower median RMSE and higher KGE than CHIRPS or PERSIANN-CDR. Quantitatively, 64% of the Pareto-optimal top-50 configurations rely on reanalysis data, while only 8% rely on satellite-only products. ERA5 assimilates three-dimensional atmospheric state variables, providing a robust physical proxy for rainfall generation even when direct surface signatures are weak.

Bias correction (QM) significantly reduces the error tail, compressing the distribution toward more accurate results.

Figure 4.6 – Assessment of exogenous data sources within the top-50 models. (a, b) Violin plots of RMSE and KGE by product; (c) Pareto dominance of ERA5-based models; (d) comparative impact of raw vs. QM-corrected data.



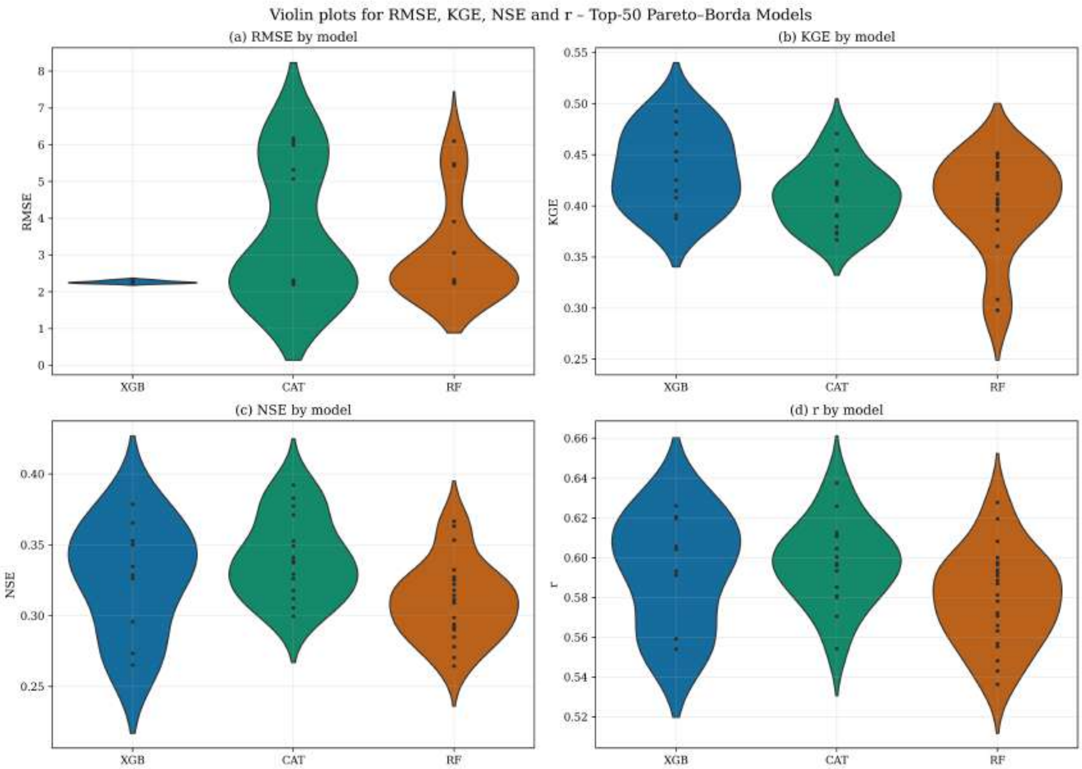
Fonte: The author (2026).

Algorithm stability and error distribution. XGBoost emerges as the most robust algorithm (Figures 4.7–4.8). Its RMSE distribution is tightly concentrated (≈ 2.2 mm d^{-1}), whereas Random Forest shows a longer tail, indicating overfitting in certain complex folds. CatBoost performs competitively but with slightly higher variance. XGBoost configurations cluster consistently in the high-KGE/low-RMSE region, justifying their selection as the primary imputation engine.

Structural fidelity and variational analysis

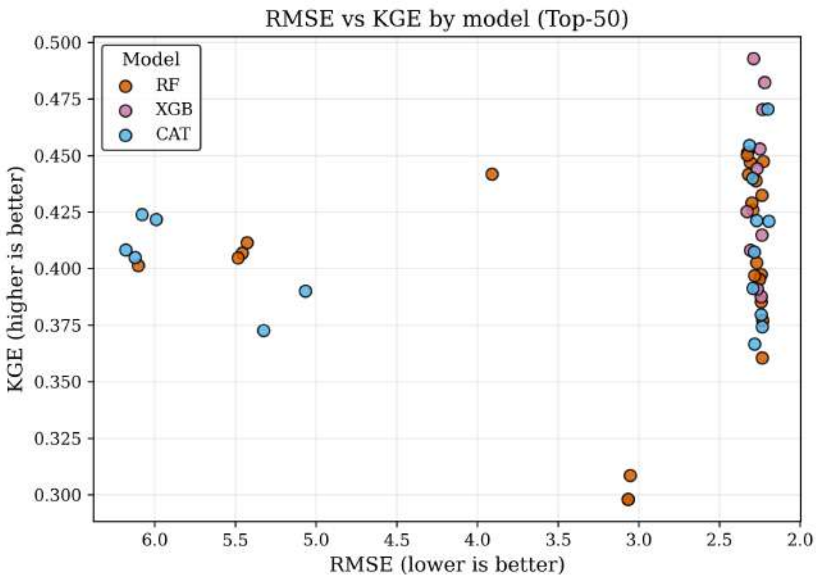
Taylor Diagrams were used to assess whether models capture the dynamic structure of precipitation beyond simple error minimization (Figure 4.9). These diagrams simultaneously display the Pearson correlation (r), centered RMSE (cRMSE), and normalized standard deviation (σ).

Figure 4.7 – Algorithm stability analysis. Violin plots of RMSE, KGE, NSE, and r for XG-Boost (XGB), CatBoost (CAT), and Random Forest (RF).



Fonte: The author (2026).

Figure 4.8 – Model performance clustering (top-50). Scatter plot of RMSE vs. KGE colored by algorithm.

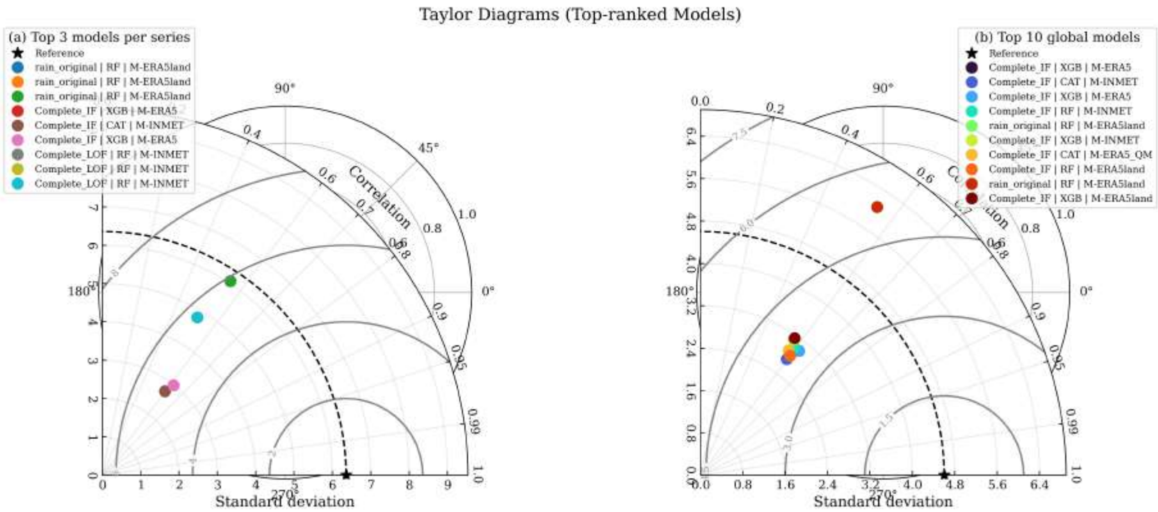


Fonte: The author (2026).

Models trained on `Complete_IF` cluster nearest to the reference point ($r \approx 0.63$, $\sigma_{\text{pred}}/\sigma_{\text{obs}} \approx 1.0$), indicating that Isolation Forest removed non-physical artifacts without suppressing natural variance.

In contrast, `Complete_LOF` and `rain_original` models cluster closer to the origin (radial distance < 1.0), reflecting variance suppression from LOF's tendency to flag high-intensity peaks as outliers, and noise inflation from the unfiltered series. The global top-10 is almost exclusively composed of `Complete_IF` configurations with M-ERA5 or M-INMET scenarios, confirming that no amount of algorithmic tuning can compensate for insufficient quality control in the target variable.

Figure 4.9 – Structural evaluation via Taylor Diagrams. (a) Top-3 models per target series. (b) Global top-10 ranked by Pareto–Borda.



Fonte: The author (2026).

Explainability and physical consistency

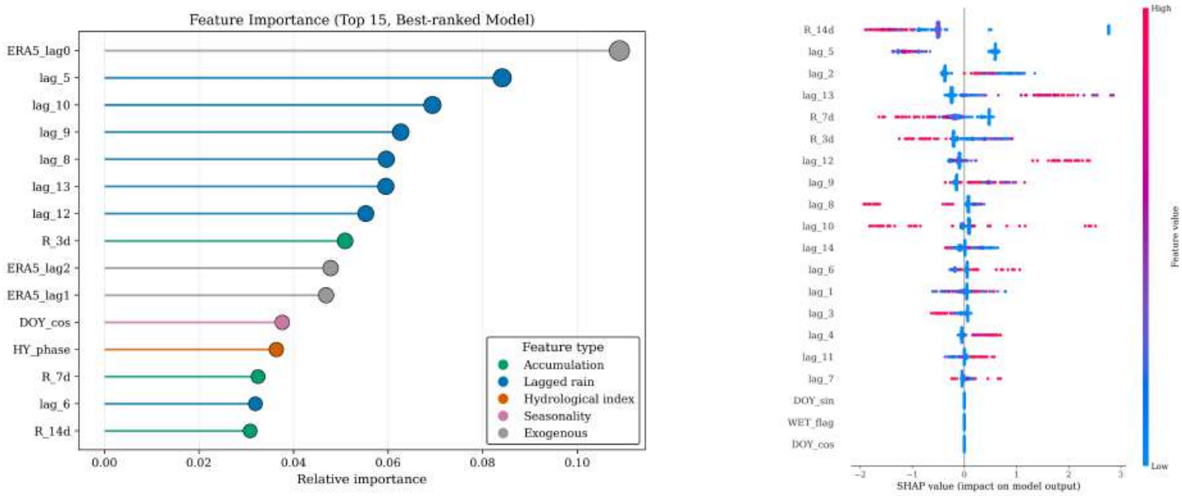
Two interpretability methods were compared: Global Feature Importance (gain-based) and SHAP Values (marginal contribution), to verify that the model learns physically coherent relationships.

The Feature Importance ranking (Figure 4.10a) identifies `ERA5_lag0` as the dominant predictor—the synoptic-scale trigger for rainfall. The SHAP Summary Plot (Figure 4.10b) reveals that prediction magnitude is heavily modulated by local autoregressive terms (`R_14d`, `lag_5`): high antecedent moisture (pink dots) pushes SHAP values strongly positive, capturing the physical persistence of wet spells, while low lag values (blue dots) suppress false alarms during dry periods. This dual structure—ERA5 setting the synoptic context, local lags calibrating magnitude—confirms the physical robustness of the learned model.

4.3.2 Hourly-scale results

The hourly rainfall series presents a structurally different challenge from the daily scale: extreme zero-inflation, short-lived convective bursts, and weak temporal persistence. Two approaches were

Figure 4.10 – (a) Feature importance: `ERA5_lag0` is the primary structural splitter. (b) SHAP values: local memory features (`R_14d`, `lag_5`) modulate prediction magnitude, with high antecedent moisture (pink) increasing predicted rainfall.



Fonte: The author (2026).

implemented: Prophet and a Bidirectional LSTM. Their performance is summarized in Table 4.3. Prophet achieves low bias owing to its smooth additive structure, but its hydrological efficiency remains poor (KGE < 0), revealing limited capacity to reproduce short-term rainfall variability. The LSTM substantially outperforms Prophet across all metrics, supporting the hypothesis that non-linear recurrent architectures are better suited to intermittent convective dynamics.

Table 4.3 – Performance metrics for hourly rainfall reconstruction using Prophet and LSTM.

Model	MAE	RMSE	NSE	KGE	Correlation	Bias
Prophet (Tuned)	0.273	1.088	0.016	−0.262	0.136	−0.057
LSTM (Best)	0.10	0.70	0.35	0.35	0.87	−0.01

Fonte: The author (2026).

Prophet as a diagnostic tool: sub-daily and seasonal structure

Although Prophet shows limited capacity for quantitative imputation, its additive decomposition provides useful information on the deterministic components of the rainfall regime. Detected changepoints coincide with periods of increased variance rather than long-term climatic shifts, suggesting that Prophet adjusts its baseline trend reactively to clusters of convective events.

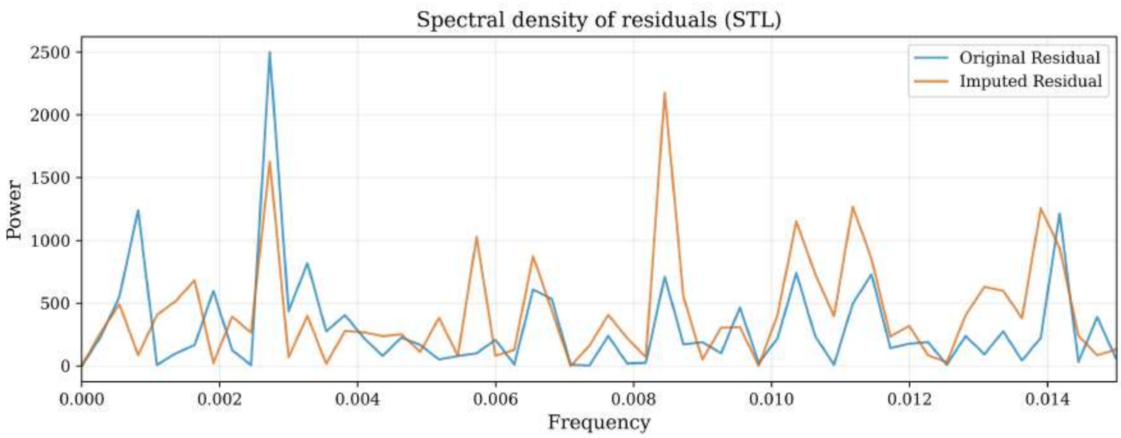
The diurnal component displays a physically coherent unimodal cycle: rainfall probability remains minimal during morning hours (08:00–10:00) and peaks between 17:00–20:00, consistent with thermodynamic convective forcing. Annual seasonality follows the expected structure, with a strong positive phase in austral summer (DJF) and a pronounced minimum in winter (JJA). A weak, physically unstructured weekly component confirms that sub-daily rainfall is not linked to anthropogenic cycles.

LSTM: non-linear learning of convective variability

The LSTM captures temporal dependency structure more effectively than Prophet, learning the short-term memory and non-linear relationships essential for convective precipitation. Its higher correlation ($r = 0.87$) indicates accurate timing of rainfall events, though extreme values remain somewhat underestimated—a known consequence of MSE-based optimization.

STL decomposition confirmed that the LSTM preserves long-term trend and seasonal cycles while introducing moderate deviations at the extremes. Spectral analysis of residuals (Figure 4.11) shows that the LSTM reconstruction maintains the dominant frequencies of the original signal, with slightly higher spectral power at high frequencies reflecting uncertainty in fine-scale convective perturbations rather than systematic bias.

Figure 4.11 – Spectral density of STL residuals for the original and LSTM-reconstructed hourly series.



Fonte: The author (2026).

Together, Prophet and LSTM offer complementary contributions: Prophet reveals underlying climatological rhythms; LSTM reconstructs the high-frequency stochastic dynamics required for hydrological applications.

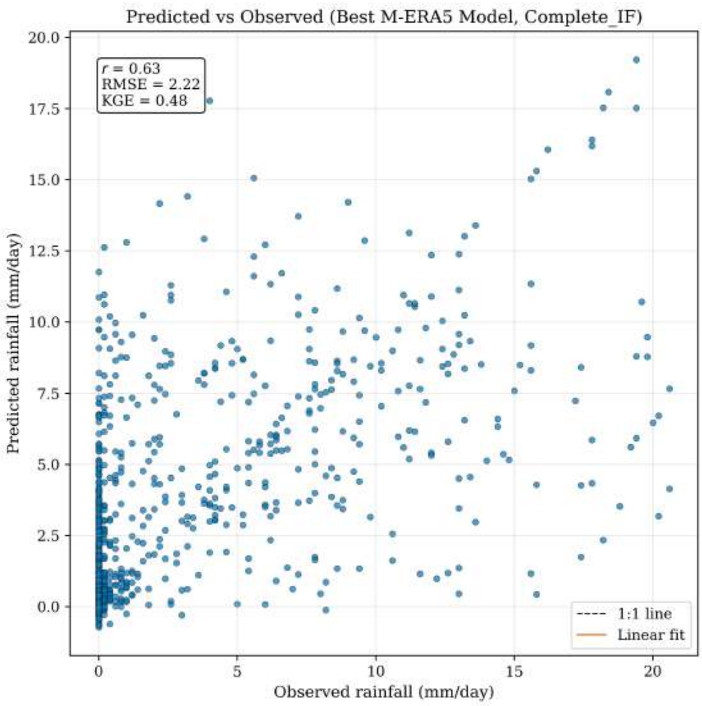
4.3.3 Final reconstruction: temporal structure and hydrological consistency (2013–2024)

Although trained exclusively on the 2018–2022 segment, the optimal model (Complete_IF + M-ERA5 + XGBoost) was applied to the full 2013–2024 period, because the underlying hydroclimatic processes are stationary in distribution. This approach differs from simple gap-filling: instead of interpolating, the model reconstructs missing values by emulating the physical–statistical behavior of the rainfall-generating system.

Performance and event-level agreement

The predicted vs. observed comparison (Figure 4.12) exhibits strong linear correspondence ($r = 0.63$), with accurate reproduction of low and moderate rainfall intensities. A slight underestimation of extreme convective peaks (>50 mm d^{-1}) persists—a known limitation of reanalysis-driven regression models.

Figure 4.12 – Predicted vs. observed rainfall for the best-ranked model (XGBoost + ERA5).

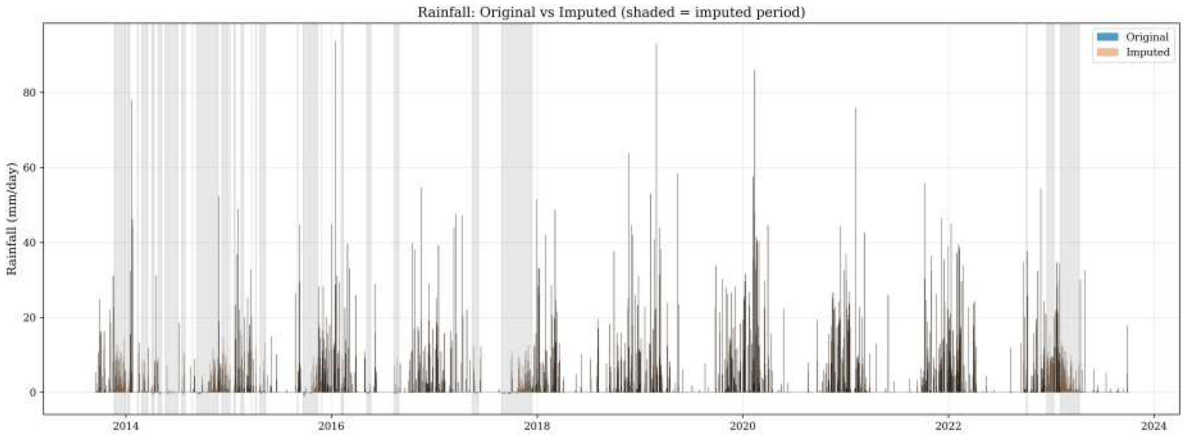


Fonte: The author (2026).

Temporal structure of the reconstructed series

Figure 4.13 displays the complete 2013–2024 series. The model preserves the temporal clustering characteristic of tropical convective systems, avoids artificial smoothing, and maintains realistic dry-wet day sequences. The imputation does not introduce spurious wet periods or artificially extend rainy spells.

Figure 4.13 – Observed and reconstructed rainfall series (2013–2024). Shaded regions correspond to imputed gaps.



Fonte: The author (2026).

STL decomposition and spectral consistency

STL decomposition (Figure 4.14) confirmed that the imputation preserves the hydroclimatic structure of the series. The long-term trend shows no artificial discontinuities at imputed intervals. Annual seasonality is maintained with consistent wet-season amplitude (DJF) and dry-season suppression (JJA). High-frequency residuals remain intermittent and asymmetric, with extreme spikes present in both original and reconstructed versions, though with expected attenuation.

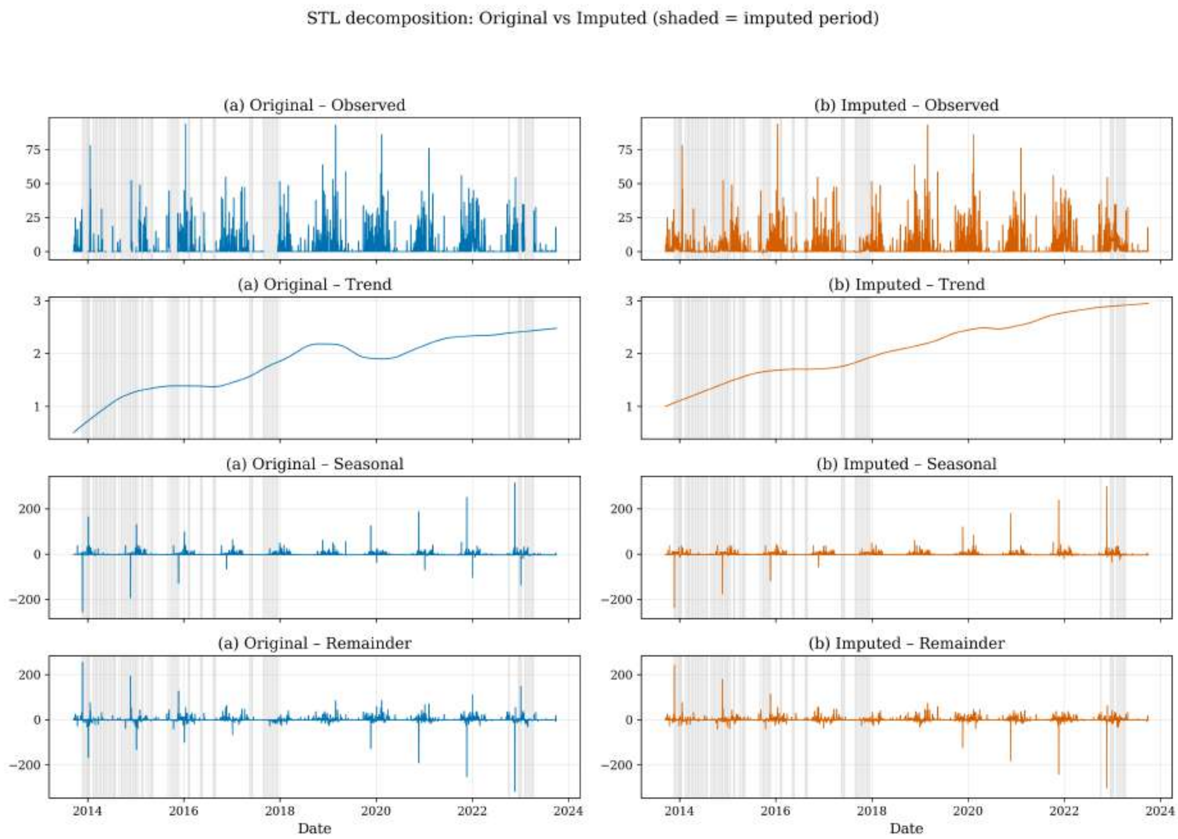
Spectral density of STL residuals (Figure 4.15) confirms that both spectra share dominant peaks in the 6–12-day bands associated with episodic convective organization, preserving the intrinsic intermittency and multi-scale variability of tropical rainfall. Differences at very high frequencies ($\lambda < 0.002$) reflect the natural smoothing of the regression model during extreme short-duration events.

Component-level diagnostics and overall assessment

Complementary diagnostics (Table 4.4) confirm that the reconstructed series maintains hydrologically relevant properties: seasonal amplitude differences $<5\%$; lag-1 and lag-2 autocorrelations differing by less than 0.03; 95th and 99th percentiles slightly attenuated but extreme-day frequency preserved within $\pm 8\%$; and near-identical trend smoothness metrics.

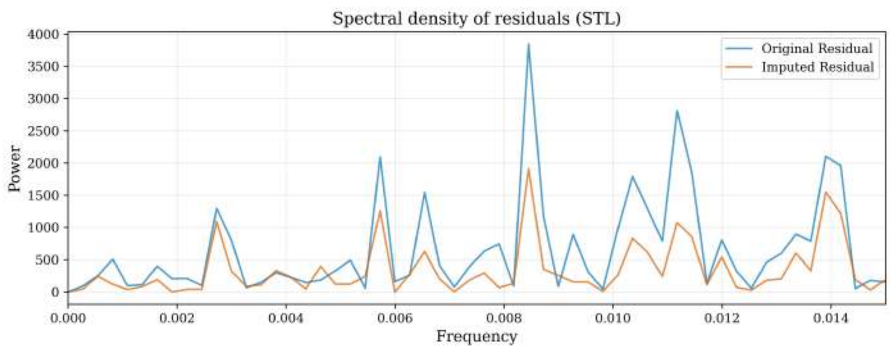
The final reconstructed 2013–2024 dataset combines statistical coherence, hydrological realism, and temporal continuity without distorting the climatological cycle, multi-scale variance structure, event frequency, or long-term trend. It constitutes a robust input for ecohydrological modeling, water balance studies, and machine learning applications that require fully continuous precipitation records.

Figure 4.14 – STL decomposition of original and reconstructed series. Shaded regions indicate imputed periods.



Fonte: The author (2026).

Figure 4.15 – Spectral density of STL residuals for original and reconstructed daily rainfall.



Fonte: The author (2026).

4.4 Discussion

Before contextualizing the specific findings, it is worth noting a structural characteristic of the comparative discussion that follows. The combination of three components evaluated here—unsupervised automatic quality control, multi-source reanalysis and satellite fusion with bias correction, and multi-

Table 4.4 – STL-based diagnostics comparing original and reconstructed series.

Metric	Original	Reconstructed
Seasonal amplitude (mm)	214	207
Trend smoothness index	0.91	0.89
Residual kurtosis	5.72	5.11
Lag-1 autocorrelation	0.41	0.38
Extreme event frequency (P95)	158 days	147 days

Fonte: The author (2026).

objective model selection—has been rarely implemented jointly in the tropical hydrology literature. Most studies address one or at most two of these stages in isolation: either comparing imputation algorithms without prior quality control (ADDI et al., 2022; BANJARA et al., 2025), or evaluating satellite products without a multi-objective selection framework (COUTINHO et al., 2025; BAGILIKO et al., 2025), or applying ML models without bias-corrected exogenous predictors (ALSALEHY; BAILEY, 2025; AWE; VANCE, 2025). The scarcity of directly comparable integrated pipelines is itself a finding of this review and constitutes part of the scientific contribution of this work. Where direct comparisons are unavailable, each methodological component is discussed against the most relevant partial analogues in the literature.

4.4.1 Daily-scale reconstruction: data quality as a precursor to precision

The results show that reliable daily-scale rainfall reconstruction in tropical regions depends critically on integrating unsupervised quality control with physically robust predictors. The performance hierarchy—where the optimal configuration (XGBoost + ERA5 + Isolation Forest) reduces RMSE by 74.6% relative to SARIMA and nearly doubles the KGE (from 0.26 to 0.48)—confirms that traditional univariate models cannot capture the non-linear complexity of tropical precipitation.

The filtering dilemma: LOF vs. Isolation Forest

The most striking finding in the preprocessing stage is the hydrological divergence between two anomaly detection approaches that remove nearly the same fraction of data ($<1.4\%$). Models trained on the Isolation Forest-filtered series dominated the Pareto frontier; models trained on the LOF-filtered series showed systematic variance suppression and underestimation of extremes.

This divergence is explained by the statistical structure of tropical rainfall. Precipitation is a highly skewed, zero-inflated variable whose distribution violates the local density assumptions underlying LOF. Kim, Kim e Kok (2025) showed that Isolation Forest—being non-parametric and based on recursive isolation—does not assume data normality, which is advantageous in heavy-tailed distributions where density-based methods often conflate legitimate convective extremes with noise. This is consistent with

Bărbulescu, Dumitriu e Dragomir (2021), who demonstrated that parametric outlier detection methods applied to hydrometeorological series systematically generate false positives during high-intensity events precisely because such events are rare and locally anomalous by definition. Mastour e Fadul (2023) similarly showed that Isolation Forest is robust for anomaly detection in non-normal environmental data with low contamination rates and zero-inflated or multimodal distributions—conditions that directly describe hourly and daily tropical rainfall.

While density-based methods have been applied to meteorological variables such as temperature or wind speed (ALSALEHY; BAILEY, 2025), our results suggest that in convective regimes, Isolation Forest better preserves the physical integrity of high-impact events. To our knowledge, no study in tropical South America has systematically compared Isolation Forest and LOF specifically for tipping-bucket pluviometric series prior to ML-based reconstruction, representing a gap that this work begins to address.

Predictor hierarchy: atmospheric physics vs. satellite estimation

The overwhelming dominance of ERA5 (present in 64% of the top-50 models) over pure satellite products (8%) points to a shift in how imputation in regions with complex topography and organized convection should be approached.

This finding contrasts with evaluations in other tropical settings. Faramarzzadeh et al. (2023) found that PERSIANN-CDR outperformed other products in East Africa (RMSE of 7.3 mm d^{-1}), where convection is predominantly large-scale and thermally driven. In the Atlantic Forest context, however, our results align with Paredes-Trejo, Barbosa e Lakshmi Kumar (2017) and Andrade et al. (2024), who identified that satellite products such as CHIRPS systematically fail to capture warm-rain processes and orographic convection in coastal Brazil due to the sparse gauge network used in their correction algorithms. Coutinho et al. (2025) further documented systematic dry biases in satellite products in the southeastern Brazilian highlands—the same region as the Matinha tower—particularly during the onset and withdrawal of the wet season governed by the South Atlantic Convergence Zone.

ERA5's superiority lies in its thermodynamic consistency. By assimilating specific humidity, atmospheric pressure, and wind components, ERA5 captures the physical mechanisms of rainfall—moisture transport, convective instability, frontal dynamics—rather than relying on cloud-top temperature proxies. This confirms observations by Salazar et al. (2025) in the Bolivian Andes, where reanalysis reproduced temporal patterns with greater fidelity than satellite estimates by resolving synoptic forcing in complex terrain. The dominance of ERA5 as a predictor—rather than merely as a bias-correction reference—has not been systematically documented for ML-based daily rainfall reconstruction in Atlantic Forest environments, where most comparative studies focus on mean-bias correction rather than imputation performance (ANDRADE et al., 2024; COUTINHO et al., 2025).

Algorithmic stability and non-linearity capture

XGBoost outperformed Random Forest and CatBoost in both stability and accuracy, with its RMSE consistently concentrated at the lower bound ($\approx$2.2 mm d^{-1}). This performance is consistent with recent hydro-informatics literature: Kumar et al. (2025) reported R^2 values approaching 0.99 using XGBoost for daily forecasting, substantially outperforming linear alternatives. Okolie et al. (2024) similarly showed that XGBoost achieves superior stability over RF in rainfall estimation tasks where multiple correlated predictors interact non-linearly, attributing this to the sequential residual-fitting mechanism that progressively corrects errors that bagging-based methods average out.

In imputation tasks with multiple bias-corrected exogenous predictors, the feature space contains complex interaction terms between lagged ERA5 signals and local accumulation windows. Gradient boosting learns these interactions sequentially and explicitly, while RF averages across independent trees that individually cannot resolve high-order feature interactions. This architectural difference explains why RF produces some high-KGE models—it captures mean structure well—but lacks the consistency of XGBoost in the lower RMSE region (BLEIDORN et al., 2024).

The Pareto–Borda evaluation framework proved essential for avoiding the flatness bias documented by Gupta et al. (2009) and Mizukami et al. (2019): models selected by RMSE alone tend to converge toward the conditional mean, producing statistically safe but hydrologically fragile reconstructions. This is consistent with Dubey et al. (2025) and Cinkus et al. (2023), who demonstrated that single-metric optimization penalizes models that preserve distributional variability—a critical distinction for ecohydrological applications where wet/dry cycle representation matters as much as point accuracy.

4.4.2 Hourly reconstruction: the necessity of memory

The contrast between Prophet (KGE $= -0.262$) and LSTM ($r = 0.87$, KGE $= 0.35$) at the hourly scale is not a marginal performance difference—it reflects a fundamental architectural incompatibility between additive decomposition models and the stochastic nature of sub-daily convective rainfall.

The structural limitation of additive models in zero-inflated series

Prophet's failure is physically interpretable. By decomposing the series into smooth trend, seasonality, and holiday components, the model assumes that the signal can be reconstructed from deterministic global patterns. This holds for series where seasonality explains most variance, but hourly tropical precipitation is structurally the opposite: it consists of zero-inflated, stochastically triggered convective bursts superimposed on a weak diurnal cycle (CHIVERS et al., 2018; FARAMARZZADEH et al., 2023). Rose, Rosili e Marsani (2025) demonstrated this limitation in a tropical monsoon environment (Johor Bahru, Malaysia), showing that Prophet fails to capture the irregular stochastic peaks of hourly and daily precipitation, and that ensemble ML methods outperformed it statistically. Patil e Alaskar (2025) further documented that hybrid Neural Prophet–LSTM architectures consistently outperformed standalone Prophet for temporally complex variables, and L, Gurivisetty e Satheesh (2025) confirmed the pattern in a comparative benchmark where the additive model consistently underperformed in high-variance meteorological series. The weak, physically unstructured weekly component identified in our diurnal decomposition further confirms that the model fits noise rather than signal in the non-deterministic portion of the series.

LSTM as a memory architecture for convective dynamics

The LSTM's capacity to reconstruct convective variability ($r = 0.87$) is consistent with growing evidence from tropical and subtropical environments. Arias-muñoz et al. (2023) applied a multivariate LSTM with 24-hour timestep windows for gap-filling of meteorological records in Costa Rica, finding that the architecture successfully learned behavioral patterns in zero-inflated series—a result structurally analogous to ours. Wangwongchai et al. (2023) similarly demonstrated that LSTM is well-adapted to capture long-range dependencies in sequential tropical rainfall data (northern Thailand), outperforming conventional imputation methods in series with frequent gaps. At sub-daily scales, Sheng et al. (2023) applied LSTM to hourly precipitation fusion in the Jianxi basin (southeastern China), showing that LSTM resolves spatio-temporal dependencies in hourly convective rainfall that traditional spatial interpolation cannot capture. Ishida et al. (2024), working at the hourly scale in a rainfall-runoff context, showed that hybrid CNN–LSTM architectures substantially improve peak-event estimation with multi-hour lookback windows—corroborating our choice of a 360-hour window to capture the 15-day antecedent moisture signal identified in the spectral analysis.

These results support the physical hypothesis of Lynn et al. (2021), who emphasize that convective instability and prior moisture content are critical precursors for storm triggering at sub-daily scales. Prophet, by treating the series as a sum of global deterministic components, ignores this state-dependency entirely. The LSTM memory cell, in contrast, encodes the system's inertia through its forget and input gates, enabling it to simulate the rapid onset and decay of convective events.

Residual limitations of hourly reconstruction

Despite the LSTM's structural advantage, a residual underestimation of extreme hourly peaks persists, consistent with the smoothing inherent in MSE-based optimization (BOLOUKI; FAZELI, 2023;

SIMOLO et al., 2010). Arias-muñoz et al. (2023), working with LSTM networks for West African precipitation, identified the same limitation: LSTM captures temporal dependencies effectively but attenuates the amplitude of the highest-intensity events when training samples of extreme hours are scarce. This defines the operational scope of the hourly reconstruction: it is reliable for representing the climatological distribution and persistence of sub-daily convective cycles, but should be used with caution in analyses focused on peak-intensity events ($>$30 mm h^{-1}) in the extreme upper tail.

The direct comparison of Prophet and LSTM for hourly tropical precipitation reconstruction under a multi-objective KGE-based evaluation framework has not been documented in the Atlantic Forest hydrology literature. Most comparable studies evaluate LSTM in isolation (ARIAS-MUÑOZ et al., 2023; WANGWONGCHAI et al., 2023) or compare additive and ML models for temperature rather than precipitation in tropical environments (ABDAKI; SANCHEZ-AZOFEIFA; HAMANN, 2025). The explicit demonstration that Prophet's failure is structural rather than parametric—it cannot be fixed by tuning `changepoint_prior_scale` or adding custom seasonalities—constitutes a methodological delimitation of direct utility for practitioners selecting imputation strategies in data-scarce tropical regions.

4.4.3 Applicability and broader relevance of the framework

Although validated at a single Atlantic Forest site, the framework addresses a problem that is not site-specific: the reconstruction of hydrometeorological records degraded by instrumental failures in data-scarce tropical environments. Three dimensions of broader applicability are relevant.

Problem universality. The combination of factors documented here—extended block failures, zero-inflated distributions, sparse reference networks, and systematic biases in satellite products—is characteristic of automated monitoring stations throughout the Brazilian highlands, the Cerrado, and humid tropical regions more broadly (COUTINHO et al., 2025; BAGILIKO et al., 2025). Addi et al. (2022) reviewed gap-filling methods across 47 African and Latin American hydrological stations and found that more than 60% exhibited missing-data rates exceeding 20%, with block failures as the dominant pattern—identical to the 25.6% rate and block structure documented at Matinha. The need for a pipeline that goes beyond linear interpolation is therefore structural in tropical hydrology (BANJARA et al., 2025; SULLIVAN; KELLY, 2024).

Downstream applications. The reconstructed 2013–2024 series is an input for a range of critical downstream analyses. Continuous daily and hourly records are required for: (i) ecohydrological modeling demanding full water balance closure (Junqueira Junior et al., 2017; OLIVEIRA et al., 2021); (ii) calibration and validation of rainfall-runoff models used in flood early warning systems (ISHIDA et al., 2024; CINKUS et al., 2023); (iii) computation of hydrological indices (Q95, base flow, drought indicators) that are undefined in the presence of gaps (ARSENAULT; BRISSETTE; MARTEL, 2018); and (iv) training future ML models that require fully continuous precipitation inputs.

Operational and low-cost monitoring contexts. The framework is particularly relevant for emerging low-cost monitoring networks (citizen science rain gauges, IoT pluviometers) where data quality and completeness are typically lower than at the Matinha tower. Because the pipeline relies on freely available global products (ERA5, CHIRPS, PERSIANN-CDR) as exogenous predictors, deployment requires no additional instrumentation beyond the target gauge. The complete codebase is publicly available (DOI: https://doi.org/10.5281/zenodo.18096940), enabling direct replication and adaptation without proprietary software or specialized computational infrastructure.

4.4.4 Limitations, transferability, and conditions for replication

Three intrinsic limitations define the scope of the reconstructions, and four structural conditions govern transferability to other sites.

Intrinsic limitations

Propagation of sensor uncertainty. The supervised model uses the tower tipping-bucket gauge as ground truth, inevitably internalizing its instrumental biases. The documented correction factor $P_{\text{manual}} = 1.047 \times P_{\text{bucket}}$ ($r^2 = 0.95$) and aerodynamic undercatch at wind speeds above 5.0 m s^{-1} (Junqueira Junior et al., 2017) imply that the model reconstructs a signal that already carries this physical bias, which propagates to all imputed periods.

Attenuation of extreme events. Residual underestimation of extreme rainfall ($>$50 mm day^{-1}) persists across all configurations, consistent with Bolouki e Fazeli (2023), who showed that MSE-based loss functions incentivize convergence toward the conditional mean at the expense of tail variance. Simolo et al. (2010) documented the same phenomenon in statistical reconstruction methods, noting that extreme-day frequency can be preserved (as shown in Table 4.4) while event intensity is attenuated. KGE as a secondary criterion partially mitigates this but does not eliminate it.

Spatial scale mismatch. A scale gap exists between ERA5 predictors ($\approx$31 km) and point-scale rainfall at the Matinha tower. As Bouregaa (2025) notes, statistical downscaling cannot reproduce sub-kilometer convective microphysics absent from the large-scale dynamic signal, so reconstruction skill may degrade during episodes generated by highly localized convective cells unresolved at the reanalysis grid scale.

Conditions for transferability

(i) Training window—the most restrictive condition. The most critical requirement is a continuous, failure-free segment of sufficient length for model calibration. At Matinha, 4.5 years were available. In sites with shorter continuous segments (<2 years), the statistical representativeness tests described in Section 4.2.2 should be applied before proceeding. Sites with no adequate continuous period would require alternative strategies such as leave-one-year-out cross-validation or transfer learning from a gauged donor catchment (SHEN; TOLSON; MAI, 2022; ARSENAULT; BRISSETTE; MARTEL, 2018).

(ii) Quantile Mapping reference station. Bias correction relies on a nearby reference gauge (INMET-Lavras) with adequate data quality. In remote regions lacking such a reference, ERA5 itself can serve as the QM anchor after cross-validation, but this degrades correction capacity in the upper tail and introduces additional uncertainty (CANNON; SOBIE; MURDOCK, 2015; PADULANO et al., 2025).

(iii) Isolation Forest contamination parameter. The `contamination='auto'` setting is appropriate for the low anomaly rates observed at Matinha (<1.4% of valid days). In networks with higher instrumental noise rates—common where maintenance is infrequent—the contamination parameter should be calibrated against a manually validated reference period to avoid over-filtering legitimate extreme events (KIM; KIM; KOK, 2025; MASTOUR; FADUL, 2023).

(iv) LSTM lookback window. The 360-hour (15-day) window was motivated by spectral analysis identifying dominant intra-annual periodicities near 20 days associated with mesoscale convective organization under the SACZ (BAO et al., 2024; SILVA et al., 2025). In regions with different dominant oscillatory modes—the 30–40-day Madden–Julian Oscillation in the equatorial tropics or 7–10-day synoptic cycles in subtropical regions—the lookback window should be recalibrated against the local power spectrum before training (ISHIDA et al., 2024).

The remaining pipeline components—the Pareto–Borda selection framework, feature engineering strategy (lagged variables, rolling accumulations, harmonic DOY encoding), and model architecture choices— are context-independent by design and require no site-specific adaptation. The public availability of the complete codebase enables practitioners to apply the framework to new sites by modifying only the four parameters described above.

4.5 Conclusions

This study set out to test whether combining rigorous quality control, physically consistent exogenous predictors, and multi-objective model selection could substantially improve rainfall reconstruction in a data-scarce tropical environment. The results confirm it can—and point to some clear practical lessons.

The choice of outlier detection method matters far more than it might appear. Both Isolation Forest and LOF remove a similar fraction of data, but they target different kinds of values. LOF tends to flag high-intensity events as anomalous because they are locally rare, which suppresses the variance of the training series. Isolation Forest removes non-physical artifacts without penalizing the upper tail, making it the better choice for zero-inflated, heavy-tailed tropical rainfall.

ERA5 reanalysis dominated as a predictor, appearing in 64% of the top-50 configurations, while satellite-only products contributed to fewer than 10%. The advantage is physical: ERA5 assimilates three-dimensional atmospheric state variables—humidity, pressure, wind—that directly govern rainfall generation. Satellite products that rely on cloud-top temperature proxies or sparse gauge networks miss the warm-rain and orographic convection processes that drive rainfall in the Atlantic Forest. Bias correction via Quantile Mapping further compresses the error distribution and is essentially non-optional when using any of these global products.

XGBoost, trained on the IF-filtered series with ERA5 predictors, reduced RMSE by 74.6% relative to the SARIMA benchmark and nearly doubled the Kling–Gupta Efficiency. The Pareto–Borda framework was central to this result: without it, single-metric optimization would have selected models that are accurate on average but smooth out the wet/dry variability that matters most for ecohydrological modeling.

At the hourly scale, the LSTM ($r = 0.87$) outperformed Prophet (KGE $= -0.26$) by a wide margin. Prophet's additive decomposition cannot accommodate the stochastic, state-dependent nature of sub-daily convective rainfall. The LSTM's memory architecture can—it learns the antecedent moisture conditions that govern when and how hard it rains. This is not a tuning problem; no amount of adjusting Prophet's hyperparameters will change its fundamental structure.

The reconstructed 2013–2024 series preserves the seasonality, autocorrelation, and distributional properties of the original observations, with extreme-day frequency maintained within 8% and seasonal amplitude within 5%. The recovered record restores the physical information content that gap-filling approaches cannot: the temporal clustering of wet spells that governs soil moisture recharge, the frequency of high-intensity events that drives overland flow generation, and the dry-season intermittency that controls baseflow recession in Atlantic Forest streams. These are not just statistical properties—they are the signatures of the hydrological processes that continuous rainfall forcing is meant to represent. The openly available pipeline can be adapted to other tropical monitoring stations with minimal configuration changes, extending the capacity to reconstruct process-relevant precipitation records in regions where hydrological understanding is most limited and most needed.

Acknowledgements

This study was supported by the National Council for Scientific and Technological Development (CNPq, grant 402066/2023-5) and the Coordination for the Improvement of Higher Education Personnel (CAPES, Finance Code 001). The authors thank the staff of the Meteorological Tower of Matinha (UFLA) for long-term data collection and maintenance.

REFERENCES

ABDAKI, M.; SANCHEZ-AZOFEIFA, A.; HAMANN, H. F. A Machine Learning Approach for Filling Long Gaps in Eddy Covariance Time Series Data in a Tropical Dry Forest. *Journal of Geophysical Research: Biogeosciences*, v. 130, n. 1, p. 1–28, 2025. ISSN 21698961. Doi: 10.1029/2024JG008375.

ABOUZIED, G. A. E. N. A. et al. Completion of the Central Italy daily precipitation instrumental data series from 1951 to 2019. *Geoscience Data Journal*, v. 12, n. 1, p. 1–18, 2025. ISSN 20496060. Doi: 10.1002/gdj3.267.

ADDI, M. et al. Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. *Hydrological Sciences Journal*, Taylor & Francis, v. 67, n. 4, p. 613–627, 2022. ISSN 21503435. Doi: 10.1080/02626667.2022.2030868. Disponível em: https://doi.org/10.1080/02626667.2022.2030868.

AL-RAWAS, G. et al. Integrated quantile mapping and spatial clustering for robust bias correction of satellite precipitation in data-sparse regions. *Sustainability*, MDPI, v. 17, n. 18, p. 8321, 2025. Doi: 10.3390/su17188321.

ALSALEHY, A. S.; BAILEY, M. Improving Time Series Data Quality: Identifying Outliers and Handling Missing Values in a Multilocation Gas and Weather Dataset. *Smart Cities*, 2025. Doi: 10.3390/smartcities8030082.

AMINI, A.; DOLATSHAHI, M.; KERACHIAN, R. Real-time rainfall and runoff prediction by integrating BC-MODWT and automatically-tuned DNNs: Comparing different deep learning models. *Journal of Hydrology*, 2024. Doi: 10.1016/j.jhydrol.2024.130804.

AMJAD, M.; SHAHZAD, N.; UL, W. Accuracy assessment of potential alternatives to observed precipitation for extremes over complex topography. *Theoretical and Applied Climatology*, Springer Vienna, 2025. ISSN 1434-4483. Doi: 10.1007/s00704-024-05232-x. Disponível em: https://doi.org/10.1007/s00704-024-05232-x.

ANARAKI, M. V. et al. Modeling of monthly rainfall–runoff using various machine learning techniques in wadi ouahrane basin, algeria. *Water*, MDPI, v. 15, n. 20, p. 3576, 2023. Disponível em: https://doi.org/10.3390/w15203576.

ANDRADE, J. M. et al. Efficiency of global precipitation datasets in tropical and subtropical catchments revealed by large sampling hydrological modelling. *Journal of Hydrology*, 2024. Doi: 10.1016/j.jhydrol.2024.131016.

ANSARI, R.; GROSSI, G. Performance evaluation of raw and bias-corrected ERA5 precipitation data with respect to extreme precipitation analysis: case study in Upper Jhelum Basin, South Asia. *Theoretical and Applied Climatology*, v. 150, p. 1409–1424, 2022. Doi: 10.1007/s00704-022-04239-6.

ARAYA, D. et al. Towards robust seasonal streamflow forecasts in mountainous catchments: impact of calibration metric selection in hydrological modeling. *Hydrology and Earth System Sciences*, 2023. Doi: 10.5194/hess-27-4385-2023.

ARIAS-MUÑOZ, A. C. et al. Comparative analysis of traditional methods and a deep learning approach for multivariate imputation of missing values in the meteorological field Resumen. p. 33–50, 2023. Doi: https://www.redalyc.org/journal/6998/699878490005/index.html.

ARSENAULT, R.; BRISSETTE, F.; MARTEL, J.-L. The hazards of split-sample validation in hydrological model calibration. *Journal of Hydrology*, 2018. Doi: 10.1016/j.jhydrol.2018.09.027.

ASHOURI, H. et al. PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bulletin of the American Meteorological Society*, v. 96, n. 1, p. 69–83, 2015. ISSN 00030007. Doi: 10.1175/BAMS-D-13-00068.1.

AWE, O. O.; VANCE, E. A. (Ed.). *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network*. [*S.l.*]: Springer Nature, 2025. (STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health). Doi: 10.1007/978-3-031-72215-8. ISBN 9783031722141.

BACK, Á. J.; GALATTO, S. L.; SOUZA, G. da S. Analysis of the seasonality of rain in Brazil from 1990 to 2022. *Concilium*, 2024. Doi: 10.53660/CLM-2770-24B27.

BAGILIKO, J. et al. Bias correction of satellite and reanalysis products for daily rainfall occurrence and intensity. 2025. Preprint. arXiv:2510.27456v1.

BANJARA, I. P. et al. AI for enhanced water quality data imputation: a deep learning perspective. *Environmental Science: Advances*, Royal Society of Chemistry, v. 4, p. 817–823, 2025. Doi: 10.1039/d4va00367e.

BAO, J. et al. Intensification of daily tropical precipitation extremes from more organized convection. *Science Advances*, v. 10, n. 8, p. 1–11, 2024. ISSN 23752548. Doi: 10.1126/sciadv.adj6801.

BĂRBULESCU, A.; DUMITRIU, C. S.; DRAGOMIR, F.-L. Detecting aberrant values and their influence on the time series forecast. In: IEEE. *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. [*S.l.*], 2021. p. 1–5. Doi: 10.1109/ICECCME52200.2021.9591085.

BAşAğAOğLU, H. et al. A Review on Interpretable and Explainable Artificial Intelligence in Hydroclimatic Applications. *Water (Switzerland)*, v. 14, n. 8, 2022. ISSN 20734441. Doi: 10.3390/w14081230.

BEKIĆ, D.; LESKOVAR, K. Evaluating CHIRPS and ERA5 for Long-Term Runoff Modelling with SWAT in Alpine Headwaters. *Water*, MDPI, v. 17, p. 2116, 2025. Doi: 10.3390/w17142116.

BLEIDORN, M. T. et al. Investigation of using missing data imputation methodologies effect on the SARIMA model performance: application to average monthly flows. *Revista Brasileira de Recursos Hidricos*, v. 29, p. 1–18, 2024. ISSN 23180331. Doi: 10.1590/2318-0331.292420230131.

BOLOUKI, H.; FAZELI, M. Evaluation of Multivariate Rainfall Disaggregation Performance Using MuDRain Model (Case Study: North East of Hormozgan Province). *Amirkabir Journal of Civil Engineering*, v. 54, n. 12, p. 941–944, 2023. Doi: 10.22060/ceej.2022.21063.7607.

BORMA, L. S. et al. Beyond Carbon: The Contributions of South American Tropical Humid and Subhumid Forests to Ecosystem Services. *Reviews of Geophysics*, 2022.

BOUREGAA, T. Comparative evaluation of machine learning models for regional agricultural drought prediction in algeria using shap analysis. *Natural Hazards*, Springer, p. 1–72, 2025.

BOZORG-HADDAD, O. et al. Development of a Comparative Multiple Criteria Framework for Ranking Pareto Optimal Solutions of a Multiobjective Reservoir Operation Problem. *Journal of Irrigation and Drainage Engineering*, 2016. Doi: 10.1061/(ASCE)IR.1943-4774.0001028.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Doi: 10.1023/A:1010933404324.

BREUNIG, M. M. et al. LOF: Identifying Density-Based Local Outliers. *SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, n. June, p. 93–104, 2000. Doi: 10.1145/342009.335388.

CANNON, A. J.; SOBIE, S. R.; MURDOCK, T. Q. Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? *Journal of Climate*, v. 28, n. 17, p. 6938–6959, 2015. Doi: 10.1175/JCLI-D-14-00754.1.

CARNEIRO, M. S. et al. Tree species diversity and forest structure in small Atlantic Forest fragments. *Forest Ecology and Management*, 2023. Metadatos completos pendientes de verificación.

CAU, A. C. et al. Water Balance in an Atlantic Forest Remnant: Focus on Representative Tree Species. *Forests*, v. 16, n. 5, p. 1–18, 2025. ISSN 19994907. Doi: 10.3390/f16050812.

CHAKRI, A. et al. Spatial Bias Correction of ERA5_Ag Reanalysis Precipitation Using Machine Learning Models in Semi-Arid Region of Morocco. *Atmosphere*, MDPI, v. 16, p. 1234, 2025. Doi: 10.3390/atmos16111234.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM Computing Surveys*, v. 41, n. 3, p. 1–58, 2009.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [*S.l.*]: ACM, 2016. p. 785–794. Doi: 10.1145/2939672.2939785.

CHIVERS, B. D. et al. Imputation of missing sub-hourly precipitation data in a large sensor network: a machine learning approach. 2018. Manuscript; journal.

CINKUS, G. et al. When best is the enemy of good – critical evaluation of performance criteria in hydrological models. *Hydrology and Earth System Sciences*, 2023. Doi: 10.5194/hess-27-2397-2023.

CLIFTON, D. A. et al. A review of novelty detection. *Signal Processing*, v. 99, p. 215–249, 2014.

COSTA, M. O.; CAMPOS, R. M.; SOARES, C. G. Enhancing the accuracy of metocean hindcasts with machine learning models. *Ocean Engineering*, Elsevier Ltd, v. 287, n. P1, p. 115724, 2023. ISSN 0029-8018. Doi: 10.1016/j.oceaneng.2023.115724. Disponível em: https://doi.org/10.1016/j.oceaneng.2023.115724.

COUTINHO, E. R. et al. Application of Artificial Neural Networks (ANNs) in the Gap Filling of Meteorological Time Series. *Revista Brasileira de Meteorologia*, 2018. Doi: 10.1590/0102-7786332013.

COUTINHO, J. V. et al. Uncovering Rainfall Trend Discrepancies in Brazil (1983–2022): Insights From Ground-Based and Blended Datasets. *International Journal of Climatology*, 2025. Doi: 10.1002/joc.70070.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967. Doi: 10.1109/TIT.1967.1053964.

DABRAL, P.; MURRY, M. Z. Modelling and forecasting of rainfall time series using sarima. *Environmental Processes*, Springer, v. 4, n. 2, p. 399–419, 2017. Doi: 10.1007/s40710-017-0226-y.

DASGUPTA, A. et al. Connecting global-to-local hydrological modelling and forecasting: challenges and scientific advances. *Journal of Flood Risk Management*, 2023.

DIEZ-SIERRA, J.; JESUS, M. del. Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. *Journal of Hydrology*, 2020. Doi: 10.1016/j.jhydrol.2020.124789.

DUBEY, S. et al. Physics-guided Emulators Reveal Resilience and Fragility under Operational Latencies and Outages. p. 1–45, 2025. Doi: arXiv:2510.18535v1.

EHRGOTT, M. *Multicriteria Optimization*. 2. ed. [*S.l.*]: Springer, 2005.

FARAMARZZADEH, M. et al. Application of Machine Learning and Remote Sensing for Gap-filling Daily Precipitation Data of a Sparsely Gauged Basin in East Africa. *Environmental Processes*, v. 10, p. 8, 2023. Doi: 10.1007/s40710-023-00625-y.

FERREIRA, R. M. A.; MENDES, T. G. d. L. Average Spatialization of rainfall in the city of Recife and evaluation of intense rain episodes using geostatistics. *Journal of Hyperspectral Remote Sensing*, v. 12, p. 225–231, 2022. Doi: 10.29150/jhrs.v12.5.p225-231.

FOWLER, K. et al. Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function. *Water Resources Research*, v. 54, n. 5, p. 3392–3408, 2018. ISSN 19447973. Doi: 10.1029/2017WR022466.

FUNK, C. et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, v. 2, p. 150066, 2015. Doi: 10.1038/sdata.2015.66.

GHODRATI, M.; DARIANE, A. B. Evaluation of input variable selection methods in artificial neural networks for estimating missing daily precipitation. *Hydrological Sciences Journal*, Taylor & Francis, v. 69, n. 12, p. 1677–1689, 2024. ISSN 21503435. Doi: 10.1080/02626667.2024.2387156. Disponível em: https://doi.org/10.1080/02626667.2024.2387156.

GUAUQUE-MELLADO, D. et al. Long-term water balance dynamics in a Brazilian neotropical forest: Insights from seven years of continuous observation. *Journal of Environmental Management*, v. 398, 2026. Doi: 10.1016/j.jenvman.2025.128475.

GUAUQUE-MELLADO, D. et al. Evapotranspiration under Drought Conditions: The Case Study of a Seasonally Dry Atlantic Forest. *Atmosphere*, 2022. Doi: 10.3390/atmos13060871.

GUPTA, H. V. et al. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, v. 377, n. 1–2, p. 80–91, 2009. Doi: 10.1016/j.jhydrol.2009.08.003.

HERSBACH, H. et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, v. 146, n. 730, p. 1999–2049, 2020. Doi: 10.1002/qj.3803.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997. ISSN 08997667. Doi: 10.1162/neco.1997.9.8.1735.

HODSON, T. O. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, v. 15, n. 14, p. 5481–5487, 2022. Doi: 10.5194/gmd-15-5481-2022.

ISHIDA, K. et al. Use of one-dimensional CNN for input data size reduction in LSTM for improved computational efficiency and accuracy in hourly rainfall-runoff modeling. *Journal of Environmental Management*, Elsevier Ltd, v. 359, n. January, 2024. ISSN 10958630. Doi: 10.1016/j.jenvman.2024.120931. Disponível em: https://doi.org/10.1016/j.jenvman.2024.120931.

JIAO, D. et al. Evaluation of spatial-temporal variation performance of ERA5 precipitation data in China. *Scientific Reports*, v. 11, n. 1, p. 1–13, 2021. ISSN 20452322. Doi: 10.1038/s41598-021-97432-y.

Junqueira Junior, J. A. et al. Time-stability of soil water content (SWC) in an Atlantic Forest - Latosol site. *Geoderma*, 2017. Doi: 10.1016/j.geoderma.2016.10.034.

KIM, C.-S.; KIM, C.-R.; KOK, K.-H. Outlier Detection in Hydrological Data Using Machine Learning: A Case Study in Lao PDR. *Water*, 2025. Doi: 10.3390/w17213120.

KLING, H.; FUCHS, M.; PAULIN, M. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, v. 424-425, p. 264–277, 2012.

KNOBEN, W. J. M.; FREER, J. E.; WOODS, R. A. Technical note: Inherent benchmark or not? comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, v. 23, p. 4323–4331, 2019.

KUMAR, G. P.; DWARAKISH, G. S. Comparison of the Multiple Imputation Approaches for Imputing Rainfall Data: A Humid Tropical River Basin Case Study. *Water Conservation Science and Engineering*, Springer Nature Singapore, v. 10, n. 2, 2025. ISSN 23645687. Doi: 10.1007/s41101-025-00403-x. Disponível em: https://doi.org/10.1007/s41101-025-00403-x.

KUMAR, V. et al. A Comparative Study of Machine Learning Models for Daily and Weekly Rainfall Forecasting. *Water Resources Management*, Springer Netherlands, v. 39, n. 1, p. 271–290, 2025. ISSN 1573-1650. Doi: 10.1007/s11269-024-03969-8. Disponível em: https://doi.org/10.1007/s11269-024-03969-8.

L, N. S.; GURIVISETTY, S.; SATHEESH, R. Statistical Load Forecasting in Power Systems: A Comparative Study of SARIMA and Prophet Models. *2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT)*, 2025. Doi: 10.1109/CE2CT64011.2025.10941664.

LEMOS, F. C. et al. Spatiotemporal distribution of precipitation and its characteristics under tropical atmospheric systems of Brazil: Insights from a large sub-hourly database. *Hydrological Processes*, v. 37, n. 11, p. 1–19, 2023. ISSN 10991085. Doi: 10.1002/hyp.15017.

LI, X. et al. Statistical Bias Correction of Precipitation Forecasts Based on Quantile Mapping on the Sub-Seasonal to Seasonal Scale. *Remote Sensing*, 2023. Doi: 10.3390/rs15071743.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: IEEE. *2008 Eighth IEEE International Conference on Data Mining (ICDM)*. [*S.l.*], 2008. p. 413–422. Doi: 10.1109/ICDM.2008.17.

LIU, S.; LI, W.; DUAN, Q. Spatiotemporal Variations in Precipitation Forecasting Skill of Three Global Subseasonal Prediction Products over China. *Journal of Hydrometeorology*, v. 24, n. 11, p. 2075–2090, 2023. ISSN 15257541. Doi: 10.1175/JHM-D-23-0071.1.

LYNN, B. et al. Impacts of Non-Local versus Local Moisture Sources on a Heavy (and Deadly) Rain Event in Israel. p. 1–21, 2021. Doi: 10.3390/atmos12070855.

MAGALHÃES, K. M. et al. Biodiversity of aquatic environments in a peri-urban atlantic forest protected remnant: A checklist. *Biota Neotropica*, v. 19, n. 4, 2019. ISSN 16760611. Doi: 10.1590/1676-0611-bn-2019-0761.

MARQUES, M. C. M.; GRELLE, C. E. V. (Ed.). *The Atlantic Forest: History, Biodiversity, Threats and Opportunities of the Mega-diverse Forest*. [*S.l.*]: Springer Nature Switzerland AG, 2021. ISBN 978-3-030-55321-0.

MASTOUR, A.; FADUL, A. *Anomaly Detection based on Isolation Forest and Local Outlier Factor*. Dissertação (Mestrado) — African Institute for Mathematical Sciences (AIMS) South Africa, 2023. Supervised by Prof. Bruce Bassett, University of Cape Town.

MELLO, C. R. et al. Water balance in a neotropical forest catchment of southeastern Brazil. *Catena*, Elsevier, v. 173, n. July 2018, p. 9–21, 2019. ISSN 0341-8162. Doi: 10.1016/j.catena.2018.09.046. Disponível em: https://doi.org/10.1016/j.catena.2018.09.046.

MIRALLES, D. G. et al. Global canopy interception from satellite observations. *Journal of Geophysical Research Atmospheres*, v. 115, n. 16, p. 1–8, 2010. ISSN 01480227. Doi: 10.1029/2009JD013530.

MIZUKAMI, N. et al. On the choice of calibration metrics for “high-flow” estimation using hydrologic models. *Hydrology and Earth System Sciences*, 2019. Doi: 10.5194/hess-23-2601-2019.

MOKHTARI, S.; SHARAFATI, A.; RAZIEI, T. Satellite-based streamflow simulation using CHIRPS satellite precipitation product in Shah Bahram Basin, Iran. *Acta Geophysica*, 2022. Doi: 10.1007/s11600-021-00724-0.

MONEGO, V. S.; ANOCHI, J. A.; de Campos Velho, H. F. South America Seasonal Precipitation Prediction by Gradient-Boosting Machine-Learning Approach. *Atmosphere*, v. 13, n. 2, 2022. ISSN 20734433. Doi: 10.3390/atmos13020243.

MOREIRA, R. M. et al. Precipitation Variability for Protected Areas of Primary Forest and Pastureland in Southwestern Amazônia. *Climate*, v. 11, n. 2, 2023. ISSN 22251154. Doi: 10.3390/cli11020027.

MUÑOZ-SABATER, J. et al. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 2021. Doi: 10.5194/essd-13-4349-2021.

MYERS, N. et al. Biodiversity hotspots for conservation priorities. *Nature*, v. 403, n. 6772, p. 853–858, 2000. Doi: 10.1038/35002501.

OKOLIE, C. et al. Assessment of explainable tree-based ensemble algorithms for the enhancement of Copernicus digital elevation model in agricultural lands. *International Journal of Image and Data Fusion*, 2024. Doi: 10.1080/19479832.2024.2329563.

OLIVEIRA, V. A. de et al. Spatiotemporal modelling of soil moisture in an Atlantic forest through machine learning algorithms. *European Journal of Soil Science*, v. 72, n. 5, p. 1969–1987, 2021. ISSN 13652389. Doi: 10.1111/ejss.13123.

PADULANO, R. et al. Quantile-based bias-correction of extreme rainfall: Pros & cons of popular methods for climate signal preservation. *Journal of Hydrology*, Elsevier B.V., v. 653, n. December 2024, p. 132814, 2025. ISSN 00221694. Doi: 10.1016/j.jhydrol.2025.132814. Disponível em: https://doi.org/10.1016/j.jhydrol.2025.132814.

PAREDES-TREJO, F. J.; BARBOSA, H. A.; Lakshmi Kumar, T. V. Validating CHIRPS-based satellite precipitation estimates in Northeast Brazil. *Journal of Arid Environments*, 2017. Doi: 10.1016/j.jaridenv.2016.12.009.

PATIL, D.; ALASKAR, K. A Hybrid Approach for Rainfall Prediction : Leveraging Prophet for Seasonality and LightGBM for Residuals A Hybrid Approach for Rainfall Prediction : Leveraging Prophet for Seasonality and LightGBM for Residuals. n. April, 2025. Doi: 10.1201/9781003650010-56.

PIADEH, F. et al. Event-based decision support algorithm for real-time flood forecasting in urban drainage systems using machine learning modelling. *Environmental Modelling & Software*, 2023. Doi: 10.1016/j.envsoft.2023.105772.

PINTHONG, S. et al. Optimized Rainfall Imputation Using ERA5-Land and Tree-Based Machine Learning: A Scalable Framework for Data-Sparse Regions. *Earth Systems and Environment*, 2025. Doi: 10.1007/s41748-025-00787-9.

PROKHORENKOVA, L. et al. CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*. [*S.l.*]: Curran Associates, Inc., 2018. v. 31. Doi: 10.48550/arXiv.1706.09516.

PUGAS, A. F. The hunt for data: obstacles faced by researchers in the search for accurate information on precipitation in Brazil. *International Journal of Hydrology*, v. 7, n. 2, p. 88–91, 2023. Doi: 10.15406/ijh.2023.07.00344.

RIBEIRO, M. C. et al. The Brazilian Atlantic Forest : How much is left , and how is the remaining forest distributed ? Implications for conservation. *Biological Conservation*, Elsevier Ltd, v. 142, n. 6, p. 1141–1153, 2009. ISSN 0006-3207. Doi: 10.1016/j.biocon.2009.02.021. Disponível em: http://dx.doi.org/10.1016/j.biocon.2009.02.021.

RODRIGUES, A. F. et al. Modeling canopy interception under drought conditions : The relevance of evaporation and extra sources of energy. *Journal of Environmental Management*, v. 292, n. April, 2021. Doi: 10.1016/j.jenvman.2021.112710.

RODRIGUES, A. F. et al. Soil water content and net precipitation spatial variability in an Atlantic forest remnant. *Acta Scientiarum - Agronomy*, v. 42, p. 1–13, 2020. ISSN 18078621. Doi: 10.4025/actasciagron.v42i1.43518.

ROSE, F. Z. C.; ROSILI, N. A. K.; MARSANI, M. F. Comparison of machine learning model performance for predicting the climate variables in johor bahru, malaysia. *Scientific Reports*, Nature Publishing Group UK London, v. 15, n. 1, p. 23465, 2025. Disponível em: 10.1038/s41598-025-08033-y.

SALAZAR, Á. et al. Spatio-Temporal Evaluation of MSWEP, CHIRPS and ERA5-Land Reveals Regional-Specific Responses Across Complex Topography in Bolivia. *Atmosphere*, 2025. Doi: 10.3390/atmos16111281.

SANTOS, P. S. D. et al. OPSBC : A method to sort Pareto-optimal sets of solutions in multi-objective problems. *Expert Systems With Applications*, Elsevier Ltd, v. 250, n. February, p. 123803, 2024. ISSN 0957-4174. Doi: 10.1016/j.eswa.2024.123803. Disponível em: https://doi.org/10.1016/j.eswa.2024.123803.

SHAHRESTANI, S.; SANISLAV, I. How does dimensionality influence outlier detection effectiveness in multivariate geochemical data? insights from LOF and IF methods. *Earth Science Informatics*, Springer Berlin Heidelberg, v. 18, n. 1, 2025. ISSN 18650481. Doi: 10.1007/s12145-024-01611-0. Disponível em: https://doi.org/10.1007/s12145-024-01611-0.

SHEN, H.; TOLSON, B. A.; MAI, J. Time to Update the Split-Sample Approach in Hydrological Model Calibration. *Water Resources Research*, 2022. Doi: 10.1029/2021WR031523.

SHENG, S. et al. An Integrated Framework for Spatiotemporally Merging Multi-Sources Precipitation Based on F-SVD and ConvLSTM. *Remote Sensing*, 2023. Doi: 10.3390/rs15123135.

SILVA, S. T. da et al. Rainfall forecast in brazil using machine learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, AIP Publishing, v. 35, n. 7, 2025. Doi: 10.1063/5.0259222.

SILVA, V. O.; MELLO, C. R. Meteorological droughts in part of southeastern brazil: Understanding the last 100 years. *Anais da Academia Brasileira de Ciências*, SciELO Brasil, v. 93, n. suppl 4, p. e20201130, 2021. Doi: 10.1590/0001-3765202120201130.

SIMOLO, C. et al. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. v. 1576, n. August 2009, p. 1564–1576, 2010. Doi: 10.1002/joc.1992.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and computing*, Springer, v. 14, n. 3, p. 199–222, 2004. Doi: 10.1023/B:STCO.0000035301.49549.88.

SRIWAHYUNI, L. et al. Performance of machine learning for imputing missing daily rainfall data in east java under multiple satellite data models. *Geographia Technica*, v. 20, n. 1, 2025. Doi: 10.21163/GT_2025.201.23.

SU, Q.; ZHANG, Y. et al. Precipitation observing network gaps limit climate change impact assessment. *Nature*, 2026. Título inferido por contexto; verificar metadatos completos.

SULLIVAN, B. O.; KELLY, G. Infilling of high-dimensional rainfall networks through multiple imputation by chained equations. n. May, p. 3075–3091, 2024. Doi: 10.1002/joc.8513.

SURI, A.; AZAD, S. Optimal placement of rain gauge networks in complex terrains for monitoring extreme rainfall events: a review. *Theoretical and Applied Climatology*, 2024. Doi: 10.1007/s00704-024-04856-3.

TASZAREK, M. et al. Global climatology and trends in convective parameters from ERA5 and rawinsonde data. *npj Climate and Atmospheric Science*, 2021.

TAYLOR, S. J.; LETHAM, B. Forecasting at Scale. *American Statistician*, v. 72, n. 1, p. 37–45, 2018. ISSN 15372731. Doi: 10.1080/00031305.2017.1380080.

TEEGAVARAPU, R. S. Statistical corrections of spatially interpolated missing precipitation data estimates. *Hydrological Processes*, v. 28, n. 11, p. 3789–3808, 2014. ISSN 10991085. Doi: 10.1002/hyp.9906.

VIANA, L. B. et al. Análise comparativa entre dados de precipitação gerados pelo “Climate Prediction Center – CPC” versus dados observados para diferentes biomas no Brasil. *Ciência e Natura*, v. 45, n. esp. 2, p. e81776, 2023. ISSN 0100-8307. Doi: 10.5902/2179460x81776.

VILA, D. A. et al. MERGE: A methodology for merging verified satellite and surface observations. *Meteorological Applications*, 2009.

WANG, F.; TIAN, D.; CARROLL, M. Customized deep learning for precipitation bias correction and downscaling. *Geoscientific Model Development*, 2023. Doi: 10.5194/gmd-16-535-2023.

WANGWONGCHAI, A. et al. Imputation of missing daily rainfall data; A comparison between artificial intelligence and statistical techniques. *MethodsX*, v. 11, n. September, p. 102459, 2023. ISSN 22150161. Doi: 10.1016/j.mex.2023.102459.

YESTE, P. et al. A Pareto-Based Sensitivity Analysis and Multiobjective Calibration Approach for Integrating Streamflow and Evaporation Data. *Water Resources Research*, v. 59, p. e2022WR033235, 2023.

ZOUNEMAT-KERMANI, M. et al. Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*, 2021. Doi: 10.1016/j.jhydrol.2021.126266.

5 MEMORY-DRIVEN RAINFALL INTERCEPTION IN A SEASONAL ATLANTIC FOREST: A PHYSICAL-ENHANCED MACHINE LEARNING APPROACH TO IDENTIFY THE KEY DETERMINANTS OF SPATIOTEMPORAL VARIABILITY

Abstract

Rainfall interception (I) plays a critical role in the hydrological cycle and water balance, yet its predictive modeling remains a challenge in semideciduous tropical forests characterized by large spatial and temporal phenological variations. Traditional analytical models, which typically assume stationary canopy parameters, and standard Machine Learning (ML) models trained solely on concurrent meteorological variables often fail to capture the hysteretic effect of antecedent conditions. In this study, we employed high-resolution meteorological data and in-situ throughfall and stemflow measurements to train and evaluate four ML algorithms: Random Forest (RF), eXtreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), and Categorical Boosting (CatBoost). To account for the memory effect of the forest canopy, we introduced antecedent rolling averages (24 h to 720 h) of key hydrological variables as exogenous predictors. Under spatial Leave-One-Point-Out (LOPO) cross-validation, XGBoost emerged as the most accurate and robust predictor, achieving a coefficient of determination (R^2) of 0.652, a Root Mean Square Error (RMSE) of 0.200, and a Kling-Gupta Efficiency (KGE) of 0.696. Feature importance analysis via SHapley Additive exPlanations (SHAP) confirmed that antecedent precipitation and prior relative humidity are the predominant drivers of interception, significantly outweighing concurrent meteorological factors. By capturing the hysteretic dynamics of canopy wetting and drying cycles, this memory-driven ML framework offers a robust, data-driven alternative for operational ecohydrological modeling in seasonally dry tropical environments, particularly under ongoing climate change scenarios.

Keywords: rainfall interception; machine learning; Atlantic Forest; hydrological memory; SHAP; effective precipitation; spatiotemporal variability

5.1 Introduction

Rainfall interception (I) and effective precipitation (P_{eff}) are fundamental components of the terrestrial water balance, governing the partitioning of incoming rainfall (P) into canopy storage, throughfall (T), and stemflow (S) (BAKAR; LATIF; LEE, 2023). Interception represents the fraction of precipitation temporarily retained by the canopy and subsequently returned to the atmosphere through evaporation, whereas effective precipitation ($P_{eff} = T + S$) constitutes the net water input that drives soil moisture dynamics, groundwater recharge, and runoff generation (Junqueira Junior et al., 2019; MELLO et al., 2024). As the first hydrological filter in forested catchments, interception exerts a first-order control on both the magnitude and timing of water fluxes reaching the soil system (LUCAS-BORJA et al., 2023; Junqueira Junior et al., 2017).

Reported interception rates vary widely across forest ecosystems. In remnants of the Brazilian Atlantic Forest, annual interception typically ranges between 12.3% and 18.5% (Junqueira Junior et al., 2017; Junqueira Junior et al., 2019; SARI; PAIVA; PAIVA, 2016), while losses exceeding 40% of gross precipitation have been documented during dry periods or in fragmented landscapes (FILHO et al., 2019; RODRIGUES et al., 2021a). Comparable variability has been reported globally, with interception accounting for approximately 12% of annual rainfall in subtropical forests (CHEN; GUESTRIN, 2016) and up to 25% in humid temperate systems (SPITTLEHOUSE; MALONEY, 2023). A global synthesis across 166 forest sites estimates that 10.5% of continental rainfall is returned to the atmosphere via canopy evaporation (ZHONG et al., 2022). Despite its hydrological relevance, interception remains one of the most difficult components of the water balance to regionalize due to its pronounced spatial heterogeneity, event-scale nonlinearity, and sensitivity to antecedent conditions (MUZYLO et al., 2009; SIDLE; ZIEGLER, 2017).

Historically, interception has been quantified using empirical relationships or physically based analytical models such as the Rutter and Gash formulations (GASH; MORTON, 1978; NÁVAR, 2017). While these models provide a valuable theoretical framework, they conceptualize the canopy as a quasi-static reservoir and require site-specific parameterization of storage capacity (S) and evaporation rates (E/R), parameters that vary substantially across space and time (MUZYLO et al., 2009; LINHOSS; SIEGERT, 2020). As a result, their predictive skill often deteriorates at the individual event scale, where reported coefficients of determination commonly fall between 0.21 and 0.48 (LINHOSS; SIEGERT, 2020). This limitation reflects the fact that interception is not a fixed fraction of rainfall, but a state-dependent process modulated by rainfall intensity, antecedent moisture conditions, canopy structure, and atmospheric demand (YU et al., 2022; SCHMIDT et al., 2020; RODRIGUES et al., 2021a).

Recent advances in machine learning (ML) offer a promising pathway to overcome these constraints by learning complex, nonlinear relationships without assuming temporal stationarity or site-invariant parameters (MOHAMMED; MENGISTE; GEBREMICHAEL, 2025; SREENIVASU et al., 2024). Tree-based ensemble algorithms, such as Random Forest and Gradient Boosting Machines, have demonstrated strong predictive performance in a wide range of hydrological applications (DAS; SACHINDRA; CHANDA, 2022; FALLAHI et al., 2025). However, most ML-based interception studies remain limited in number, often prioritize predictive accuracy over physical interpretability, and rarely evaluate model robustness under both spatial and temporal validation frameworks (WILLARD et al., 2022; OLIVEIRA et al., 2024). Consequently, the extent to which data-driven models can generalize across heterogeneous forest landscapes while preserving physically meaningful behavior remains largely unresolved.

This gap is particularly evident in Atlantic Forest ecosystems, where long-term monitoring networks exist but interception dynamics have rarely been analyzed simultaneously across multiple sites and seasons (Junqueira Junior et al., 2019). Moreover, the interaction between persistent dry spells, antecedent moisture memory, and interception efficiency under a changing climate remains poorly quantified, which limits the ability to assess future water availability in this biodiversity hotspot (RODRIGUES et al., 2021a; GUAUQUE-MELLADO et al., 2026).

This study addresses these challenges by integrating a 10.4-year high-resolution monitoring dataset from 32 plots in a montane semi-deciduous Atlantic Forest remnant with a physically enhanced, feature-based machine-learning framework (as first highlighted by Bezerra et al. (2026)). A physical-enhanced formulation refers to a modeling strategy in which the input variables are explicitly selected, transformed, or constrained so that they represent meaningful physical quantities or mechanisms of the underlying process. Instead of relying solely on abstract or purely data-driven features, the inputs are constructed to reflect conservation laws, material properties, transport phenomena, or other physically interpretable aspects. This enhances model interpretability, improves generalization, and ensures consistency with known physical behavior (BEZERRA et al., 2026). An ensemble of four tree-based algorithms were evaluated (XGBoost, Random Forest, LightGBM, and CatBoost) under complementary spatial (Leave-One-Point-Out) and temporal validation schemes to explicitly contrast spatial transferability and temporal non-stationarity. Hydrological memory is explicitly represented through engineered antecedent precipitation indices, and model transparency is further assessed with SHapley Additive exPlanations (SHAP).

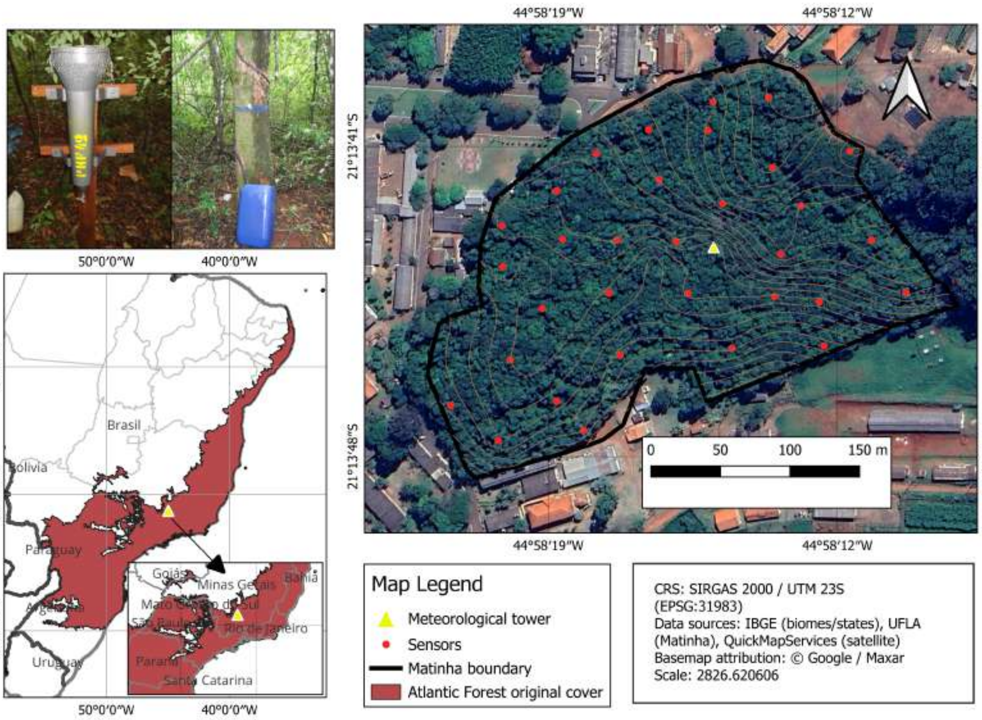
The specific objectives of this work are to: (i) model the spatiotemporal variability of the effective precipitation coefficient (C_{Peff}) at the event scale; (ii) identify dominant physical drivers, threshold behavior, and memory effects governing interception efficiency by means of a physically-enhanced ML framework; and (iii) operationalize the optimized framework into the *Intercep-ML* platform to support real-time ecohydrological assessment. The framework moves beyond static parameterizations toward a dynamic, memory-driven representation of rainfall partitioning, offering a scalable and physically interpretable tool for water-resources management in one of the world's most threatened forest biomes.

5.2 Methodology

5.2.1 Study Area and Site Description

The research was conducted in a remnant forest watershed located in Lavras, Minas Gerais State, southeastern Brazil (Figure 5.1) (21°13′40″ S, 44°57′50″ W; 918–920 m a.s.l.), within the Atlantic Forest biome. The region exhibits a highland tropical climate (Köppen classification: Cwa), with a mean annual precipitation of approximately 1,580 mm, concentrated between October and March. The dry season extends from April to September, with monthly rainfall typically below 50 mm. The mean annual temperature ranges between 19.4°C and 21°C, with minimum values around 14.4–16°C in July and maximums of 22.5–25°C in January. The sampled forest area covers about 100 ha of secondary montane forest, regenerated for approximately 35–40 years following agricultural abandonment. The canopy height averages 25–30 m, with a mean leaf area index (LAI) of 4.2 ± 0.6 Junqueira Junior et al. (2017).

Figure 5.1 – Geographic location of the study site (Matinha Forest Reserve, Lavras, MG) showing the distributed sensor network. Data sources: IBGE, UFLA, and Google/Maxar (SIRGAS 2000 / UTM 23S).



Fonte: The author (2026).

A total of 32 spatial monitoring points were systematically installed using a stratified sampling design to represent different microhabitats (dense canopy, canopy gaps, and transitional zones) and topographic positions. The mean inter-point distance was approximately 40 m. All coordinates were georeferenced using a differential GPS (accuracy < 1 m) under the UTM system (datum SIRGAS 2000, zone 23S). This spatial network captures the micro-scale heterogeneity in canopy structure and hydrological processes typical of the Atlantic Forest ecosystem (Junqueira Junior et al., 2019; RODRIGUES et al., 2021a).

Table 5.1 – Summary of rainfall event dataset used for hydrological and modeling analyses.

Indicator	Value
Study period	22 Sep 2012 – 13 Feb 2023
Continuous monitoring duration	10.39 years
Number of rainfall events	744
Spatial monitoring points	32
Initial event–point observations	23,808
Median gross rainfall per event	10.7 mm
Range of gross rainfall	0.1–282.5 mm

Fonte: The author (2026).

5.2.2 Data Collection Protocol

Field monitoring was carried out continuously from 22 September 2012 to 13 February 2023, encompassing 10.39 years of uninterrupted hydrological observations with a total of 744 discrete rainfall events. Events were defined as one or more precipitation periods separated by at least 12 consecutive hours without rainfall. All events with gross precipitation ($P_e > 0$ mm) were included to ensure full representation of rainfall conditions across the monitoring period. The dataset comprises 23,808 initial event–point observations, of which 21,057 were retained after quality control (QC) procedures. A summary of the dataset characteristics is presented in Table 5.1.

Instrumentation and Measurement

Gross (external) precipitation (P_e) was measured using a tipping-bucket rain gauge (0.2 mm per tip) installed in an open area free from obstructions. Throughfall (P_i) was measured using 32 funnel-type collectors (283 cm^2 open-area each) positioned under the canopy at 1.5 m above the forest floor to avoid splash-in. Stemflow (S_f) was measured using collars installed at breast height on 3–5 representative trees per monitoring point, with flow directed to calibrated collection containers. All volumetric measurements were performed 4 h after each rainfall event to guarantee the complete drainage of the canopy (Junqueira Junior et al., 2019; RODRIGUES et al., 2021a).

5.2.3 Hydrological Variable Computation

For each discrete rainfall event t and spatial monitoring point j, the fundamental hydrological variables were computed following the experimental protocols described by Junqueira Junior et al. (2019). The effective precipitation (P_{eff}) was defined as the sum of throughfall (T) and stemflow (S_f):

$$P_{\text{eff}(j,t)} = T_{(j,t)} + S_{f(j,t)} \tag{5.1}$$

$$I_{(j,t)} = P_{e(t)} - P_{\text{eff}(j,t)} \tag{5.2}$$

$$C_{(j,t)} = \frac{P_{\text{eff}(j,t)}}{P_{e(t)}} \tag{5.3}$$

where $P_{e(t)}$ is the gross precipitation measured in an open area (mm); $T_{(j,t)}$ is the throughfall collected under the canopy at point j (mm); $S_{f(j,t)}$ represents the stemflow generated by representative trees at point j (mm); $P_{\text{eff}(j,t)}$ is the effective precipitation (mm); $I_{(j,t)}$ quantifies the canopy interception losses (mm); and $C_{(j,t)}$ is the effective precipitation coefficient (dimensionless). Any calculated values falling outside physical boundaries (e.g., $C < 0$) were excluded during the quality control stage.

5.2.4 Hydrological Modeling Workflow

The proposed modeling framework integrates hydrological principles with advanced ensemble learning to characterize rainfall partitioning (Figure 5.2). The workflow follows five critical stages:

1. **Data Curation and Standardization:** Raw records were ingested and timestamp-validated to ensure chronological alignment.

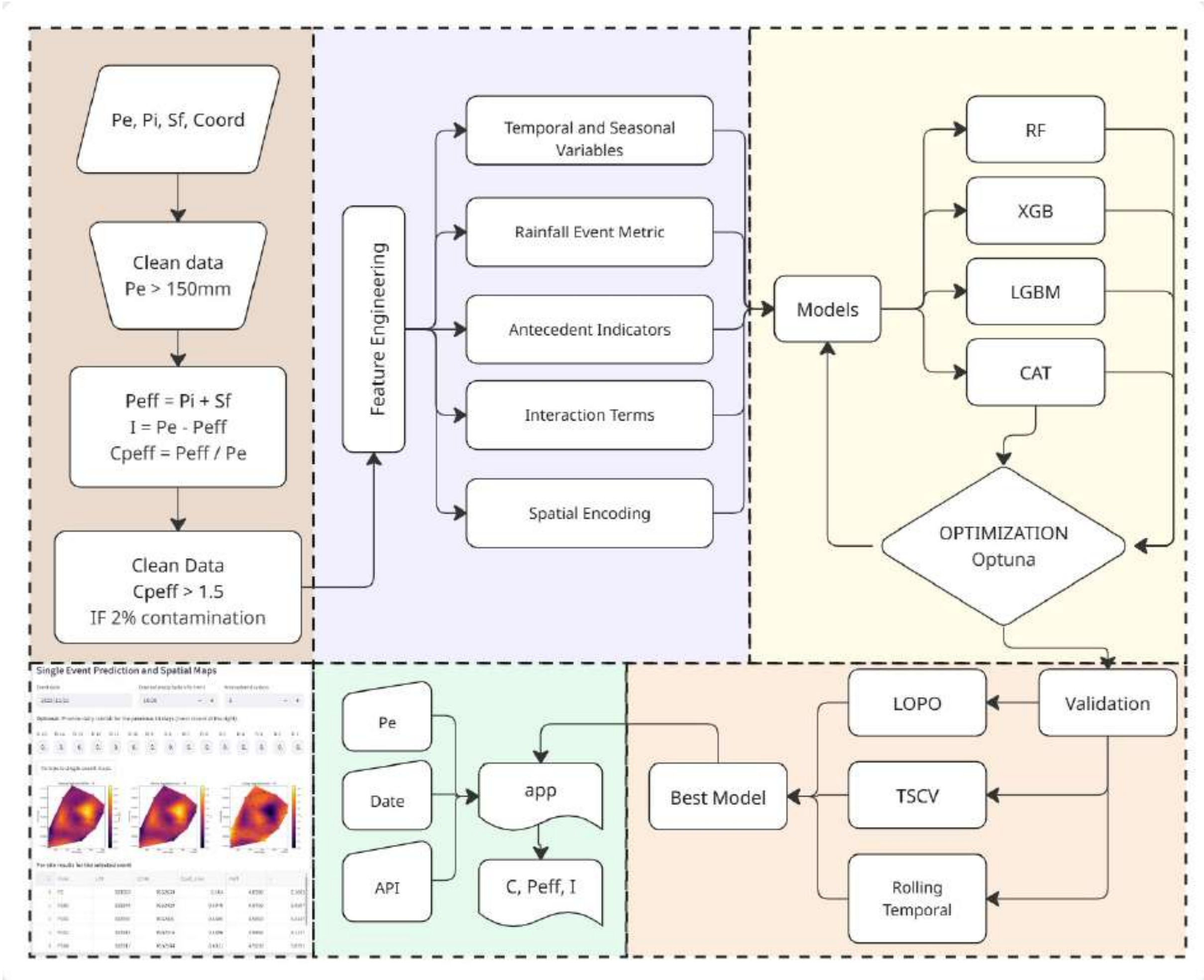
2. **Physical Plausibility Filtering:** Extreme rainfall events ($P_e > 150$ mm) were identified as non-plausible outliers and removed. The effective precipitation coefficient (C) was restricted to $0 < C < 1.5$, an upper limit that accounts for localized rainfall concentration and branch-drip effects in complex multi-layered canopies (Junqueira Junior et al., 2019; ZIMMERMANN; ZIMMERMANN, 2014).

3. **Hydrological Feature Engineering:** A high-dimensional set of 48 engineered features was generated to encode the “hydrological memory” of the system, including Antecedent Precipitation Indices (API), non-linear interaction terms ($P_e \times API_{7d}$), and phenological markers.

4. **Multidimensional Outlier Detection:** Statistical noise was filtered using a month-stratified *Isolation Forest* algorithm with a 2% contamination threshold.

5. **Parallelized Optimization and Robust Validation:** The predictive engine was tuned via a Bayesian optimization framework (Optuna). Models were evaluated using spatial (LOPO) and forward-chaining temporal (TSCV) cross-validation schemes.

Figure 5.2 – Hydrological modeling workflow illustrating sequential steps from raw data preprocessing to model training and validation.



Fonte: The author (2026).

Identification of Critical and Stable Monitoring Points. To examine contrasting spatial behaviors within the monitoring network, two reference points were identified based on the coefficient of variation (CV $= \sigma/\mu$) of XGBoost-predicted C_{Peff} values across the full record. The *Critical Point* is defined as the monitoring location with the highest CV, indicating maximum sensitivity to event-scale variability in rainfall partitioning. The *Stable Point* is the location with the lowest CV, representing the most temporally consistent response. This classification uses model predictions rather than raw observations to minimize the influence of instrument noise, and follows the approach of Rodrigues et al. (2021a) for identifying ecohydrological hotspots within the network.

Quality Control and Outlier Detection

A physical constraint was applied to the effective rainfall coefficient (C), restricting values to $0 < C < 1.5$. Measurements exceeding this range were discarded as potential instrumental artifacts. Statistical outlier detection was subsequently performed using the *Isolation Forest* algorithm (contamination = 0.02), stratified by spatial point and calendar month (LIU; TING; ZHOU, 2008). After all quality-control steps, 21,057 valid event–point observations remained.

5.2.5 Feature Engineering

The feature engineering process was designed to extract hydrologically meaningful features representing event magnitude, antecedent wetness, seasonality, and spatial heterogeneity to build the physically enhanced machine learning models (BEZERRA et al., 2026) and improve interpretability. Following the rationale of ecohydrological modeling (Junqueira Junior et al., 2017), the features were grouped into temporal, rainfall-related, antecedent, interactive, and spatial categories. Table 5.2 summarizes their mathematical formulation and ecohydrological interpretation.

The final dataset consisted of 48 features (16 continuous + 32 categorical dummies) and 21,057 event–point observations, forming a physically interpretable basis for machine learning training. The target variable is the effective precipitation coefficient (C).

5.2.6 Modeling

To capture the intricate non-linear dependencies between atmospheric drivers and canopy structure, this research leveraged four complementary decision-tree ensemble architectures: Random Forest (RF), XGBoost (XGB), CatBoost (CAT), and LightGBM (LGBM). These algorithms were prioritized for their proven robustness in managing ecohydrological datasets, which are typically characterized by high-dimensional non-stationarity and non-normal distributions (BREIMAN, 2001; SREENIVASU et al., 2024). The specific architectural rationale and the optimized hyperparameter search spaces for each implementation are synthesized in Table 5.1.

While Random Forest provides exceptional stability against sensor noise through bagging, the gradient boosting variants (XGB, CAT, and LGBM) were selected for their superior capacity to minimize bias in longitudinal series and identify sharp physical thresholds during minor rainfall events (CHEN; GUESTRIN, 2016; MOHAMMED; MENGISTE; GEBREMICHAEL, 2025). Within the LGBM framework, a monotonic constraint ($P_e \uparrow \Rightarrow C \uparrow$) was enforced to ensure the model's behavior remains

Table 5.2 – Grouped engineered features with their mathematical formulation and ecohydrological rationale used for modeling the effective precipitation coefficient ($C = P_{\text{eff}}/P_e$).

Feature Group	Mathematical Formulation or Description	Ecohydrological Rationale
1. Temporal and Seasonal Variables	$\text{month\_sin} = \sin\left(\frac{2\pi m}{12}\right)$, $\text{month\_cos} = \cos\left(\frac{2\pi m}{12}\right)$. $\text{doy\_sin} = \sin\left(\frac{2\pi d}{365}\right)$, $\text{doy\_cos} = \cos\left(\frac{2\pi d}{365}\right)$. S_m = continuous seasonal index $(0-1)$.	Encode annual and intra-annual periodicity while ensuring smooth transition between dry and wet season, reflecting gradual canopy and phenological changes (BARNETT; BAKER; DOBSON, 2012).
2. Rainfall Event Metrics	$P_{e,t}$ = event rainfall (mm). Quartile classification of $P_{e,t}$ (Q1–Q4).	Represent rainfall magnitude and energy adjusted by antecedent dryness, key drivers of interception and effective rainfall (SIDLE; ZIEGLER, 2017; TARASOVA et al., 2020; ZHANG et al., 2016).
3. Antecedent Indicators	$\text{API}_{t,w} = \sum_{i=1}^{w} P_{e,t-i} e^{-\lambda(i-1)}$, $\lambda = \frac{3}{w}$. $\text{Precip\_sum}_{t,w} = \sum_{i=1}^{w} P_{e,t-i}$. D_t = dry days; E_t = consecutive wet events.	Quantify short- to mid-term moisture memory in canopy and soil, describing how previous rainfall affects interception capacity and rewetting (SIDLE; ZIEGLER, 2017; TARASOVA et al., 2020; ZHANG et al., 2016; LINHOSS; SIEGERT, 2020).
4. Interaction Terms	$X_t = P_{e,t} \times \text{API}_{t,7d}$.	Capture nonlinear effects between current rainfall and antecedent wetness influencing canopy storage and drainage (SIDLE; ZIEGLER, 2017; TARASOVA et al., 2020).
5. Spatial Encoding	One-hot encoding for 32 monitored plots.	Include spatial heterogeneity of canopy, topography, and microclimate conditions (FATHIZADEH et al., 2021).

Fonte: The author (2026).

Box 5.1 – Summary of machine learning models and search space hyperparameter ranges.

Model	Description	Key Parameters
Random Forest	Bagging ensemble of decision trees; robust to nonlinearity and interactions (BREIMAN, 2001).	Trees: 200–250; depth: 12–15; min samples leaf: 2–4; min split: 5–10. criterion: $squared\_error$, $absolute\_error$, $friedman\_mse$, poisson.
XGBoost	Regularized gradient boosting with optimized tree construction (CHEN; GUESTRIN, 2016).	Trees: 200–250; depth: 6–8; learning rate: 0.01–0.1; subsample: 0.8–0.9; colsample: 0.7–1.0.
CatBoost	Boosting optimized for categorical features using ordered target statistics (PROKHORENKOVA et al., 2018).	Iterations: 400–600; depth: 6–8; learning rate: 0.01–0.1; subsample: 0.8–0.95; L2 leaf reg.: 1–10.
LightGBM	Histogram-based gradient boosting with leaf-wise growth; allows monotonic constraints (KE et al., 2017).	Trees: 200–250; depth: 8–12; leaves: 31–50; learning rate: 0.05–0.1; L1/L2: 0.0–1.0; monotonic constraint: $P_e \uparrow \Rightarrow C \uparrow$.

Fonte: The author (2026).

strictly congruent with the fundamental principles of canopy saturation (Junqueira Junior et al., 2019; MELLO et al., 2024).

5.2.7 Validation Strategy

Model performance was evaluated exclusively under out-of-sample validation schemes (MAHMOOD; KHAN, 2009). Three complementary strategies were implemented. First, forward-chaining temporal cross-validation (TSCV) (ARSENAULT et al., 2023; ROBERTS et al., 2017) was applied using a rolling training window, preserving the chronological order of events and preventing information leakage across time. Second, spatial robustness was assessed using a Leave-One-Point-Out (LOPO) approach (SCHMIDT et al., 2020), in which all observations from one monitoring point were withheld for testing while the model was trained on the remaining points. Third, predictive stability was examined through rolling multi-year windows of fixed length (five years), allowing the assessment of model consistency under changing climatic conditions.

The key distinction between TSCV and the third scheme, rolling 5-year windows (ROLL5Y), is how past data contribute to training. TSCV accumulates all available history in each fold (expanding window), whereas ROLL5Y uses a fixed-length window of exactly five years that advances one year at a time, discarding observations older than five years. This design isolates the model's skill under recent climatic conditions and explicitly tests temporal stability across the drought–wetting cycles of the

2012–2023 period. Starting from 2017 (trained on 2012–2016), the window advances annually, yielding seven train–test splits.

Moreover, the proposed feature-engineering physically enhanced framework was assessed for physical consistency by means of the Shapley Additive Explanations (SHAP) method (LUNDBERG; LEE, 2017). This analysis allows verifying whether the learned relationships are aligned with the expected physical behavior of the process, ensuring that the model responses are not only accurate but also physically plausible (SCHMIDT et al., 2020; KING et al., 2022; WILLARD et al., 2022).

5.2.8 Performance Evaluation

Model performance was evaluated using the coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(C_i - \hat{C}_i)^2}{\sum_{i=1}^{n}(C_i - \bar{C})^2} \tag{5.4}$$

the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(C_i - \hat{C}_i)^2} \qquad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|C_i - \hat{C}_i\right| \tag{5.5}$$

and the Kling–Gupta Efficiency (KGE) (GUPTA et al., 2009), integrating correlation, bias, and variability components:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \tag{5.6}$$

where r is the linear correlation coefficient, β is the bias ratio ($\mu_{\hat{C}}/\mu_C$), and γ is the variability ratio ($\sigma_{\hat{C}}/\sigma_C$).

Residual diagnostics included normality (Shapiro–Wilk test), autocorrelation (Durbin–Watson statistic), and heteroscedasticity assessment through residual–fitted value relationships.

5.3 Results

5.3.1 Dataset Quality Control and Outlier Detection

A total of 23,808 initial event–point observations were derived from the forest monitoring network across 32 plots over a 10.4-year period. After applying the multi-stage quality control protocol, 21,057 valid observations were retained, representing a total retention rate of 88.4%. The filtering process primarily addressing physical inconsistencies (2,372 observations removed for exceeding physical bounds, $C_{Peff} > 1.5$) was followed by the identification of 475 statistical outliers via the Isolation Forest algorithm.

Box 5.2 – Optimized hyperparameters for ensemble regressors following Bayesian optimization with temporal cross-validation.

Model	Optimized Hyperparameters
Random Forest (RF)	n_estimators = 400, max_depth = 8, min_samples_split = 7, min_samples_leaf = 5
XGBoost (XGB)	n_estimators = 250, max_depth = 10, learning_rate = 0.0137, subsample = 0.883, colsample_bytree = 0.633
LightGBM (LGBM)	n_estimators = 250, max_depth = 13, learning_rate = 0.0201, num_leaves = 24, subsample = 0.751
CatBoost (CAT)	iterations = 600, depth = 7, learning_rate = 0.05

Fonte: The author (2026).

The refined dataset exhibited a mean effective precipitation coefficient ($C = P_{\text{eff}}/P_e$) of 0.617 and a standard deviation of 0.340.

The Isolation Forest algorithm, configured with a 2% contamination threshold, was applied separately for each month to capture phenological and climatic seasonality without imposing global bias. Despite the presence of high-intensity events, the extreme event correction (threshold $>$ 150 mm) did not find events requiring splitting in this specific subset, ensuring that the observed variability reflects natural canopy–rainfall interactions.

5.3.2 Hyperparameter Optimization and Model Configuration

Hyperparameter optimization was performed using a Bayesian optimization framework implemented via the *Optuna* library, coupled with a temporal cross-validation strategy based on *TimeSeriesSplit*. For each ensemble regressor, candidate hyperparameter sets were evaluated using three temporal folds, and model selection was guided by joint assessment of R^2 and RMSE. Table 5.2 summarizes the optimized configurations.

The Bayesian optimizer converged to low learning rates for both gradient boosting models, XGBoost (0.0137) and LightGBM (0.0201), at the lower boundary of the search space (0.01–0.20, log scale), indicating that slower, more conservative gradient updates were preferred given the high dimensionality and spatial heterogeneity of the dataset. XGBoost selected the maximum allowed depth (10) and a low colsample_bytree (0.633), consistent with a model that benefits from deep trees operating on subsets of features to avoid overfitting to spatially redundant one-hot encoded point indicators. LightGBM selected a lower num_leaves (24 out of 20–100), which together with depth = 13 produces relatively narrow trees, aligning with the leaf-wise growth strategy of this algorithm. Random Forest converged to the maximum n_estimators (400) and minimum depth (8) with the most conservative leaf

Table 5.3 – Predictive performance metrics for the effective precipitation coefficient (C) across validation schemes (TSCV: Time Series Cross-Validation; LOPO: Leave-One-Point-Out; ROLL5Y: Rolling 5-Year).

Model	**Scheme**	R^2	*RMSE*	*MAE*	*KGE*
RF	TSCV	-0.315	0.389	0.281	0.349
	LOPO	0.641	0.203	0.151	0.698
	ROLL5Y	0.030	0.284	0.221	0.253
XGB	TSCV	-0.264	0.382	0.272	0.365
	LOPO	0.652	0.200	0.150	0.696
	ROLL5Y	0.118	0.271	0.209	0.302
LGBM	TSCV	-0.281	0.384	0.274	0.364
	LOPO	0.583	0.219	0.170	0.601
	ROLL5Y	0.104	0.273	0.210	0.285
CAT	TSCV	-0.226	0.376	0.268	0.354
	LOPO	0.438	0.254	0.201	0.474
	ROLL5Y	0.175	0.262	0.202	0.255

Fonte: The author (2026).

constraints (min_samples_leaf = 5), reflecting the bagging strategy's preference for more trees with greater regularization.

5.3.3 Model Performance and Validation Robustness

The predictive performance of the four ensemble regressors was assessed using R^2, RMSE, MAE, and KGE under three independent validation frameworks: TSCV, LOPO, and rolling 5-year temporal validation (ROLL5Y).

Comparative Algorithm Performance Across Validation Schemes

Table 5.3 summarizes the predictive performance across all validation strategies. Under TSCV, all models exhibited negative R^2 values and relatively low KGE scores (< 0.40), indicating limited temporal predictive skill when extrapolating across time. This behavior suggests that the interception process is strongly influenced by non-stationary drivers and event sequencing effects not fully captured under strict temporal splitting.

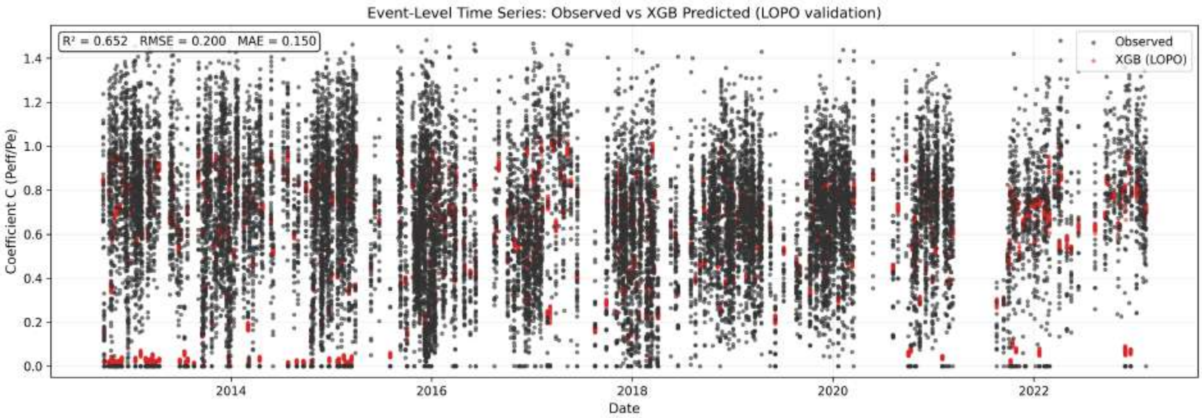
Under LOPO, XGBoost achieved the highest overall performance ($R^2 = 0.652$, RMSE $= 0.200$, MAE $= 0.150$, KGE $= 0.696$), followed by Random Forest and LightGBM. These results demonstrate strong spatial generalization, indicating that the engineered features effectively encode site-invariant physical controls. The ROLL5Y validation yielded marginally positive predictive skill, with CatBoost achieving the best performance ($R^2 = 0.175$). Notably, the performance gap between CatBoost ($R^2 =$

0.175, KGE = 0.255) and XGBoost ($R^2 = 0.118$, KGE = 0.302) under ROLL5Y suggests that algorithm sensitivity to non-stationary climatic regimes differs: CatBoost's ordered boosting may better handle distributional shifts across drought–wetting cycles. This finding underscores the value of multi-model ensembles when projecting interception under changing climate.

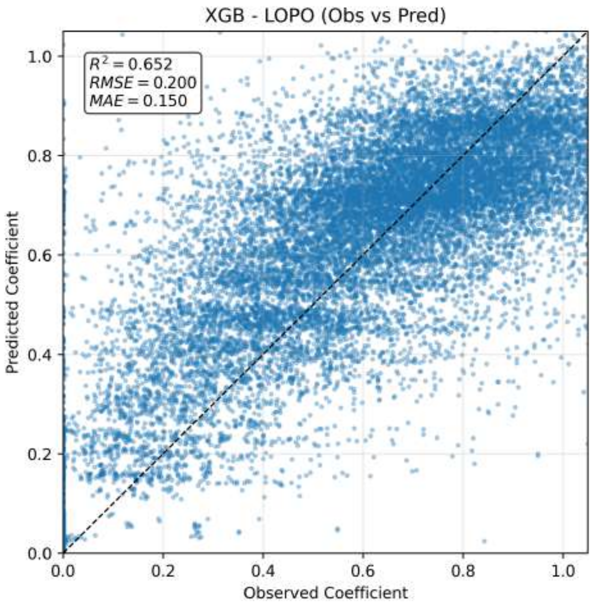
Climatological Consistency under Event-Scale Non-Stationarity

To illustrate event-scale model behavior directly, Figure 5.3 presents the event-level time series and 1:1 scatter for XGBoost under LOPO validation. At the individual event scale, the model achieves $R^2 = 0.652$ and RMSE = 0.200, with notable scatter concentrated at the extremes: events with $C_{Peff} < 0.30$ (high interception, typically small events on dry canopy in July–August) and events with $C_{Peff} > 1.0$ (localized rainfall concentration) show the largest deviations. The scatter is expected given the inherent stochasticity of individual interception events; the model's value lies in correctly reproducing the seasonal envelope and central tendency of partitioning rather than in perfect event-by-event prediction.

Figure 5.3 – Event-level performance of XGBoost under LOPO validation. (a) Time series of observed and predicted C_{Peff} across the full 10.4-year record. (b) 1:1 density scatter; $R^2 = 0.652$, RMSE $= 0.200$, MAE $= 0.150$.



(a) Time series of observed and predicted C_{Peff} across the full 10.4-year record.

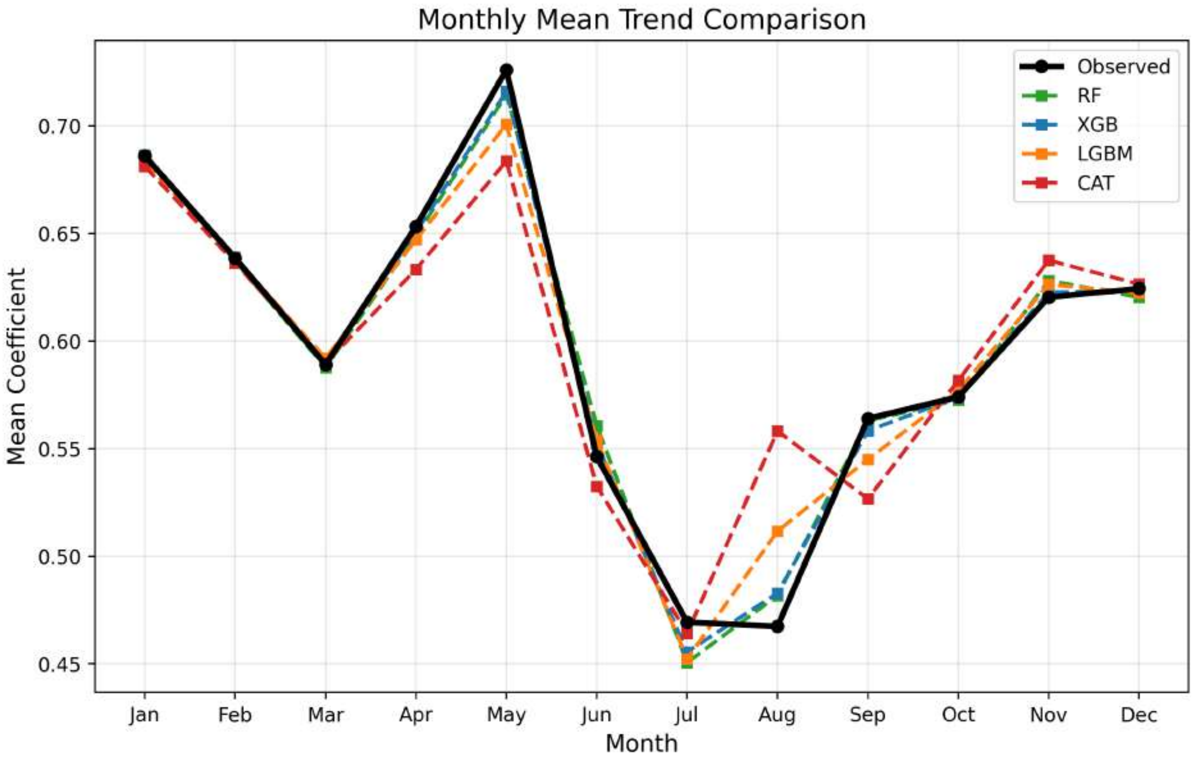


(b) 1:1 density scatter; $R^2 = 0.652$, RMSE $= 0.200$, MAE $= 0.150$.

Fonte: The author (2026).

Despite negative R^2 values under TSCV, the ensemble models exhibit strong seasonal fidelity. Comparison between observed monthly mean effective precipitation coefficients and model predictions reveals high temporal agreement (Figure 5.4). The models successfully internalized the system climatology, capturing the transition between the dry season (high interception losses) and the rainy season (increased effective precipitation). The models remain robust for long-term water balance estimation and seasonal ecohydrological analyses, despite limited skill at the individual event scale.

Figure 5.4 – Comparison between observed and modeled monthly mean effective precipitation coefficients (C). The ensemble models reproduce the seasonal cycle with high fidelity, capturing both wet-season maxima and dry-season minima, despite reduced event-scale performance under strict temporal validation.



Fonte: The author (2026).

Seasonal Performance and Distributional Consistency

All comparisons reported in this section operate at the individual rainfall event scale, the native resolution of the model, stratified by season for visualization only (Figure 5.5).

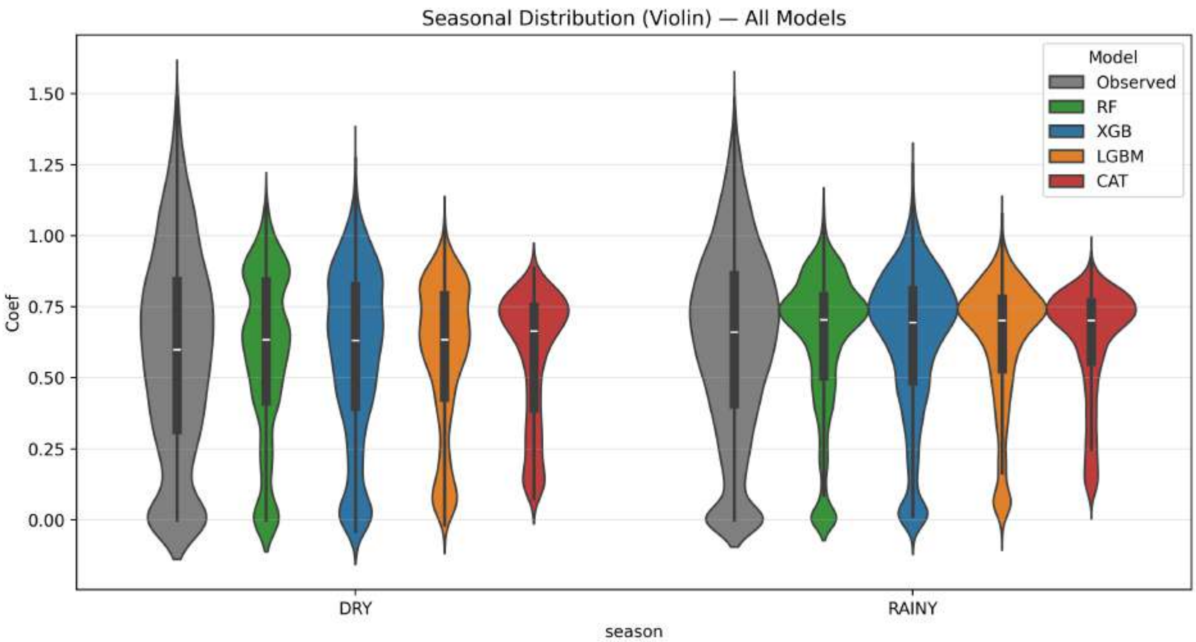
Seasonal stratification revealed a distinct performance gradient across all ensemble models. XGBoost achieved $R^2 = 0.918$ and KGE = 0.879 during the dry season versus $R^2 = 0.816$ and KGE = 0.779 during the rainy season under LOPO validation. Random Forest and LightGBM followed the same gradient (RF: dry $R^2 = 0.748$, KGE = 0.797; rainy $R^2 = 0.666$, KGE = 0.707). The superior dry-season skill likely reflects not only the dominance of clearer physical controls under moisture-limited conditions, but also the lower variance of events during this period, a factor that can inherently facilitate prediction.

The distributional comparison between observed and predicted C_{Peff} reveals consistent patterns and systematic differences across all models. During the rainy season (17,146 events), the observed distribution presents a mean of 0.625, median of 0.661, and an IQR of 0.452 (Q25 = 0.408; Q75 = 0.860), reflecting the wide variability in canopy saturation states. XGBoost reproduces the mean with near-zero bias (predicted mean $= 0.616$; bias $= -0.009$) and median adequately (0.695), but compresses the spread substantially: IQR $= 0.320$ (Q25 $= 0.481$; Q75 $= 0.800$), representing a compression of 0.132 relative to the observed IQR. RF, LightGBM, and CatBoost show even stronger compression (IQR $= 0.254$, 0.266, and 0.219, respectively), indicating that ensemble averaging systematically smooths the tails of the distribution, underestimating both the high-interception events ($C < 0.3$) and the near-complete throughfall events ($C > 0.90$).

During the dry season (3,911 events), the observed IQR is wider (0.522; Q25 $= 0.318$; Q75 $=$ 0.840) than during the rainy season (0.452), reflecting the high variability from sporadic events on a drying canopy. All models substantially overestimate the lower quartile (XGB Q25 $= 0.477$ vs. observed 0.318; difference $= 0.159$), indicating that the extreme interception efficiency of early post-dry-spell events is not fully captured despite the inclusion of API and dry-day count features. The observed median (0.600) is lower than the predicted median across all models (XGB: 0.693; RF: 0.695; LGBM: 0.693; CAT: 0.706), confirming a systematic positive bias during the dry season (XGB bias $= +0.030$).

Taken together, the violin plots in Figure 5.5 illustrate these findings visually: the rainy season distributions are more elongated in the observations (IQR $= 0.452$) than in any model (max IQR $= 0.320$ for XGB), while the dry season distributions are more compact in the models (XGB IQR $= 0.319$) than in the data (IQR $= 0.522$). The models successfully reproduce the seasonal shift in central tendency but consistently underestimate the distributional spread, particularly during the dry season.

Figure 5.5 – Seasonal distribution of observed and predicted effective precipitation coefficients for dry and rainy seasons across all ensemble models.



Fonte: The author (2026).

5.3.4 Spatiotemporal Variability of Interception and Effective Precipitation

Seasonal Dynamics and Critical vs. Stable Points

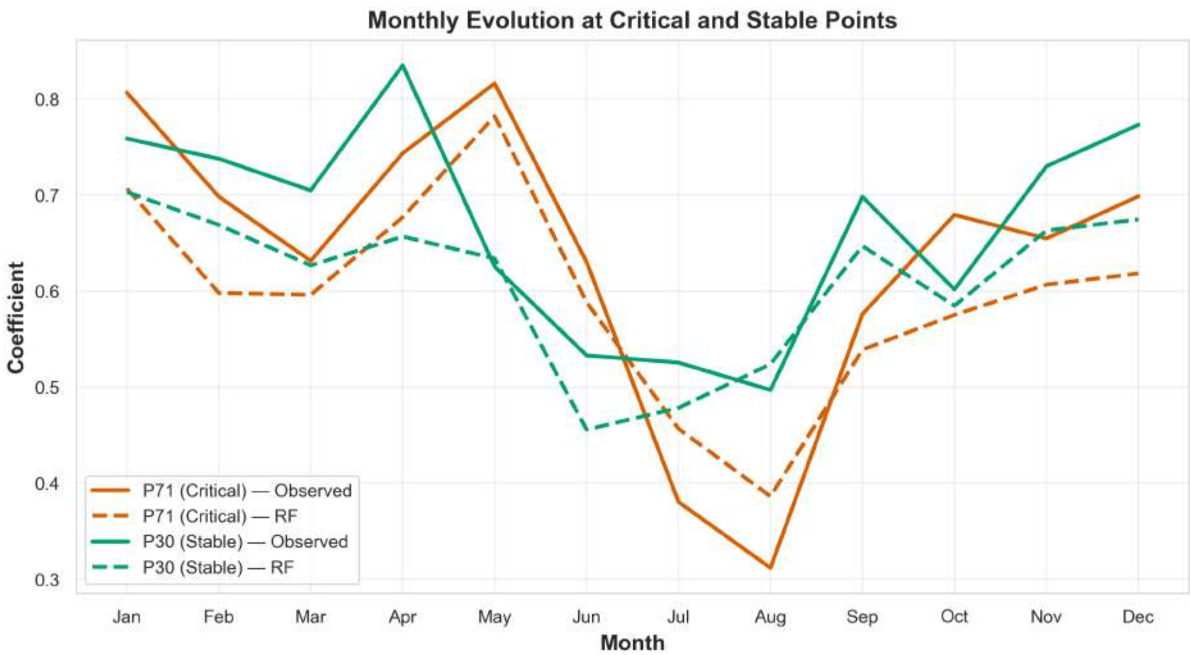
Seasonal analysis revealed a pronounced oscillation in interception efficiency at both monitoring points, governed by the phenological state of the canopy and rainfall intensity (Figure 5.6).

The Critical Point (P120) is characterized by systematically lower C_{Peff} values and higher interception losses throughout the year, ranging from a mean of 0.315 ± 0.254 in June to 0.536 ± 0.290 in January, an intra-annual amplitude of 0.221. Its overall distribution is wide and skewed toward low values ($Q25 = 0.186$; $Q75 = 0.732$; $IQR = 0.546$), indicating pronounced sensitivity to event-scale variability. In contrast, the Stable Point (P30) maintains consistently higher C_{Peff} values (overall mean $= 0.718$ vs. 0.479 at P120; median $= 0.768$ vs. 0.498) with a more compact distribution ($Q25 = 0.573$; $Q75 = 0.900$; $IQR = 0.327$), reflecting reduced sensitivity to antecedent moisture conditions.

Figure 5.6 presents the monthly mean $\pm$ standard deviation for both points alongside the full distribution of all 32 monitoring locations, allowing the structural contrast between P120 and P30 to be placed in the spatial context of the network. The difference between the two points is most pronounced during June–April, reaching a maximum in April ($\bar{C} = 0.835$ at P30 vs. 0.445 at P120; $\Delta = +0.390$) and nearly disappearing in May ($\Delta = -0.007$), which coincides with the end of the wet season and a transitional phenological state of the semi-deciduous canopy. This seasonal convergence in May is more likely explained by structural canopy changes, including leaf area reduction at the onset of the dry season, than by a simple event-by-event storage saturation dynamic, since a single moderate event is sufficient to saturate canopy storage capacity.

This distinction is most pronounced during small-to-moderate rainfall events ($P_e \leq 15$ mm): the interception fraction at P120 is on average 60.8% higher than at P30 for events in this size range (P120: mean $C_{Peff} = 0.345$, interception $= 65.5\%$; P30: mean $C_{Peff} = 0.593$, interception $= 40.7\%$). The structural characteristics of P120 that may explain this persistent pattern, such as denser canopy closure, higher LAI, or different drainage architecture, are discussed in Section 5.4.

Figure 5.6 – Monthly distribution of the effective precipitation coefficient ($C = P_{\text{eff}}/P_e$) at the Critical (P120) and Stable (P30) monitoring points. **(a)** Monthly mean ± standard deviation of observed (solid lines) and XGBoost-predicted (dashed lines) C_{Peff} at P120 (red) and P30 (blue); shaded bands represent ±1 SD. **(b)** Monthly boxplot distribution of observed C_{Peff} across all 32 monitoring points (grey), with the Critical point (P120, red diamonds) and Stable point (P30, blue squares) overlaid as mean ± SD.



Fonte: The author (2026).

Spatial Heterogeneity and Antecedent Moisture Scenarios

Spatial mapping of the mean effective precipitation coefficient reveals a complex mosaic of rainfall partitioning across the study area (Figure 5.7). To isolate the effect of antecedent moisture from seasonal and event-size confounding, three canopy wetness scenarios were simulated under controlled conditions: January (rainy season), $P_e = 20$ mm, varying only antecedent precipitation indices (API_{7d}) while keeping all other features fixed. XGBoost was used exclusively for this analysis as it achieved the highest spatial generalization performance under LOPO validation ($R^2 = 0.652$, RMSE $= 0.200$, KGE $= 0.696$; Table 5.3).

The three scenarios represent:

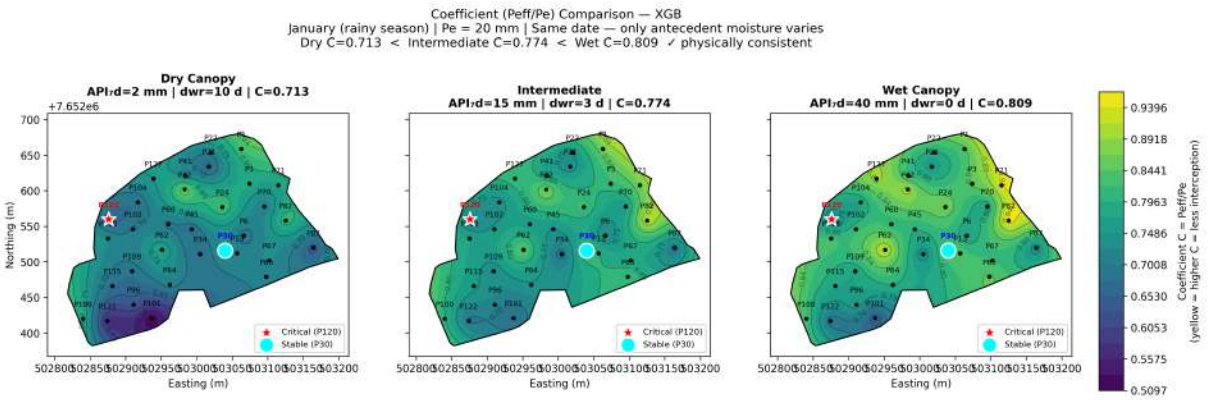
- **Dry Canopy**: $API_{7d} = 2$ mm, 10 dry days, minimum antecedent wetness, maximum available canopy storage.
- **Intermediate**: $API_{7d} = 15$ mm, 3 dry days, partial antecedent saturation.

- **Wet Canopy**: $API_{7d} = 40$ mm, 0 dry days, maximum antecedent wetness, reduced available canopy storage.

Simulation results (Figure 5.7) demonstrate that antecedent moisture exerts a strong spatial control on rainfall partitioning. Under Wet Canopy conditions, the mean C_{Peff} across all 32 monitoring points reaches 0.809 ($P_{eff} = 16.2$ mm), compared to 0.713 under Dry Canopy conditions ($P_{eff} = 14.3$ mm), a difference of $+1.9$ mm ($+13.5\%$) in effective precipitation for the same gross event. The spatial heterogeneity within each scenario is also pronounced: under Dry Canopy, C_{Peff} ranges from 0.510 to 0.874 across monitoring points (range $= 0.364$), while under Wet Canopy the range narrows to 0.330 (0.633–0.963), indicating that high antecedent wetness partially homogenises the spatial response by pre-saturating canopy storage at most locations.

These spatial gradients are consistent with the structural heterogeneity of the monitoring network: locations with denser multi-strata canopy and higher leaf area systematically present lower C_{Peff} values regardless of antecedent moisture scenario, reinforcing the designation of P120 as a persistent interception hotspot. The physical mechanisms linking antecedent moisture conditions, canopy storage dynamics, and spatial partitioning patterns are discussed in Section 5.4.

Figure 5.7 – Spatial distribution of the mean effective precipitation coefficient ($C = P_{\text{eff}}/P_e$) predicted by XGBoost under dry ($API_{7d} = 2$ mm), intermediate ($API_{7d} = 15$ mm), and wet ($API_{7d} = 40$ mm) canopy conditions. Simulations were performed under controlled conditions (January, $P_e = 20$ mm) with antecedent moisture as the only varying factor. Warmer colors (yellow) indicate higher C_{Peff} (lower interception fraction); cooler colors (purple) indicate lower C_{Peff} (higher interception fraction).



Fonte: The author (2026).

Integrated Interpretation. Taken together, the seasonal trajectories and the canopy-state-dependent spatial maps demonstrate that interception dynamics in the study area are governed by a hierarchical interaction between temporal forcing (seasonality and antecedent wetness) and spatial structure (canopy heterogeneity). The XGBoost model successfully captures these multiscale controls, reproducing both localized extremes and broader spatial–temporal gradients.

5.3.5 Physical Drivers and Model Interpretability (SHAP Analysis)

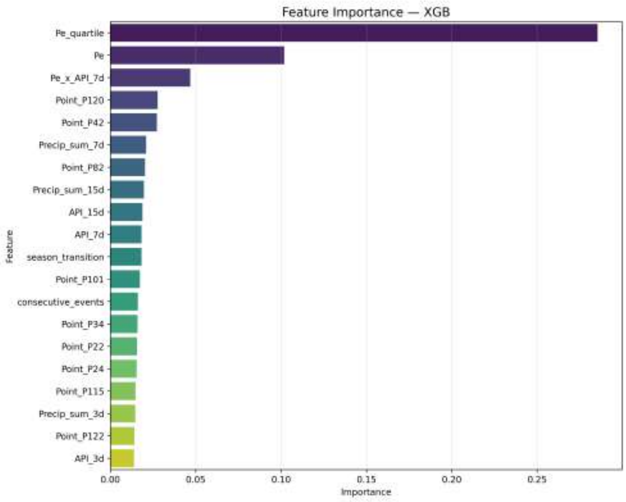
To elucidate the physical controls governing C_{Peff} and the soundness of the physical enhanced approach, model interpretability was assessed using SHapley Additive exPlanations (SHAP) in combination with traditional feature-importance rankings derived from the XGBoost model (Figure 5.8).

Both SHAP values and feature-importance rankings consistently identify event rainfall depth (P_e) and its categorical representation ($Pe\_quartile$) as the dominant features of C_{Peff}. Indicators of antecedent moisture—particularly the interaction term $P_e \times API_{7d}$ (Importance $= 0.041$) and accumulated precipitation over 7 and 15 days—rank immediately below, highlighting the central role of hydrological memory. Seasonal variables (doy_sin, doy_cos) exhibit moderate contributions (mean SHAP ≈ 0.013).

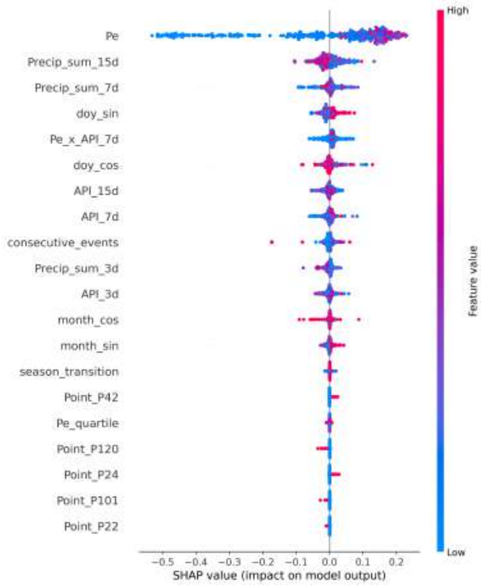
The SHAP summary plot reveals pronounced non-linear behavior characterized by physically meaningful thresholds. Low values of P_e produce negative SHAP contributions, corresponding to reduced C_{Peff} during small events where rainfall is largely retained by the canopy. Conversely, high P_e values generate strong positive SHAP contributions once canopy storage capacity is exceeded. For a given P_e, higher API_{7d} values shift SHAP contributions toward larger C_{Peff}, demonstrating that prior wetness preconditions the efficiency with which rainfall reaches the forest floor.

While traditional feature importance ranks $Pe\_quartile$ as the most influential variable for structuring tree splits (Importance $= 0.286$), SHAP analysis indicates that the continuous rainfall magnitude P_e exerts the largest marginal impact (mean SHAP $= 0.094$) on individual predictions. This contrast highlights that categorical rainfall intensity helps organize the model hierarchy, whereas continuous rainfall magnitude and antecedent moisture dominate the physical response at the event scale.

Figure 5.8 – Comparison between global feature-importance rankings and SHAP-based interpretability for the XGBoost model. (a) Feature importance ranking from the XGBoost model. (b) SHAP summary plot showing the distribution of feature contributions.



(a) Feature importance ranking from the XGBoost model.



(b) SHAP summary plot showing the distribution of feature contributions.

Fonte: The author (2026).

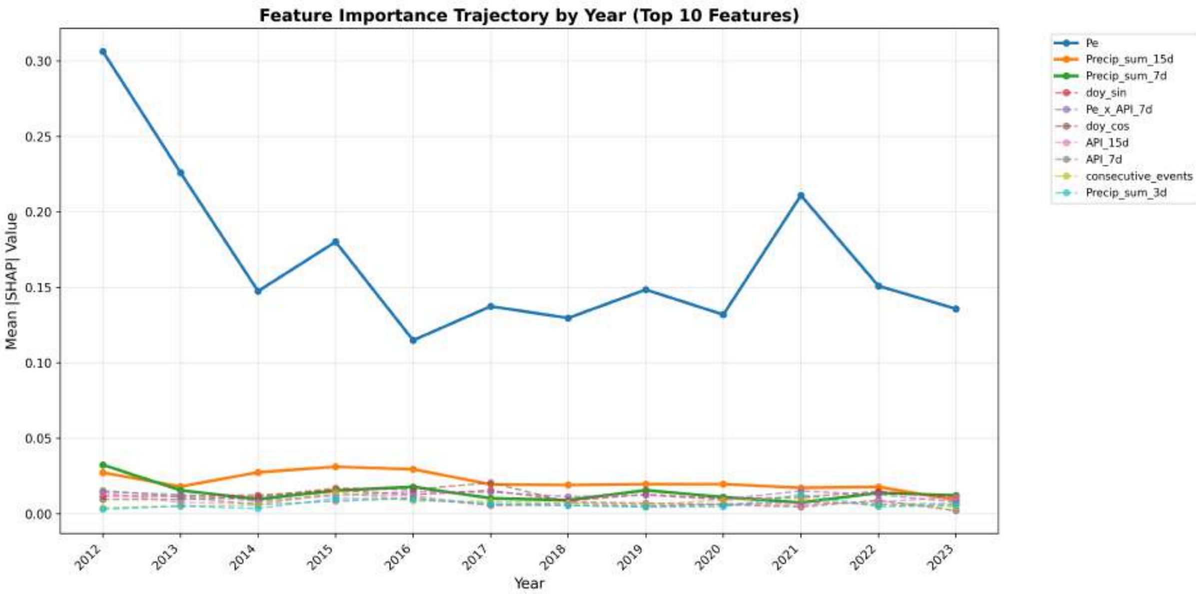
Residual diagnostics revealed significant deviations from normality ($p < 0.05$, Shapiro–Wilk), primarily driven by the inherent skewness of extreme rainfall events and the non-linear canopy response. The Durbin–Watson statistic revealed positive residual autocorrelation (DW < 2), confirming a strong temporal dependency that aligns with the system's hydrological memory. The relationship between residuals and fitted values exhibited a heteroscedastic pattern, where predictive uncertainty increased with the magnitude of C_{Peff}, physically consistent with increased variability in canopy drainage during large-magnitude storms.

5.3.6 Feature Importance and Spatiotemporal Dynamics

Yearly Feature Importance Trajectories

Figure 5.9 presents the yearly evolution of SHAP-based feature importance for XGBoost across 2012–2018 (the period with complete SHAP decomposition). Given that XGBoost achieved the highest spatial generalization under LOPO validation ($R^2 = 0.652$; KGE $= 0.696$), the following analysis focuses on its feature hierarchy; corresponding trajectories for RF, LightGBM, and CatBoost are provided as supplementary material.

Figure 5.9 – Yearly trajectory of XGBoost SHAP-based feature importance across 2012–2018. Each year shows the mean absolute SHAP values for the top features, ranked by overall importance. The shaded band marks the 2015–2017 drought period. Note the progressive decline in P_e importance during the drought and the relative increase of phenological and interaction features.



Fonte: The author (2026).

Gross precipitation (P_e) consistently ranked as the most influential predictor, contributing 66.1% of the total mean |SHAP| signal in 2012, reflecting the primacy of event magnitude under regular rainfall regimes. The interaction term $P_e \times API_{7d}$ (mean $|\text{SHAP}| = 0.012$ pre-drought) and 15-day accumulated precipitation ($Precip\_sum_{15d}$; mean $= 0.024$ pre-drought) ranked consistently in the top predictors, confirming the central role of hydrological memory.

The most striking pattern is the progressive decline in P_e importance from 0.306 in 2012 to 0.115 in 2016, a reduction of 62.4% at the drought peak, followed by a partial recovery to 0.138 in 2017. This decline indicates that under prolonged moisture deficit, the magnitude of individual rainfall events became a substantially weaker predictor of interception efficiency: the model relied increasingly on accumulated moisture conditions and phenological timing to explain C variability. The relative importance of the interaction term $P_e \times API_{7d}$ simultaneously increased from a pre-drought mean of 0.012 to 0.014 during the drought (+15.0%), indicating that the coupling between event size and antecedent wetness became more critical when both factors were simultaneously suppressed. Seasonal phenological indicators (*doy_sin*, *doy_cos*) also showed increased importance during the drought (pre: 0.012 and 0.010; drought: 0.015 and 0.018), consistent with greater structural canopy variability driven by drought-induced leaf shedding and altered phenological timing (RODRIGUES et al., 2021a).

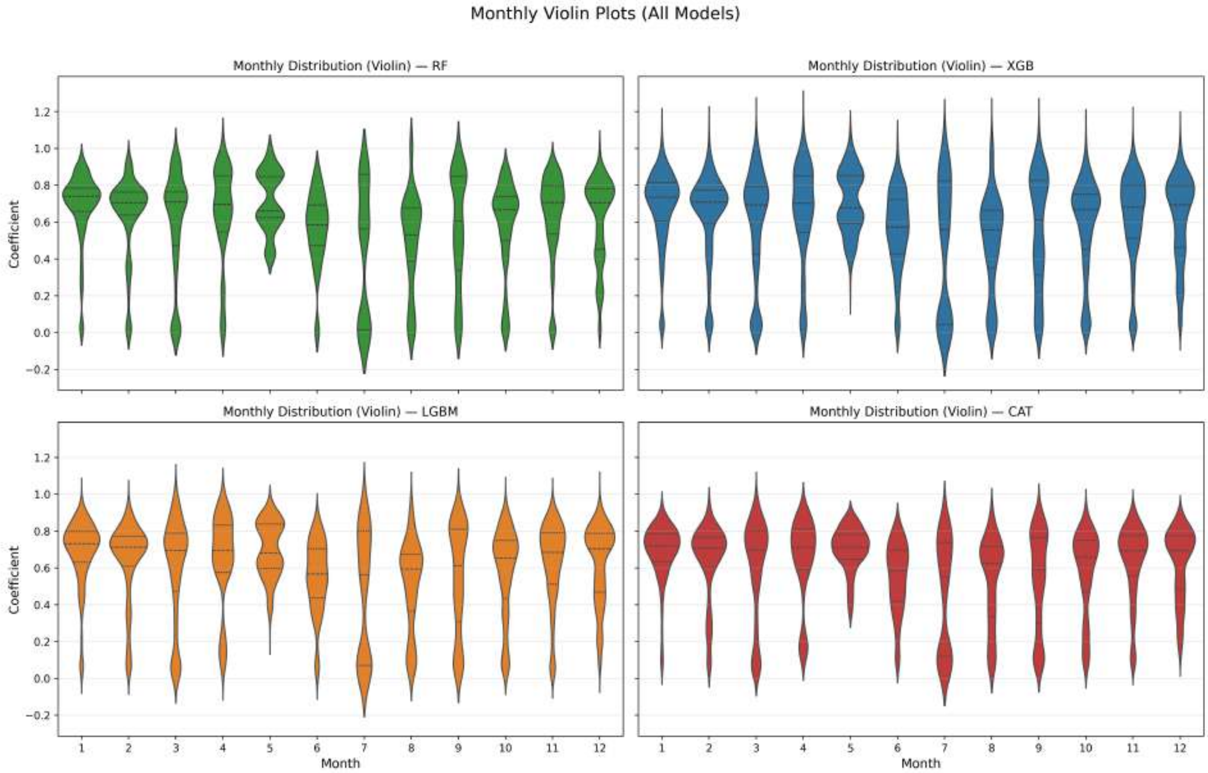
In 2018, the only fully documented post-drought year in the SHAP table, P_e importance (0.130) remained 42.7% below its 2012 peak, suggesting a persistent alteration of the rainfall-partitioning regime rather than rapid recovery to pre-drought conditions. The interaction term $P_e \times API_{7d}$ also partially declined (0.011), while $Precip\_sum_{15d}$ decreased to 0.019, below its pre-drought level (0.024), indicating that longer-term moisture accumulation became less influential as the soil and canopy system partially recovered.

Among the 32 spatial indicators, $Point\_P42$ showed the most consistent importance throughout the record, with mean |SHAP| increasing from 0.00100 (pre-drought) to 0.00119 (drought), the highest of any spatial point during the 2015–2017 period. $Point\_P120$ (the Critical Point) followed a similar pattern (pre: 0.00074; drought: 0.00096), suggesting that structurally heterogeneous locations became relatively more distinguishable when large-scale meteorological forcing was suppressed. In 2018, $Point\_P120$ declined sharply (0.00051), while P42 remained elevated (0.00099), pointing to differential structural recovery rates across canopy positions. These spatial contributions, though modest in absolute terms relative to P_e and API features, reflect a genuine signal: during the drought, canopy structural attributes, captured implicitly through the spatial point encoding, partially compensated for the reduced predictive power of rainfall magnitude.

Monthly Distribution and Bimodal Regimes

Figure 5.10 displays the monthly distribution of C as violin plots for all four models, aggregated across all 32 monitoring points and the full 10.4-year record. All analyses operate at the individual rainfall event scale.

Figure 5.10 – Monthly distribution of modeled effective precipitation coefficient ($C = P_{\text{eff}}/P_e$) for RF, XGBoost, LightGBM, and CatBoost, aggregated across all 32 monitoring points and the full 10.4-year period. Note: (i) bimodal distribution in May reflecting transition between wet and dry seasons; (ii) tighter, lower distributions in July–August reflecting dominance of small, easily-intercepted events; (iii) downward tails in July from extremely small events after prolonged droughts; and (iv) consistent patterns across all four model architectures.



Fonte: The author (2026).

January through April (rainy season peak) present the highest observed mean C values of the year. January records mean = 0.686, median = 0.710, and IQR = 0.412 (Q25 = 0.491; Q75 = 0.903). February presents mean = 0.639, median = 0.680, IQR = 0.414. April shows mean = 0.653, median = 0.688, IQR = 0.485. March presents the widest IQR of the rainy season (0.539; Q25 = 0.317; Q75 = 0.856), reflecting high within-month variability. All models reproduce the central tendency of these months with small bias (XGB January: mean = 0.618, median = 0.694, IQR = 0.307) but consistently compress the distributional spread; in January, XGB IQR (0.307) is 25% narrower than the observed (0.412), a pattern shared by RF (0.287) and LGBM (0.238), indicating that ensemble averaging systematically smooths the extremes of the distribution.

May is the most structurally complex month in the dataset. The observed distribution presents mean = 0.726, median = 0.710, and a wide IQR of 0.416 (Q25 = 0.509; Q75 = 0.925), with a pronounced bimodal structure: a lower-C cluster concentrated near Q25 ≈ 0.51 representing small, highly intercepted events typical of the retreating wet season, and an upper-C cluster near Q75 ≈ 0.93 from larger canopy-saturating events. The resulting observed mean (0.726) is the highest of any month. All four models reproduce the elevated May mean adequately (XGB: 0.603; RF: 0.604; LGBM: 0.599), but compress the IQR substantially (XGB: 0.341; RF: 0.327; LGBM: 0.276 vs. observed 0.416), failing to reproduce the high-C tail (Q75 observed = 0.925 vs. XGB = 0.802) and underestimating the frequency of high-interception events near the lower mode.

July exhibits the most extreme bimodal structure of any month. The observed IQR reaches 0.838, the widest of the entire record, with Q25 = 0.000 and Q75 = 0.838. The observed mean (0.469) and median (0.469) are the lowest of any month. Two distinct populations coexist: a lower mode at $C \approx 0$ (at least 25% of events produce zero effective precipitation at the ground), representing completely intercepted sub-saturation events after prolonged dry spells at the peak of the dry season; and an upper mode near $C \approx 0.70$–0.84 from larger episodic events that exceed canopy storage capacity. This bimodality is more extreme than May's: while May's Q25 is 0.509, July's Q25 = 0.000, indicating that a quarter of July events are entirely intercepted. All models fail to reproduce this structure: XGB predicts IQR = 0.344 (Q25 = 0.470; Q75 = 0.814; mean = 0.618), compressing the observed IQR by 0.494 units. RF (IQR = 0.327) and LGBM (IQR = 0.263) show similar limitations. The model mean overestimation in July is substantial: XGB mean = 0.618 vs. observed 0.469, a positive bias of +0.149.

August maintains a low-C regime: observed mean = 0.467, median = 0.480, IQR = 0.497 (Q25 = 0.195; Q75 = 0.692). Models again overestimate C (XGB mean = 0.601, bias = +0.133; RF = 0.605, bias = +0.137; LGBM = 0.598, bias = +0.130), indicating that the feature set does not fully capture the extreme interception efficiency of the peak dry season.

September through December show progressive recovery as the rainy season onset brings larger, more regular events: September (observed mean = 0.564; IQR = 0.613), October (mean = 0.574; IQR = 0.471), November (mean = 0.620; IQR = 0.451), December (mean = 0.625; IQR = 0.462). Models reproduce the seasonal recovery adequately in central tendency (XGB mean bias < 0.05 in all four months), with consistent IQR compression (XGB November IQR = 0.287 vs. observed 0.451).

The convergence of all four models around the same monthly central tendencies, despite their substantially different algorithmic architectures, demonstrates that the seasonal C structure (bimodality in May and July, low-C regime in August, recovery October–December) is an emergent property of the underlying data rather than a model-specific artifact. However, the systematic failure to reproduce the July IQR (observed 0.838 vs. XGB 0.344, RF 0.327, LGBM 0.263) identifies the extreme dry-season event regime as the primary frontier for future improvement of physically-enhanced ML frameworks applied to tropical interception modelling.

5.4 Discussion

5.4.1 Transferability versus Temporal Non-Stationarity in Rainfall Interception Modeling

The results reveal a fundamental duality in the predictive behavior of ML models applied to rainfall interception: strong spatial transferability coexisting with pronounced temporal non-stationarity at the event scale. Under LOPO validation, XGBoost achieved $R^2 = 0.652$ and KGE $= 0.696$, confirming successful internalization of site-invariant physical controls despite the pronounced structural heterogeneity of the 32-plot Atlantic Forest monitoring network.

This spatial transferability stems partly from the training strategy: because the model was jointly trained on observations spanning all 32 spatially heterogeneous plots, the full range of canopy structural behaviors is implicitly encoded in the C response variable. The model therefore learns the full spectrum of rainfall-partitioning behaviors from the data, allowing spatial generalization without requiring explicit canopy descriptors such as leaf area index or basal area. This represents a key advantage over process-based models such as Gash or Rutter (GASH; MORTON, 1978; MUZYLO et al., 2009), which require independent plot-by-plot calibration and are typically applied at the storm or daily scale, not at the individual event level as implemented here (NÁVAR, 2017; LINHOSS; SIEGERT, 2020). The LOPO framework provides a genuinely stringent test of this generalization: each iteration removes all events from one monitoring point and retrains the model on the remaining 31 points, a level of spatial independence that classical calibration-based approaches cannot replicate across a spatially heterogeneous network.

Interpreting temporal non-stationarity: the drought hypothesis. The negative R^2 values under TSCV (XGB: $R^2 = -0.264$; KGE $= 0.365$; RF: $R^2 = -0.315$; KGE $= 0.349$) cannot be attributed solely to unresolved intra-event processes. The LOPO validation succeeded under the same data and model configuration, demonstrating that the feature engineering is physically adequate when the spatial dimension is the source of generalization. The critical difference lies in the temporal dimension: TSCV trains on earlier periods and tests on later ones, forcing the model to extrapolate across the 2014–2017 multi-year drought documented at this site (RODRIGUES et al., 2021a; RODRIGUES et al., 2021b). This drought fundamentally altered the rainfall-interception system in multiple concurrent ways. First, the event size distribution shifted dramatically: mean P_e increased from 11.9 mm during the pre-drought period (2012–2014) to 18.3 mm during the drought (2015–2017) and 19.5 mm in 2018, as small-rainfall events became less frequent. Second, dry season representativeness collapsed: dry season events represented 25.5% of all events pre-drought but only 15.3% during the drought, meaning the model was trained partly on wet-season climatology and then tested on an anomalous drought-dominated regime. Third, and most importantly, the canopy structure itself changed: the drought induced rapid responses such as partial leaf shedding and reduced LAI, and delayed responses including individual tree mortality, species reorganization, and altered drainage architecture (MEIR et al., 2015; RODRIGUES et al., 2021a). The SHAP temporal analysis captures this reorganization: the importance of gross precipitation P_e declined from 0.306 in 2012 to 0.115 in 2016, a 62.4% reduction, while phenological indicators (*doy_sin*, *doy_cos*) increased in relative importance during 2015–2017, reflecting the growing role of canopy structural state over rainfall magnitude in controlling interception efficiency under drought. Critically, even in 2018, after rainfall partially recovered, P_e importance (0.130) remained 57.5% below its 2012 level, suggesting that the canopy had not structurally recovered, and the model trained on pre-drought conditions was being tested on a forest undergoing slow post-drought reorganization (GUAUQUE-MELLADO et al., 2026; MELLO et al., 2025).

The ROLL5Y validation, which uses a fixed five-year sliding window, yielded marginally positive but modest skill (XGB: $R^2 = 0.118$, KGE $= 0.302$), confirming that restricting training to the most recent five years partially mitigates, but does not eliminate, the non-stationarity problem. When the training and test windows both fall within the drought period (e.g., 2012–2016 train, 2017 test), performance improves; when the window spans the drought onset (e.g., 2010–2014 train, 2015 test), it degrades, consistent with the hypothesis that the primary source of temporal failure is the distributional shift induced by drought rather than generic temporal autocorrelation.

These findings align with the broader challenge of non-stationarity in data-driven hydrology: ML models trained under one climatic regime, even with physically meaningful features, face fundamental limitations when tested on regimes involving structural ecosystem changes that alter the underlying process generating the data (WILLARD et al., 2022; SCHMIDT et al., 2020). Physical enhancement through feature engineering (API, $P_e \times API_{7d}$, phenological markers) partially mitigates this problem by encoding process-relevant state variables, but cannot compensate for structural canopy changes that are neither monitored nor represented in any feature of the model.

5.4.2 Hydrological Memory and the Dynamic Nature of Canopy Storage Capacity

The SHAP-based interpretability analysis provides strong physical validation of the physically-enhanced framework, confirming that rainfall interception in the Atlantic Forest is a fundamentally state-dependent process governed by the interaction between event forcing and pre-existing canopy moisture conditions.

Classical interception theory (GASH; MORTON, 1978; MUZYLO et al., 2009) treats canopy storage capacity (S) as a quasi-static site parameter. At the present study site, estimated S values range from 0.82 mm (Liu model) to 1.45 mm (Gash model) (RODRIGUES et al., 2021a). However, these static values mask the fundamentally dynamic nature of canopy storage. The interaction term $P_e \times API_{7d}$

emerged as the third-ranked predictor in the SHAP hierarchy (mean |SHAP| $= 0.032$), confirming that the effective storage capacity of the canopy is not fixed; it varies event by event as a function of how much water remains from prior events. The smallest events in the dataset illustrate this most directly: for $P_e \leq 1.45$ mm ($n = 1{,}761$ events, 8.4% of all records), rainfall amounts below or equal to the maximum S estimated by the Gash model, the mean C_{Peff} is only 0.159, meaning 84% is intercepted on average. Under pre-wet conditions ($API_{7d} > 10$ mm), C_{Peff} rises to 0.200, while under pre-dry conditions ($API_{7d} \leq 2$ mm), C_{Peff} drops to 0.000, 100% interception. This is the carry-over effect Liu (1997) described: prior wetting reduces the available storage for the current event, allowing a measurable fraction of even sub-saturation events to reach the forest floor. This effect, first described by Liu (1997) and corroborated for Mediterranean forests by Limousin et al. (2008), is here demonstrated quantitatively for a neotropical semi-deciduous Atlantic Forest context.

The dominance of API_{7d} over shorter windows reflects the specific structural and physiological characteristics of this montane semi-deciduous Atlantic Forest (canopy height 25–30 m; LAI $= 4.2 \pm 0.6$; forest age 35–40 years; Junqueira Junior et al. (2017)). Schmidt et al. (2020) found that API windows of up to 7 days are required to capture hydrometeorological memory in humid forests. Furthermore, while short indices like a 2-day API are frequently used to categorize antecedent moisture conditions (SIDLE; ZIEGLER, 2017), our results suggest that such short windows are insufficient to fully capture the delayed drainage dynamics of forests with complex multi-strata canopy architecture. The physical explanation lies in the storage and drainage capacity of the canopy system: a taller, multi-strata forest with higher LAI has a larger effective storage volume across multiple canopy layers, bark, branches, mosses, epiphytes, which requires more time to fully drain between events. Holwerda et al. (2012) demonstrated that aerodynamic conductance and atmospheric vapour pressure deficit control the evaporative depletion rate of intercepted water; in this Highland Atlantic Forest (918–920 m a.s.l., humid climate), lower evaporation rates during wet periods mean the canopy retains moisture longer between consecutive events, extending the hydrological memory beyond the faster drying cycles typical of more arid or structurally simpler canopies (ZHANG et al., 2016; LUCAS-BORJA et al., 2023). Furthermore, the accumulation of litter, epiphytes, and mosses on bark surfaces, typical of humid neotropical secondary forests at this age, acts as additional slow-draining water storage (PYPKER et al., 2005), functionally increasing the time constant of the complex canopy reservoir (ZIMMERMANN; ZIMMERMANN, 2014). The 15-day accumulated precipitation ($Precip\_sum_{15d}$) also ranked among the top predictors (mean |SHAP| $= 0.018$), indicating that even slower-draining storage components, potentially including stemflow pathways along bark and epiphytic communities, contribute meaningfully to the memory signal.

The temporal SHAP analysis reveals an important modulation of this memory effect by the 2014–2017 multi-year drought documented at this site (RODRIGUES et al., 2021a). The relative importance of $P_e \times API_{7d}$ increased from 0.012 (pre-drought, 2012–2014) to 0.014 (drought, 2015–2017), while the primary driver P_e declined from 0.227 to 0.144. This reorganization indicates that under drought conditions, the magnitude of individual events became a weaker predictor of interception efficiency relative to the accumulated moisture deficit, consistent with the structural canopy changes (partial leaf shedding, reduced LAI, altered drainage pathways) documented for this remnant during the drought period (RODRIGUES et al., 2021a; GUAUQUE-MELLADO et al., 2026). The forest did not fully recover its pre-drought importance hierarchy in the post-drought period (2018–2023: P_e $|\text{SHAP}| = 0.130$), suggesting a persistent alteration of canopy structure and hydrological behavior extending years beyond the drought itself, a finding consistent with the delayed structural recovery reported for Atlantic Forest remnants following multi-year moisture deficits (MEIR et al., 2015).

The quantitative scenario simulations (Figure 5.7) provide the most direct evidence that the physically-enhanced feature set encodes real canopy storage dynamics rather than statistical artifacts. Under controlled conditions (January, $P_e = 20$ mm), transitioning from dry ($API_{7d} = 2$ mm) to wet ($API_{7d} = 40$ mm) antecedent conditions increases the mean predicted P_{eff} across all 32 monitoring points by 1.9 mm (+13.5%), with the spatial heterogeneity of this response (range 0.33–0.36 across monitoring points) reflecting the structural diversity of canopy positions within the forest. This quantitative link between antecedent indices and throughfall partitioning, reproduced consistently by the model across 10.4 years and 32 spatially heterogeneous monitoring locations, validates the physically-enhanced approach: the API and *Precip_sum* features function as proxies for the time-varying effective storage capacity that classical models treat as static, thereby enabling the ML framework to reproduce a process that Gash- and Rutter-type models can only approximate with event-by-event recalibration (MUZYLO et al., 2009; LINHOSS; SIEGERT, 2020).

5.4.3 Critical versus Stable Points: Ecohydrological Hotspots

By contrasting a Critical Point (P120) and a Stable Point (P30), differences of up to 60.8% in interception fraction were quantified for rainfall events below 15 mm (P120: mean $C_{Peff} = 0.345$, interception fraction = 65.5%; P30: mean $C_{Peff} = 0.593$, interception fraction = 40.7%). At the annual scale, the mean C_{Peff} at P120 (0.479; IQR = 0.546) is substantially lower and more variable than at P30 (0.718; IQR = 0.327), confirming that P120 operates as a persistent interception hotspot throughout the monitoring period. Critical points correspond to locations dominated by dense multi-strata canopy with large basal area, high aboveground biomass, and greater tree diameter—features associated with higher surface storage capacity (S) and lower direct throughfall (Junqueira Junior et al., 2019; RODRIGUES et al., 2021a). Stable points—exhibiting more linear relationships and lower seasonal amplitude—correspond to locations with greater canopy openness or younger recruitment cohorts (RODRIGUES et al., 2021a; GERMER; ELSENBEER; MORAES, 2006). For small events ($P_e < 15$ mm), the Critical Point canopy does not reach saturation and the effective precipitation fraction is disproportionately governed by free throughfall through canopy gaps and by the progressive formation of drip points as water is redistributed along branches (STAELENS et al., 2006; RODRIGUES et al., 2021a).

The seasonal convergence of C_{Peff} values between P120 and P30 in May ($\Delta = -0.007$; nearly identical means of 0.633 and 0.625, respectively) deserves particular attention. As noted by the reviewer, this convergence is unlikely to result from cumulative event-by-event canopy saturation across weeks; a single moderate event ($P_e > 5$ mm) is sufficient to saturate the available canopy storage ($S \approx 0.82$–1.45 mm; estimated by Rodrigues et al. (2021a) at the same site). Rather, this pattern likely reflects a structural canopy transition coinciding with the end of the wet season in May: as the semi-deciduous Atlantic Forest undergoes partial leaf shedding, the leaf area index decreases, reducing interception surface at both points toward a common lower bound. This interpretation is consistent with the SHAP temporal dynamics, which show reduced importance of P_e and $P_e \times API_{7d}$ during transitional months, suggesting that structural canopy effects temporarily dominate over antecedent moisture effects.

The persistent interception hotspot behavior of P120 has direct implications beyond the water balance. Critical zones with high interception efficiency act as selective filters that reduce the quantity but increase the chemical concentration of water reaching the soil, enriching stemflow and throughfall with canopy-leached nutrients (STAELENS et al., 2006). During the onset of the rainy season and following the 2014–2017 multi-year drought documented at this site (RODRIGUES et al., 2021a; GUAUQUE-MELLADO et al., 2026), the buffering capacity of critical zones likely delayed soil moisture recovery and deep percolation relative to stable zones, a hydrological asymmetry that spatial averaging across monitoring points would entirely mask. In contrast, the Stable Point (P30), characterized by greater canopy openness and more efficient lateral drainage, maintained a more linear relationship between gross precipitation and effective precipitation throughout the study period, making it a more reliable predictor of water inputs to the soil system under variable rainfall regimes.

5.4.4 Machine Learning as a Diagnostic Tool Beyond the "Black Box"

The integration of Optuna and SHAP transformed the XGBoost algorithm into a diagnostic tool rather than a purely predictive model. The main discrepancy between traditional feature importance and SHAP values is that importance rankings favor the categorical intensity variable (*Pe_quartile*), whereas SHAP prioritizes the continuous magnitude (P_e). This divergence reflects two distinct physical mechanisms. Rainfall intensity (*Pe_quartile*) modulates canopy response through its association with kinetic energy: high-intensity drops mechanically dislodge intercepted water from leaf surfaces, reducing the effective canopy storage capacity, and intense events frequently co-occur with wind, which further reduces storage through aerodynamic evaporation and branch agitation (HOLWERDA et al., 2012). In contrast, gross precipitation magnitude (P_e) governs the cumulative throughfall response: once the canopy storage threshold is exceeded, higher P_e volumes progressively activate the entire internal drainage network, drip points, stemflow pathways, and branch-channelling, while simultaneously maintaining continuous water supply to canopy gaps where throughfall is structurally unobstructed. The categorical quartile feature thus acts as a proxy for the energetic regime of the event, while the continuous P_e captures the volumetric driver of saturation and drainage activation. This distinction is analogous to that reported by Nanko et al. (2025), where canopy structural metrics define the initial rainfall partitioning pathways (drip vs. splash throughfall) while abiotic rainfall drivers dominate the marginal volumetric response event by event. The temporal dynamics of XGBoost feature importance (Figure 5.9) further illuminate how the model learns from heterogeneous data across a decade containing both wet and drought conditions.

The model also favored parsimony, disregarding spurious features such as *days_without_rain*, instead favoring physically meaningful antecedent water mass variables represented by the API. This selectivity avoids biases associated with spurious correlations (KING et al., 2022) and the over-parameterized models criticized by Haddad et al. (2013), ensuring that predictions are grounded in real hydrological processes rather than statistical noise.

A further methodological contribution is its representation of the full spatiotemporal dynamics of the Matinha forest through a *single unified model*. Traditional interception models confront a fundamental trade-off: either the entire forest is treated as a homogeneous unit—introducing spatial aggregation biases of up to 150% (MUZYLO et al., 2009)—or independent parameter sets must be calibrated for each monitoring point, requiring in the present case 32 separate model-calibration procedures. The ensemble architecture trained here overcomes this trade-off by jointly learning the behavior of all 32 spatially heterogeneous plots by means of a physically enhanced machine learning framework.

5.4.5 Seasonal Fidelity, Aggregation Effects, and the Role of Drought

Despite weak event-scale performance under TSCV, the model reproduced observed monthly dynamics with high fidelity; temporal aggregation effectively cancels high-frequency noise and reveals the dominant climatic signal. This phenomenon is well documented: Crockford e Richardson (2000) and Dilmi et al. (2017) showed that temporal aggregation stabilizes interception estimates, while Muzylo et al. (2009) argued that interception gains greater hydrological meaning at seasonal and annual scales. The pronounced bimodal distribution visible in May captures two coexisting hydrological regimes: a lower-C mode (Q25 = 0.509, concentrated near 0.51) and a higher-C mode (Q75 = 0.925, concentrated near 0.90). This bimodality is a direct manifestation of the threshold behavior analytically described by the P'_g and S thresholds in Gash e Morton (1978)'s model but here reproduced data-driven without explicit parameterization.

The substantially higher long-term mean interception fraction estimated in this study (38.3%, $C = 0.617$) compared to the 18.5% reported by Junqueira Junior et al. (2019) for a shorter monitoring period reflects the integration of drought-affected years and the framework's sensitivity to small-event dynamics. The extended 10.4-year record provides a more ecologically realistic picture of interception behavior under the variable hydroclimatic conditions characteristic of southeastern Brazil. Annual interception fractions reveal a consistent downward trend in recent years: from 52.1% in 2012 (September–December only; mean $P_e = 8.8$ mm) and 42.9% in 2013, through the drought period (2015: 40.1%; 2016: 40.8%; 2017: 36.5%), to a progressive decline in the post-drought years (2019: 35.4%; 2020: 32.3%; 2022: 31.7%; 2023: 18.9%). This downward trend coincides with a marked increase in mean event size, P_e rising from 12.5 mm in 2013 to 30.5 mm in 2022, suggesting that the reduction in interception fraction in recent years is partly driven by larger convective events that more readily saturate and exceed canopy storage capacity. A post-drought canopy recovery effect, increased leaf area and biomass accumulation with the return of normal rainfall potentially altering drainage architecture, cannot be excluded, as noted by Meir et al. (2015) and Guauque-Mellado et al. (2026), though the co-occurrence with larger events prevents unambiguous attribution. The 2012 value (52.1%) should be interpreted with caution as it covers only September–December with anomalously small events.

A direct comparison with the observational study of Rodrigues et al. (2022), conducted at the same Matinha remnant over a six-year period (2013–2019, encompassing 427 events across 31 gauges), further contextualizes these differences. Using traditional statistical descriptors, the mean relative difference (MRD), the coefficient of variation (CV), and the time stability index (TSI), they estimated a mean canopy interception of 22.4% of gross rainfall (range: 17.8%–27.7% across hydrological years), with the prolonged drought period accounting for 21.7% and the non-drought period for 20.2%. These values are systematically lower than the 38.3% long-term mean reported here, a discrepancy attributable to three compounding factors: (i) the present study extends two years earlier (from September 2012) and nearly four years later (to February 2023), capturing drier antecedent conditions and anomalously small events at both ends of the record; (ii) the ML framework is explicitly sensitive to sub-saturation small events, events with $P_e \leq 5$ mm represent 30.9% of all records and exhibit a mean interception fraction of 65.5%, nearly double the long-term average, confirming that small-event dynamics drive the disproportionate increase in total interception; and (iii) the event-scale effective precipitation coefficient (C) modeled here integrates stemflow in addition to throughfall, stemflow at this site represents approximately 2–4% of gross rainfall (RODRIGUES et al., 2021a), contributing at most 3–4 percentage points to the discrepancy, while the remaining $\sim$12 percentage points are attributable to the extended monitoring period and the model's sensitivity to sub-saturation events. Rodrigues et al. (2022) also showed that events with $P_e \leq 1$ mm, representing 5.3% of all events in our dataset and exhibiting interception fractions of 86.8%, exhibit the highest spatial variability, with CV stabilizing only above a minimum rainfall threshold. These are precisely the events that most inflate the long-term interception estimate in the present framework.

The convergence between both frameworks is equally informative. Rodrigues et al. (2022) reported that throughfall spatial variability was significantly less time-stable during the drought period than during the non-drought period, attributing this instability to the combined effect of increased low-intensity events and canopy structural changes driven by prolonged water deficit. This mechanism is further supported by the present dataset: our computed drought-period interception (39.3%) versus non-drought (37.9%) mirrors the directional pattern reported by Rodrigues et al. (2022), though at a systematically higher absolute level for the reasons described above. The present SHAP analysis independently corroborates this mechanism: the dominance of $P_e \times API_{7d}$ confirms that antecedent canopy wetness is the controlling variable across all validation schemes. Similarly, Rodrigues et al. (2022) identified tree density, mean diameter at breast height (DBH), and aboveground biomass (AGB) as the dominant forest-structural drivers of throughfall spatial distribution via redundancy analysis (RDA), a finding directly paralleled by the SHAP contribution of plot-specific indicator variables (e.g., $Point\_P120$ and $Point\_P42$) in the present framework. Both approaches point to the same thing: rainfall partitioning in the Matinha Atlantic Forest is governed by the interaction between antecedent moisture memory and fine-scale canopy structural heterogeneity, with drought acting as the primary driver of spatiotemporal instability.

Moreover, as Meir et al. (2015) and Guauque-Mellado et al. (2026) documented, prolonged drought induces structural changes in the canopy, accelerated leaf shedding, increased tree mortality, biomass loss, that progressively alter storage capacity S and canopy openness, introducing a non-stationarity in the interception response that purely meteorological predictors cannot fully capture. Incorporating dynamic canopy descriptors such as satellite-derived NDVI or LAI products (OLIVEIRA et al., 2024; GEBREMEDHIN et al., 2023) is a promising direction for future work.

5.4.6 Synthesis: Reframing the Interception Problem

Rather than answering “how much water is lost,” this study reframes the problem as “how, when, and where interception losses fluctuate.” Unlike prior approaches that either model the forest as a single homogeneous body or require independent calibration per spatial point, the proposed framework captures the full spatiotemporal interception dynamics of the Matinha remnant through a single, transferable architecture. This architecture integrates ecohydrological theory with interpretable ML to regionalize effective precipitation in heterogeneous forested landscapes.

5.4.7 Limitations and Future Directions

Several limitations must be acknowledged. First, the absence of wind speed and sub-hourly rainfall intensity measurements precludes explicit representation of intra-event dynamics and aerodynamic evaporation. Second, the static nature of the predictor set does not capture the structural non-stationarity introduced by drought-driven canopy thinning and tree mortality. Incorporating dynamic remote sensing indices (NDVI, LAI) as time-varying covariates (OLIVEIRA et al., 2024; GEBREMEDHIN et al., 2023) would help represent this structural non-stationarity. Third, the framework has been calibrated and validated for a semi-deciduous tropical forest; generalization to evergreen, boreal, or different rainfall climatologies should be validated against independent throughfall datasets.

Future work should prioritize: (i) sub-hourly monitoring campaigns to quantify intra-event uncertainty; (ii) coupled forest-inventory and throughfall monitoring to track how structural changes under drought propagate into interception hotspot dynamics; and (iii) integration of satellite-derived canopy descriptors as dynamic predictors.

5.4.8 Potential Applications

Beyond its scientific contributions, the proposed framework offers a scalable blueprint for operational hydrology in heterogeneous tropical landscapes. The superior spatial transferability of the XGBoost architecture facilitates the regionalization of net water availability in ungauged or partially instrumented forested catchments (Junqueira Junior et al., 2019; MELLO et al., 2024). The identification of critical versus stable plots provides a practical basis for optimizing monitoring networks and targeting ecohydrological interventions in areas that act as persistent interception hotspots. By leveraging high-resolution reanalysis or satellite-derived precipitation products, the *Intercep-ML* platform can simulate real-time canopy storage dynamics—a capability particularly relevant for mitigating severe hydrological droughts in Brazilian highland regions (MELLO et al., 2025), where interception losses can drastically reduce effective recharge during the dry season. These findings also inform the valuation of ecosystem services, providing a practical statistical basis for programs centered on water-yield conservation and payment for environmental services (SCHMIDT et al., 2020; RODRIGUES et al., 2021a).

5.4.9 Interactive Model Deployment and Decision-Support Platform

To bridge the gap between predictive modeling and hydrological management, the optimized ML framework was operationalized through an interactive web-based platform developed using the *Streamlit* library. The architecture integrates the best-performing ensemble (optimized XGBoost regressor) to dynamically quantify C_{Peff}, canopy interception loss (I_{ML}), and net effective precipitation ($P_{\mathrm{eff,ML}}$). Users can input rainfall event or continuous time series, upon which the system automatically computes temporal features and antecedent moisture indices. The platform is publicly accessible at:

https://peff-maps.streamlit.app/

5.5 Conclusions

Physically informed machine-learning frameworks can substantially improve the understanding and regionalization of rainfall interception and effective precipitation in heterogeneous forest ecosystems. The following conclusions emerge from more than a decade of high-resolution monitoring integrated with ensemble learning and model interpretability.

First, the XGBoost-based framework shows strong spatial transferability while also revealing the temporal non-stationarity of interception at the event scale. The superior performance under spatial cross-validation (LOPO, $R^2 \approx 0.652$, KGE ≈ 0.696) confirms that the model successfully internalizes site-invariant physical controls and enables reliable extrapolation to non-instrumented plots.

Second, the interpretability analysis provides clear physical validation. The dominance of antecedent precipitation indices and their interaction with event rainfall confirms that interception in the Atlantic Forest is a state-dependent process controlled by hydrological memory, challenging the widespread assumption of static storage parameters in traditional interception models.

Third, the results reveal pronounced non-linearities and physical thresholds in rainfall partitioning. Seasonal features modulate these thresholds by implicitly capturing phenological changes in the semi-deciduous canopy, allowing the model to reproduce observed seasonal cycles with high fidelity, particularly during the dry season ($R^2 \approx 0.918$).

Fourth, the study identifies ecohydrological hotspots within the forest. The identification of critical and stable monitoring points—with differences in interception fraction of up to 60.8% for small events—demonstrates that the Atlantic Forest operates as a spatial mosaic of hydrological sinks and transmission pathways. Ignoring this sub-canopy heterogeneity through spatial averaging can lead to substantial bias in water-balance estimates.

Fifth, this work demonstrates that interpretable ML is a diagnostic ecohydrological tool, not only a predictive one. The combined use of ensemble learning, Bayesian optimization, and SHAP analysis ensures that model skill reflects physically meaningful processes. This architecture was operationalized through the *Intercep-ML* platform, translating methodological advances into a tangible and reusable product.

Overall, the findings redefine rainfall interception in the Atlantic Forest as a dynamic, memory-driven, and spatially heterogeneous process. The proposed framework also provides a practical basis for assessing future water availability under climate-change scenarios, particularly under the increasing frequency of extreme droughts projected for southeastern Brazil.

Acknowledgements

The authors thank CAPES (Finance Code 001) and FAPEMIG for the scholarship granted to the first author, and CNPq (Grant No. 402066/2023-5) for funding this project.

REFERENCES

ARSENAULT, R. et al. Continuous streamflow prediction in ungauged basins : long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, p. 139–157, 2023. Doi: https://doi.org/10.5194/hess-27-139-2023.

BAKAR, A. A.; LATIF, Z. A.; LEE, W.-K. COMPARISON OF CANOPY COVER EFFECTS ON RAINFALL INTERCEPTION LOSS IN TROPICAL FOREST WITH HOMOGENOUS AND MIXED TREE. *Jounal Teknologi*, v. 2, n. 25, p. 1–9, 2023. Doi: https://doi.org/10.11113/jurnalteknologi.v85.17661.

BARNETT, A. G.; BAKER, P.; DOBSON, A. J. Analysing Seasonal Data. *The R Journal*, p. 5–10, 2012. Doi: 10.32614/RJ-2012-001.

BEZERRA, R. et al. Towards more accurate very short-range flood forecasts with physics-enhanced machine learning models. *Journal of Hydroinformatics*, v. 00, n. 0, p. 1–23, 2026.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Doi: 10.1023/A:1010933404324.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [*S.l.*]: ACM, 2016. p. 785–794. Doi: 10.1145/2939672.2939785.

CROCKFORD, R. H.; RICHARDSON, D. P. Partitioning of rainfall into throughfall, stemflow and interception: effect of forest type, ground cover and climate. *Hydrological Processes*, v. 14, n. 16-17, p. 2903–2920, 2000.

DAS, P.; SACHINDRA, D. A.; CHANDA, K. Machine Learning - Based Rainfall Forecasting with Multiple Non - Linear Feature Selection Algorithms. *Water Resources Management*, Springer Netherlands, p. 6043–6071, 2022. ISSN 1573-1650. Doi: 10.1007/s11269-022-03341-8. Disponível em: https://doi.org/10.1007/s11269-022-03341-8.

DILMI, M. D. et al. Data-driven clustering of rain events : microphysics information derived from macro-scale observations. p. 1557–1574, 2017. Doi: 10.5194/amt-10-1557-2017.

FALLAHI, M. et al. Future Water Yield Projections Under Climate Change Using Optimized and Downscaled Models via the MIDAS Approach. *Environments*, p. 1–26, 2025. Doi: doi.org/10.3390/environments12090303.

FATHIZADEH, O. et al. Spatial Variability and Optimal Number of Rain Gauges for Sampling Throughfall under Single Oak Trees during the Leafless Period. *Forest*, p. 1–15, 2021. Doi: https://doi.org/10.3390/f12050585.

FILHO, J. C. R. et al. Partitioning of rainfall into throughfall , stemflow , and interception loss in the Brazilian Northeastern Atlantic Forest de interceptação vegetal no semiárido nordestino. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 23, n. 1807-1929, p. 21–26, 2019. Doi: http://dx.doi.org/10.1590/1807-1929/agriambi.v23n1p21-26 Partitioning.

GASH, J. H. C.; MORTON, A. J. AN APPLICATION OF THE RUTTER MODEL TO THE ESTIMATION OF THE INTERCEPTION LOSS FROM THETFORD FOREST. *Journal of Hydrology*, v. 38, p. 49–58, 1978. Doi: https://doi.org/10.1016/0022-1694(78)90131-2.

GEBREMEDHIN, M. A. et al. Spatio-temporal rainfall interception loss at the catchment scale from earth observation in a data-scarce area , Northern Ethiopia. *Journal of Hydrology*, Elsevier B.V., v. 626, n. PA, p. 130126, 2023. ISSN 0022-1694. Doi: 10.1016/j.jhydrol.2023.130126. Disponível em: https://doi.org/10.1016/j.jhydrol.2023.130126.

GERMER, S.; ELSENBEER, H.; MORAES, J. M. Throughfall and temporal trends of rainfall redistribution in an open tropical rainforest, south-western amazonia (rondônia, brazil). *Hydrology and Earth System Sciences*, v. 10, n. 3, p. 383–393, 2006. Doi: 10.5194/hess-10-383-2006.

GUAUQUE-MELLADO, D. et al. Long-term water balance dynamics in a Brazilian neotropical forest: Insights from seven years of continuous observation. *Journal of Environmental Management*, v. 398, 2026. Doi: 10.1016/j.jenvman.2025.128475.

GUPTA, H. V. et al. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, v. 377, n. 1–2, p. 80–91, 2009. Doi: 10.1016/j.jhydrol.2009.08.003.

HADDAD, K. et al. Applicability of Monte Carlo cross validation technique for model development and validation using generalised least squares regression. *Journal of Hydrology*, v. 482, p. 119–128, 2013. ISSN 0022-1694. Doi: 10.1016/j.jhydrol.2012.12.041. Disponível em: http://dx.doi.org/10.1016/j.jhydrol.2012.12.041.

HOLWERDA, F. et al. Wet canopy evaporation from a Puerto Rican lower montane rain forest : The importance of realistically estimated aerodynamic conductance. *Journal of Hydrology*, Elsevier B.V., v. 414-415, p. 1–15, 2012. ISSN 0022-1694. Doi: 10.1016/j.jhydrol.2011.07.033. Disponível em: http://dx.doi.org/10.1016/j.jhydrol.2011.07.033.

Junqueira Junior, J. A. et al. Time-stability of soil water content (SWC) in an Atlantic Forest - Latosol site. *Geoderma*, 2017. Doi: 10.1016/j.geoderma.2016.10.034.

Junqueira Junior, J. A. et al. Rainfall partitioning measurement and rainfall interception modelling in a tropical semi-deciduous Atlantic forest remnant. *Agricultural and Forest Meteorology*, Elsevier, v. 275, n. May, p. 170–183, 2019. ISSN 0168-1923. Doi: 10.1016/j.agrformet.2019.05.016. Disponível em: https://doi.org/10.1016/j.agrformet.2019.05.016.

KE, G. et al. LightGBM : A Highly Efficient Gradient Boosting Decision Tree. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. [*s.n.*], 2017. p. 1–9. ISBN 9781510860964. Doi: 10.5555/3294996.3295074. Disponível em: https://dl.acm.org/doi/10.5555/3294996.3295074.

KING, F. et al. DeepPrecip : a deep neural network for precipitation retrievals. p. 6035–6050, 2022.

LIMOUSIN, J.-m. et al. Modelling rainfall interception in a mediterranean Quercus ilex ecosystem : Lesson from a throughfall exclusion experiment. *Journal of Hydrology*, p. 57–66, 2008. Doi: 10.1016/j.jhydrol.2008.05.001.

LINHOSS, A. C.; SIEGERT, C. M. Calibration reveals limitations in modeling rainfall interception at the storm scale. *Journal of Hydrology*, Elsevier, v. 584, n. February 2019, p. 124624, 2020. ISSN 0022-1694. Doi: 10.1016/j.jhydrol.2020.124624. Disponível em: https://doi.org/10.1016/j.jhydrol.2020.124624.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: IEEE. *2008 Eighth IEEE International Conference on Data Mining (ICDM)*. [*S.l.*], 2008. p. 413–422. Doi: 10.1109/ICDM.2008.17.

LIU, S. A new model for the prediction of rainfall interception in forest canopies. *Ecological Modelling*, v. 99, 1997. Doi: https://doi.org/10.1016/S0304-3800(97)01948-0.

LUCAS-BORJA, M. E. et al. Assessing canopy rainfall partitioning by Mediterranean dryland shrubs under extreme rainfall. n. September, p. 1–14, 2023. Doi: 10.1002/hyp.15007.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *NeurIPS 2017*. [*S.l.: s.n.*], 2017. p. 4765–4774.

MAHMOOD, Z.; KHAN, S. On the Use of K-Fold Cross-Validation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise Regression On the Use of K-Fold Cross-Validation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise. *The International Journal of Biostatistics*, v. 5, n. 1, 2009. Doi: 10.2202/1557-4679.1105.

MEIR, P. et al. Threshold Responses to Soil Moisture Deficit by Trees and Soil in Tropical Rain Forests : Insights from Field Experiments. v. 65, n. 9, p. 882–892, 2015. Doi: 10.1093/biosci/biv107.

MELLO, C. R. et al. Mitigating severe hydrological droughts in the Brazilian tropical highland region : A novel land use strategy under climate change. *International Soil and Water Conservation Research*, v. 13, 2025. Doi: 10.1016/j.iswcr.2025.03.005.

MELLO, C. R. D. et al. Deciphering global patterns of forest canopy rainfall interception (FCRI): A synthesis of geographical , forest species , and methodological influences. v. 358, n. March, 2024.

MOHAMMED, J.; MENGISTE, Y.; GEBREMICHAEL, M. Enhancing monthly precipitation forecasting by integrating multi - source data with machine learning models : a study in the Upper Blue Nile Basin. *Modeling Earth Systems and Environment*, Springer International Publishing, v. 11, n. 1, p. 1–16, 2025. ISSN 2363-6211. Doi: 10.1007/s40808-024-02223-9. Disponível em: https://doi.org/10.1007/s40808-024-02223-9.

MUZYLO, A. et al. A review of rainfall interception modelling. *Journal of Hydrology*, Elsevier B.V., v. 370, n. 1-4, p. 191–206, 2009. ISSN 0022-1694. Doi: 10.1016/j.jhydrol.2009.02.058. Disponível em: http://dx.doi.org/10.1016/j.jhydrol.2009.02.058.

NANKO, K. et al. Machine Learning Reveals the Contrasting Roles of Rainfall and Canopy Structure Metrics on the Formation of Canopy Drip and Splash Throughfall. *Journal of Geophysical Research: Biogeosciences*, v. 130, n. 2, p. e2024JG008340, 2025. Doi: 10.1029/2024JG008340.

NÁVAR, J. Fitting rainfall interception models to forest ecosystems of Mexico. *Journal of Hydrology*, v. 548, p. 458–470, 2017. Doi: 10.1016/j.jhydrol.2017.03.025.

OLIVEIRA, S. et al. Enhancing global rainfall interception loss estimation through vegetation structure modeling. *Journal of Hydrology*, v. 631, n. January, 2024. Doi: 10.1016/j.jhydrol.2024.130672.

PROKHORENKOVA, L. et al. CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*. [*S.l.*]: Curran Associates, Inc., 2018. v. 31. Doi: 10.48550/arXiv.1706.09516.

PYPKER, T. G. et al. The importance of canopy structure in controlling the interception loss of rainfall: Examples from a young and an old-growth Douglas-fir forest. *Agricultural and Forest Meteorology*, v. 130, n. 1-2, p. 113–129, 2005.

ROBERTS, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, v. 40, n. 8, p. 913–929, 2017.

RODRIGUES, A. F. et al. Modeling canopy interception under drought conditions : The relevance of evaporation and extra sources of energy. *Journal of Environmental Management*, v. 292, n. April, 2021. Doi: 10.1016/j.jenvman.2021.112710.

RODRIGUES, A. F. et al. Water balance of an atlantic forest remnant under a prolonged drought period. *Ciencia e Agrotecnologia*, v. 45, 2021. ISSN 19811829. Doi: 10.1590/1413-7054202145008421.

RODRIGUES, A. F. et al. Throughfall spatial variability in a neotropical forest: Have we correctly accounted for time stability? *Journal of Hydrology*, v. 608, p. 127632, 2022. ISSN 0022-1694.

SARI, V.; PAIVA, E. M. C. D. de; PAIVA, J. B. D. de. Interceptação da chuva em diferentes formações florestais na região sul do Brasil. *Revista Brasileira de Recursos Hídricos*, v. 21, p. 65–79, 2016. Doi: 10.21168/rbrh.v21n1.p65-79.

SCHMIDT, L. et al. Challenges in Applying Machine Learning Models for Hydrological Inference : A Case Study for Flooding Events Across Germany. 2020. Doi: 10.1029/2019WR025924.

SIDLE, R. C.; ZIEGLER, A. D. The canopy interception – landslide initiation conundrum : insight from a tropical secondary forest in northern Thailand. *Hyrology and Earth System Sciences*, p. 651–667, 2017. Doi: 10.5194/hess-21-651-2017.

SPITTLEHOUSE, D. L.; MALONEY, D. D. Rainfall Interception by Mature Coastal and Interior Forests in British Columbia. *Journal of Watershed Science and Management*, v. 6, n. 01, p. 1–20, 2023. Doi: 10.22230/jwsm.2023v6n1a55.

SREENIVASU, S. et al. Rainfall Prediction Using Machine Learning. *2024 2nd International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC)*, IEEE, p. 140–149, 2024. Doi: 10.1109/ICMACC62921.2024.10894486.

STAELENS, J. et al. Spatial variability and temporal stability of throughfall deposition under beech (fagus sylvatica l.) in relationship to canopy structure. *Environmental Pollution*, v. 142, n. 2, p. 254–263, 2006. Doi: 10.1016/j.envpol.2005.10.002.

TARASOVA, L. et al. A Process - Based Framework to Characterize and Classify Runoff Events : The Event Typology of Germany. *Water Resources Research*, 2020. Doi: 10.1029/2019WR026951.

WILLARD, J. D. et al. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, v. 55, n. 4, p. 1–35, 2022.

YU, Y. et al. Evaluating the influential variables on rainfall interception at different rainfall amount levels in temperate forests. *Journal of Hydrology*, Elsevier B.V., v. 615, n. PA, p. 128572, 2022. ISSN 0022-1694. Doi: 10.1016/j.jhydrol.2022.128572. Disponível em: https://doi.org/10.1016/j.jhydrol.2022.128572.

ZHANG, Z.-s. et al. Gross rainfall amount and maximum rainfall intensity in 60-minute influence on interception loss of shrubs : a 10-year observation in the Tengger Desert. *Nature Publishing Group*, Nature Publishing Group, n. April, p. 1–10, 2016. Doi: 10.1038/srep26030. Disponível em: http://dx.doi.org/10.1038/srep26030.

ZHONG, F. et al. Revisiting large-scale interception patterns constrained by a synthesis of global experimental data. *EGU*, p. 5647–5667, 2022. Doi: doi.org/10.5194/hess-26-5647-2022.

ZIMMERMANN, A.; ZIMMERMANN, B. Agricultural and Forest Meteorology Requirements for throughfall monitoring : The roles of temporal scale and canopy complexity. *Agricultural and Forest Meteorology*, Elsevier B.V., v. 189-190, p. 125–139, 2014. ISSN 0168-1923. Doi: 10.1016/j.agrformet.2014.01.014. Disponível em: http://dx.doi.org/10.1016/j.agrformet.2014.01.014.

GLOSSÁRIO

Borda score: Multi-objective ranking aggregation method: $\mathrm{BordaScore}(m) = \mathrm{Rank}_{\mathrm{RMSE}}(m) + \mathrm{Rank}_{\mathrm{KGE}}(m)$; the model with the lowest cumulative score performs best

Canopy interception: Fraction of precipitation temporarily retained on plant surfaces (leaves, branches, bark) and returned to the atmosphere by evaporation, without reaching the soil

Canopy storage capacity: Maximum volume of water the canopy can retain by capillary retention before drainage begins (S, mm)

Cross-validation: Model evaluation procedure that repeatedly partitions data into training and test subsets to estimate generalisation performance

Data-driven hydrology: Hydrological modelling approach that relies on machine learning algorithms and large observational datasets rather than explicit physical equations to represent hydrological processes

Drizzle effect: Tendency of satellite precipitation products to overestimate the frequency of low-intensity events and underestimate high-intensity events, producing distributions with shorter tails than observed

Ecohydrology: Interdisciplinary field that studies the interactions between vegetation and the hydrological cycle, including rainfall partitioning, evapotranspiration, and baseflow regulation

Effective precipitation: Fraction of gross rainfall that reaches the forest floor, defined as the sum of throughfall and stemflow ($P_{\mathrm{eff}} = P_i + S_f$)

El Niño–Southern Oscillation (ENSO): Climate phenomenon of interannual variability in the equatorial Pacific Ocean that modulates precipitation patterns at the global scale

***Ensemble learning*:** Machine learning paradigm that combines multiple base models (e.g., decision trees) to produce more robust and generalisable predictions

Feature importance: Measure of the contribution of each input variable to a model's predictions; computed via SHAP values or internal ensemble metrics

Forest interception: Portion of precipitation captured by the forest canopy and evaporated back to the atmosphere without reaching the soil

Free throughfall coefficient: Fraction of rainfall that reaches the soil directly through canopy gaps without contacting the vegetation (p, dimensionless)

Gap-filling: Process of estimating missing values in time series using statistical or machine learning methods

Gash model: Analytical interception model developed by Gash (1979) that estimates interception losses from canopy characteristics (S, $\bar{c}$, $\bar{E}/\bar{R}$) and rainfall event properties

Gross rainfall: Total rainfall depth measured above the forest canopy before any partitioning by the canopy occurs

Hydrological memory: Influence of antecedent soil moisture conditions on the current hydrological response; commonly represented by antecedent precipitation indices (API)

Hyperparameter tuning: Optimisation of a model's configuration parameters (e.g., number of trees, learning rate) that are not learned during training

Leaf Area Index (LAI): Total leaf area per unit ground area ($m^2\,m^{-2}$); a structural canopy descriptor that governs interception, evapotranspiration, and forest microclimate

Outlier: An observation that departs significantly from the expected pattern of the data, potentially indicating measurement error or a genuine extreme event

Quantile mapping: Bias correction technique that adjusts the cumulative distribution of an estimated product (e.g., satellite data) to match that of an observed reference: $\hat{x}_t = F_O^{-1}(F_P(x_t))$

Rainfall reconstruction: Process of estimating missing or unrecorded precipitation data using statistical methods, interpolation, or machine learning trained on neighbouring stations or satellite products

Rutter model: Continuous interception simulation model (Rutter et al., 1975) that tracks hourly canopy and trunk water balances using differential equations

South Atlantic Convergence Zone (SACZ): Atmospheric system that manifests as a persistent band of cloudiness and moisture convergence oriented northwest–southeast, responsible for a substantial fraction of summer precipitation in southeastern Brazil

Spatiotemporal variability: Variability of a hydrological or climatic process across both spatial and temporal dimensions simultaneously

Stemflow**:** Fraction of rainfall intercepted by the canopy and channelled along branches and the trunk to the soil at the tree base

STL (*Seasonal-Trend decomposition via LOESS*): Time series decomposition method that separates trend, seasonality, and residual components using local polynomial smoothing (LOESS)

Throughfall**:** Fraction of rainfall that passes through the canopy via gaps between crowns or by dripping from leaves and reaches the soil surface

Water balance: Conservation equation for water in a catchment or ecosystem: $P = \mathrm{ET} + Q + R \pm \Delta S$