UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Instituto de Ciência e Tecnologia - ICTS- campus de Sorocaba

RENAN DE PAULO

Quantificação de Incertezas em Previsões Pluviométricas com Conformal Prediction

Sorocaba

2026

Renan de Paulo

Quantificação de Incertezas em Previsões Pluviométricas com Conformal Prediction

Trabalho de Conclusão de Curso, apresentada à Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia - ICTS, Sorocaba, para obtenção do título de Bacharelem Engenharia de Controle e Automação.

Orientador: Prof. Dr. Leopoldo André Dutra Lusquino Filho

Sorocaba
2026

P331q

Paulo, Renan de
Quantificação de incertezas em previsões pluviométricas com Conformal Prediction / Renan de Paulo. -- Sorocaba, 2026
59 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Engenharia de Controle e Automação) - Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia, Sorocaba
Orientador: Leopoldo André Dutra Lusquino Filho

1. Análise de regressão. 2. Distribuição (Probabilidades). 3. Mineração de dados (Computação). 4. Biometeorologia. 5. Hydrography. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

RENAN DE PAULO

QUANTIFICAÇÃO DE INCERTEZAS EM PREVISÕES PLUVIOMÉTRICAS COM CONFORMAL PREDICTION

Trabalho de Conclusão de Curso apresentado à Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia - ICTS, Sorocaba, para obtenção do título de Bacharel em Engenharia de Controle e Automação.

Data de defesa: 03/06/2026

BANCA EXAMINADORA

Prof. Dr. Leopoldo André Dutra Lusquino Filho
UNESP – Instituto de Ciência e Tecnologia - ICTS – Campus de Sorocaba

---

Prof. Dr. William Dantas Vichete
UNESP – Instituto de Ciência e Tecnologia - ICTS – Campus de Sorocaba

---

Prof. Dr. Leandro Santiago de Araujo
UFF – Universidade Federal Fluminense

Dedico este trabalho à minha companheira Giovana, que foi a grande incentivadora para que eu continuasse no curso, e à República Magnatas.

RESUMO

Este trabalho investiga a aplicação de métodos de *Conformal Prediction* (CP) na construção de intervalos preditivos para previsão pluviométrica diária, com foco na quantificação de incerteza em múltiplos horizontes de previsão. O problema é motivado pela natureza altamente irregular e não linear da precipitação, em que a obtenção de estimativas pontuais isoladas é insuficiente para apoiar decisões confiáveis. Quatro modelos preditivos são comparados: *Seasonal Autoregressive Integrated Moving Average with Exogenous Variables* (SARIMAX), como *benchmark* estatístico clássico; *Random Forest* (RF), como método baseado em *ensemble* de árvores; e duas arquiteturas neurais recorrentes, *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU). Cada modelo é combinado a três estratégias conformes: *Inductive Conformal Prediction* (ICP), *Mondrian Conformal Prediction* e *Conformalized Quantile Regression* (CQR). Os experimentos são conduzidos com dados meteorológicos reais da estação de Brasília, disponibilizados pelo Instituto Nacional de Meteorologia (INMET), para os horizontes de um, sete e trinta dias. Uma contribuição metodológica central do trabalho é o tratamento explícito da violação da hipótese de permutabilidade que séries temporais impõem ao CP: na ausência dessa garantia teórica, a validade dos intervalos é verificada empiricamente por meio da cobertura observada no conjunto de teste. Os resultados mostram que LSTM, GRU e SARIMAX superam o Random Forest tanto em acurácia pontual quanto na qualidade dos intervalos conformes. No horizonte de curto prazo, a combinação LSTM com ICP se destaca por reunir cobertura empírica próxima ao nível nominal de noventa por cento e a menor largura média entre as combinações competitivas, constituindo o resultado mais relevante do trabalho. Nos horizontes de médio e longo prazo, o Mondrian se mostra a estratégia conforme mais consistente em termos de equilíbrio entre cobertura e largura dos intervalos, enquanto o CQR apresenta maior sensibilidade a picos e oscilações locais da série. Os achados indicam que a qualidade do preditor base influencia diretamente a utilidade dos intervalos conformes e que diferentes estratégias de CP favorecem dimensões distintas da qualidade intervalar, o que reforça a importância de avaliar métodos de quantificação de incerteza sob critérios complementares em problemas de previsão pluviométrica.

Palavras-Chave: previsão pluviométrica; quantificação de incerteza; *conformal prediction*; aprendizado de máquina; redes neurais recorrentes; séries temporais.

ABSTRACT

This work investigates the application of Conformal Prediction (CP) methods for constructing predictive intervals in daily rainfall forecasting, with a focus on uncertainty quantification across multiple prediction horizons. The problem is motivated by the highly irregular and nonlinear nature of precipitation, where point estimates alone are insufficient to support reliable decision-making. Four predictive models are compared: Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX), as a classical statistical *benchmark*; Random Forest (RF), as an *ensemble-based* tree method; and two recurrent neural architectures, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Each model is combined with three conformal strategies: Inductive Conformal Prediction (ICP), Mondrian Conformal Prediction, and Conformalized Quantile Regression (CQR). Experiments are conducted using real meteorological data from the Brasília weather station, made available by the Brazilian National Institute of Meteorology (INMET), for prediction horizons of one, seven, and thirty days. A central methodological contribution of this work is the explicit treatment of the exchangeability assumption violation that time series impose on CP: in the absence of this theoretical guarantee, the validity of the intervals is assessed empirically through the *coverage* observed on the test set. Results show that LSTM, GRU, and SARIMAX outperform Random Forest in both point accuracy and conformal interval quality. At the short-term horizon, the combination of LSTM with ICP stands out by achieving empirical *coverage* close to the nominal ninety percent level alongside the smallest average interval width among competitive combinations, representing the most relevant finding of this work. At medium and long-term horizons, Mondrian proves to be the most consistent conformal strategy in terms of the balance between *coverage* and interval width, while CQR shows greater sensitivity to peaks and local oscillations in the series. The findings indicate that the quality of the base predictor directly influences the utility of conformal intervals, and that different CP strategies favor distinct dimensions of interval quality, reinforcing the importance of evaluating uncertainty quantification methods under complementary criteria in rainfall forecasting problems.

Keywords: rainfall forecasting; uncertainty quantification; conformal prediction; machine learning; recurrent neural networks; time series.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CP	Conformal Prediction
CQR	Conformalized Quantile Regression
CPU	Central Processing Unit
DDR5	Double Data Rate 5
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
ICP	Inductive Conformal Prediction
INMET	Instituto Nacional de Meteorologia
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MSE	Mean Squared Error
OOB	Out-of-Bag
QRF	Quantile Regression Forest
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Exogenous Variables
VS Code	Visual Studio Code

LISTA DE SÍMBOLOS

α	Nível de significância do intervalo preditivo
$\hat{\delta}_{1-\alpha}$	Quantil conformal do CQR
ε_t	Termo de erro no instante t
$\hat{f}$	Modelo preditivo ajustado
f_t	Porta de esquecimento da LSTM no instante t
h	Horizonte de previsão
h_t	Estado oculto da rede recorrente no instante t
H	Horizonte de previsão nos experimentos (1, 7 ou 30 dias)
i_t	Porta de entrada da LSTM no instante t
$I(\cdot)$	Função indicadora
L_i	Limite inferior do intervalo preditivo para a observação i
LB	Comprimento da janela histórica (365 dias)
$\overline{W}$	Largura média dos intervalos preditivos
P	Operador de probabilidade
p	Número de defasagens consideradas ou ordem autorregressiva não sazonal, conforme o contexto
$\hat{q}_{1-\alpha}$	Quantil empírico de nível $1-\alpha$
$\hat{q}_{lo}$	Limite inferior estimado pelo modelo quantílico
$\hat{q}_{hi}$	Limite superior estimado pelo modelo quantílico
r_t	Porta de reinicialização da GRU no instante t
R^2	Coeficiente de determinação
s_i	Escore de não conformidade da observação i
σ	Função sigmoide
T_b	Previsão da b-ésima árvore do Random Forest

U_i	Limite superior do intervalo preditivo para a observação i
W_i	Largura do intervalo preditivo para a observação i
x_t	Vetor de covariáveis no instante t
$\hat{y}_t$	Valor previsto no instante t
y_t	Valor observado no instante t
z_t	Porta de atualização da GRU no instante t
$\odot$	Multiplicação elemento a elemento (produto de Hadamard)
ϕ	Função de ativação aplicada ao estado oculto na RNN básica
ψ	Função de ativação aplicada à saída da RNN básica
$\tilde{c}_t$	Estado candidato da célula de memória da LSTM no instante t
$\Phi_P(B^s)$	Polinômio autorregressivo sazonal de ordem P
$\phi_p(B)$	Polinômio autorregressivo não sazonal de ordem p
$(1-B)^d$	Operador de diferenciação não sazonal de ordem d
$(1-B^s)^D$	Operador de diferenciação sazonal de ordem D
$\Theta_Q(B^s)$	Polinômio de médias móveis sazonal de ordem Q
$\theta_q(B)$	Polinômio de médias móveis não sazonal de ordem q
B	Operador defasagem
s	Período sazonal da série temporal

SUMÁRIO

1 INTRODUÇÃO

A capacidade de prever fenômenos naturais com elevado grau de confiança constitui um requisito fundamental em diversas áreas de aplicação prática. Na gestão de recursos hídricos, no planejamento urbano, na agricultura e na resposta a desastres naturais, decisões críticas dependem não apenas de estimativas pontuais sobre eventos futuros, mas também da confiança que se pode depositar nessas estimativas. No caso específico da precipitação pluviométrica, essa exigência é ainda mais evidente: erros de previsão podem resultar em falhas no dimensionamento de sistemas de drenagem, em respostas inadequadas a eventos extremos ou em alocações ineficientes de recursos hídricos. Nesse contexto, prever com precisão não é suficiente; é igualmente necessário quantificar a incerteza associada à previsão, de modo a fornecer informações completas e confiáveis para o processo decisório (Papacharalampous; Tyralis, 2022).

A previsão de séries temporais pluviométricas é, contudo, uma tarefa intrinsecamente difícil. A precipitação diária apresenta comportamento altamente não linear, com forte intermitência, grande frequência de valores nulos, heterocedasticidade marcada e ocorrência esporádica de eventos extremos. Essas características impõem desafios consideráveis aos métodos de previsão tradicionais, que frequentemente pressupõem estruturas lineares ou distribuições paramétricas bem comportadas. Modelos estatísticos clássicos, como os da família ARIMA e suas extensões sazonais, podem capturar componentes estruturais da série, mas tendem a apresentar limitações diante de relações não lineares complexas entre a precipitação e as covariáveis atmosféricas que a influenciam (Hyndman; Athanasopoulos, 2021). Essa não linearidade, combinada à natureza estocástica do fenômeno, torna a construção de intervalos preditivos confiáveis um problema aberto e de relevância tanto científica quanto aplicada.

Os algorítimos da área de aprendizado de máquina surgem como uma alternativa promissora para a previsão de séries temporais com alta não linearidade, oferecendo modelos capazes de capturar relações complexas entre variáveis sem a necessidade de especificações funcionais explícitas (Papacharalampous; Tyralis, 2022). Arquiteturas recorrentes como *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU) demonstraram capacidade de modelar dependências temporais de longo alcance, tornando-se referências em tarefas de *forecasting* hidrológico e meteorológico (Hewamalage; Bergmeir; Bandara, 2021). Métodos baseados em *ensemble*, como o *Random Forest*, também têm sido aplicados com sucesso em problemas de previsão temporal, por sua robustez a sobreajuste e sua capacidade de lidar com interações não lineares entre atributos (Bontempi; Taieb; Borgne, 2013). Entretanto, apesar dos avanços em acurácia pontual, modelos de aprendizado de máquina introduzem novas fontes de incerteza que não estavam presentes nos métodos estatísticos tradicionais. A estocasticidade inerente ao processo de treinamento, a sensibilidade à inicialização dos parâmetros, a variabilidade decorrente de diferentes partições dos dados e a dificuldade de calibrar distribuições preditivas a partir de modelos determinísticos tornam a quantificação de incerteza um desafio adicional e frequentemente

negligenciado nesse contexto (Tyralis; Papacharalampous, 2022).

O *Conformal Prediction* (CP) apresenta-se como uma solução metodologicamente atrativa para esse problema (Shafer; Vovk, 2008). Trata-se de um arcabouço estatístico para a construção de intervalos preditivos com garantias de cobertura em amostras finitas, agnóstico em relação ao modelo de base utilizado, o que permite sua aplicação sobre qualquer preditor, desde modelos lineares clássicos até arquiteturas de aprendizado profundo (Angelopoulos; Bates, 2021). Em sua formulação padrão, o CP garante que os intervalos cubram a observação real com probabilidade de pelo menos $1 - \alpha$ sob a hipótese de permutabilidade (*exchangeability*) dos dados. Embora séries temporais violem essa hipótese de forma estrutural, pela impossibilidade de trocar o passado pelo futuro sem alterar a natureza do problema, o CP ainda pode produzir intervalos empiricamente calibrados quando o conjunto de calibração é construído de forma temporalmente coerente. Nesse caso, a validade dos intervalos deixa de ser uma garantia teórica e passa a ser verificada empiricamente por meio da cobertura observada no conjunto de teste, o que constitui a hipótese central deste trabalho (Xu; Xie, 2021).

A escolha do tema também foi motivada pelo contato prévio com problemas de previsão pluviométrica no contexto da LINCE, Liga de Inteligência Neurocomputacional na Engenharia do campus de Sorocaba, onde esse tipo de aplicação já vinha sendo discutido e estudado. Esse contato despertou o interesse em investigar o problema sob a perspectiva da quantificação de incerteza, combinando métodos de aprendizado de máquina com o arcabouço de *Conformal Prediction*. Nesse contexto, este trabalho propõe uma comparação sistemática entre quatro modelos preditivos de famílias distintas — *SARIMAX*, *Random Forest*, *LSTM* e *GRU* — combinados a três estratégias de *Conformal Prediction*: *Inductive Conformal Prediction* (ICP), *Mondrian Conformal Prediction* e *Conformalized Quantile Regression* (CQR). A avaliação é conduzida para três horizontes de previsão — curto, médio e longo prazo — com dados reais de precipitação diária da estação de Brasília, disponibilizados pelo Instituto Nacional de Meteorologia (INMET). Uma contribuição metodológica central do trabalho é o tratamento explícito da violação da hipótese de permutabilidade (*exchangeability*) que séries temporais impõem ao *Conformal Prediction*: em vez de assumir validade teórica formal, o trabalho adota uma perspectiva empírica, avaliando a utilidade dos intervalos pela cobertura observada no conjunto de teste. A contribuição do estudo é principalmente comparativa e aplicada — o objetivo não é propor um novo método ou uma nova variante de *Conformal Prediction*, mas investigar como diferentes combinações entre preditor base e estratégia conforme se comportam em termos de acurácia pontual, validade empírica dos intervalos e informatividade, em um problema pluviométrico real com múltiplos horizontes de previsão. Este trabalho está organizado da seguinte forma. O Capítulo 2 apresenta a revisão dos trabalhos relacionados, cobrindo os fundamentos teóricos do *Conformal Prediction* e suas principais variantes, os aspectos centrais da previsão de séries temporais com aprendizado de máquina, as arquiteturas LSTM e GRU, o uso de *Random Forest* em *forecasting* temporal, o modelo SARIMAX como *benchmark* estatístico e trabalhos recentes que integram CP e aprendizado de máquina. O Capítulo 3 discute a metodologia adotada, com ênfase na questão da *exchangeability*

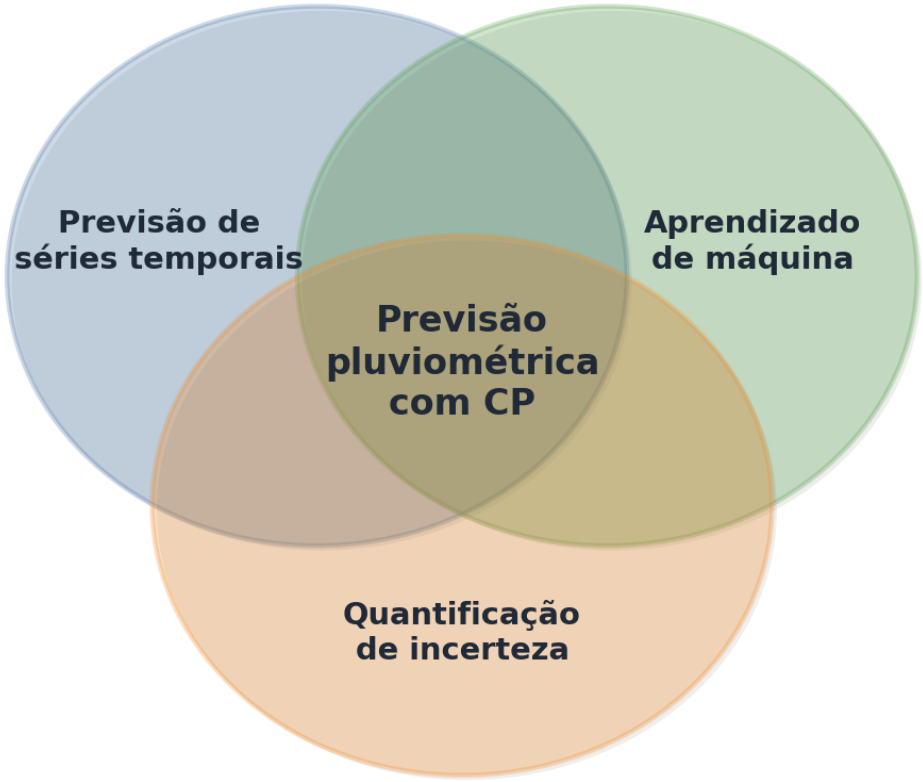
em séries temporais e na abordagem empírica utilizada para verificar a utilidade do CP nesse contexto. O Capítulo 4 descreve o *setup* experimental, o conjunto de dados meteorológicos da estação de Brasília disponibilizado pelo INMET, o *pipeline* de experimentos, as métricas de avaliação e os resultados obtidos para os horizontes de curto, médio e longo prazo, seguidos de uma discussão integrada dos achados. Por fim, o Capítulo 5 apresenta as conclusões do trabalho e as perspectivas para estudos futuros.

2 TRABALHOS RELACIONADOS

A quantificação de incerteza em previsão pluviométrica situa-se na interseção entre três frentes principais de pesquisa: (i) a teoria de *Conformal Prediction* (CP), voltada à construção de intervalos preditivos com garantias em amostras finitas; (ii) o uso de modelos de aprendizado de máquina na previsão de séries temporais hidrológicas e meteorológicas; e (iii) a integração recente entre CP, métodos de *machine learning* e arquiteturas de *deep learning* (Papacharalampous; Tyralis, 2022).

Nesse contexto, este capítulo revisa os fundamentos de *Conformal Prediction*, os principais aspectos da previsão de séries temporais com aprendizado de máquina, o emprego de arquiteturas recorrentes como LSTM e GRU, o uso de *Random Forest* em *forecasting* temporal e o papel do modelo SARIMAX como linha de base estatística clássica. Ao final, são discutidos trabalhos recentes que combinam modelos de aprendizado de máquina e CP. Como sintetiza a Figura 1, o tema deste trabalho se situa justamente na interseção entre previsão temporal, aprendizado de máquina e quantificação de incerteza.

Figura 1 – Interseção entre previsão de séries temporais, aprendizado de máquina e quantificação de incerteza no contexto da previsão pluviométrica



Legenda: Diagrama conceitual representando a interseção entre previsão de séries temporais, aprendizado de máquina e quantificação de incerteza, destacando a previsão pluviométrica com *Conformal Prediction* como foco do trabalho.

2.1 CONFORMAL PREDICTION

Conformal Prediction é um arcabouço estatístico para construção de conjuntos ou intervalos preditivos com garantias de cobertura em amostras finitas, sob a hipótese de permutabilidade (*exchangeability*) das observações (Vovk; Gammerman; Shafer, 2005). Em regressão, essa garantia pode ser expressa por

$$P\big(y_{n+1} \in \mathcal{C}_\alpha(x_{n+1})\big) \geq 1 - \alpha \tag{1}$$

em que $\alpha \in (0, 1)$ é o nível de significância, que controla o grau de conservadorismo do intervalo, e $1 - \alpha$ é a cobertura nominal desejada, ou seja, a fração mínima de observações futuras que se espera que o intervalo cubra. Na prática, valores típicos de α situam-se entre $0{,}05$ e $0{,}20$, correspondendo a coberturas nominais entre 80% e 95%. O termo $\mathcal{C}_\alpha(x_{n+1})$ representa o intervalo preditivo construído para uma nova entrada x_{n+1}, e y_{n+1} é o valor real a ser observado (Vovk; Gammerman; Shafer, 2005).

A lógica central da CP consiste em medir o quanto uma observação é não conforme em relação ao comportamento observado anteriormente. Para isso, define-se um escore de não conformidade. Em regressão, uma escolha simples e amplamente utilizada é o resíduo absoluto:

$$s_i = |y_i - \hat{f}(x_i)| \tag{2}$$

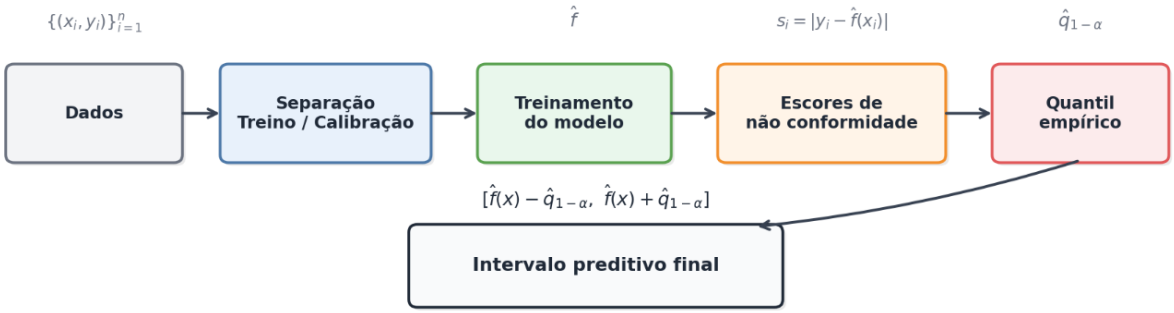
onde $\hat{f}$ é o modelo ajustado (Angelopoulos; Bates, 2021). Esses escores são então calibrados empiricamente.

Na variante mais usada em aplicações práticas, denominada *split conformal prediction* ou *inductive conformal prediction* (ICP), os dados são divididos em treino e calibração. O modelo $\hat{f}$ é ajustado no conjunto de treino, e os escores $s_1, \ldots, s_m$ são calculados no conjunto de calibração. A partir deles, estima-se um quantil empírico $\hat{q}_{1-\alpha}$, produzindo o intervalo

$$\mathcal{C}_\alpha(x_{n+1}) = \Big[\hat{f}(x_{n+1}) - \hat{q}_{1-\alpha},\ \hat{f}(x_{n+1}) + \hat{q}_{1-\alpha}\Big] \tag{3}$$

Essa formulação é particularmente atrativa por sua simplicidade computacional e por ser agnóstica em relação ao modelo de base (Shafer; Vovk, 2008). A sequência de etapas desse procedimento é resumida na Figura 2.

Figura 2 – Fluxo geral do *split conformal prediction*



Legenda: Fluxograma ilustrando as principais etapas do *split conformal prediction*, incluindo a separação dos dados em treino e calibração, o ajuste do modelo base, o cálculo dos escores de não conformidade e a obtenção do intervalo preditivo final por meio de um quantil empírico.

Embora robusta, a CP global pode gerar intervalos pouco adaptativos em cenários heteroscedásticos, isto é, quando a variabilidade do erro não é constante no espaço de entrada (Romano; Patterson; Candès, 2019). Nesses casos, a utilização de um único quantil global pode tornar os intervalos excessivamente largos em regiões mais estáveis e insuficientemente informativos em outras.

Uma alternativa é a *Mondrian Conformal Prediction*, em que a calibração é realizada separadamente em grupos ou regimes (Boström; Johansson; Löfström, 2021). Se $g(x)$ identifica o grupo de uma observação, então o intervalo pode ser expresso como

$$\mathcal{C}_\alpha(x_{n+1}) = \left[\hat{f}(x_{n+1}) - \hat{q}_{1-\alpha}^{(g(x_{n+1}))},\ \hat{f}(x_{n+1}) + \hat{q}_{1-\alpha}^{(g(x_{n+1}))}\right] \tag{4}$$

Essa ideia é especialmente relevante quando há regimes distintos, como períodos secos e chuvosos.

Outra extensão importante é a *Conformalized Quantile Regression* (CQR), proposta por Romano, Patterson e Candès (Romano; Patterson; Candès, 2019). Nesse caso, o modelo base estima quantis inferior e superior, $\hat{q}_{lo}(x)$ e $\hat{q}_{hi}(x)$, e o escore de não conformidade é dado por

$$s_i = \max\{\hat{q}_{lo}(x_i) - y_i,\ y_i - \hat{q}_{hi}(x_i)\} \tag{5}$$

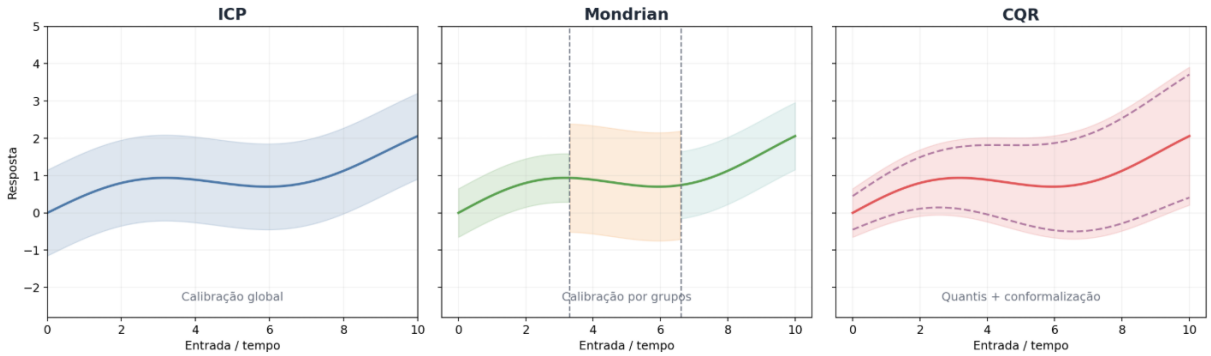
O intervalo final passa então a ser

$$\mathcal{C}_\alpha(x_{n+1}) = \left[\hat{q}_{lo}(x_{n+1}) - \hat{\delta}_{1-\alpha},\ \hat{q}_{hi}(x_{n+1}) + \hat{\delta}_{1-\alpha}\right] \tag{6}$$

onde $\hat{\delta}_{1-\alpha}$ é obtido a partir dos escores calculados no conjunto de calibração (Romano; Patterson; Candès, 2019). A principal vantagem do CQR é permitir que a largura inicial do intervalo já reflita a incerteza condicional aprendida pelo modelo. Como mostra a Figura 3, ICP, Mondrian e CQR diferem principalmente na forma como a calibração é realizada e na adaptabilidade

resultante dos intervalos.

Figura 3 – Comparação conceitual entre ICP, Mondrian e CQR



Legenda: Comparação entre três estratégias de *Conformal Prediction*. No ICP, a calibração é realizada globalmente com um único quantil; no Mondrian, a calibração é feita separadamente por grupos ou regimes; no CQR, limites quantílicos estimados pelo modelo são posteriormente ajustados por uma etapa de conformalização.

Em séries pluviométricas, nas quais a variabilidade muda ao longo do tempo e entre regimes de chuva, essas extensões são especialmente relevantes. Por essa razão, ICP, Mondrian e CQR constituem estratégias adequadas para investigar o compromisso entre validade estatística e informatividade dos intervalos.

2.2 PREDIÇÃO DE SÉRIES TEMPORAIS USANDO APRENDIZADO DE MÁQUINA

A previsão de séries temporais consiste em estimar um valor futuro y_{t+h} a partir de observações passadas da própria série e, eventualmente, de covariáveis exógenas (Hyndman; Athanasopoulos, 2021). Em termos gerais,

$$y_{t+h} = f(y_t, y_{t-1}, \ldots, y_{t-p}, x_t, x_{t-1}, \ldots) + \varepsilon_t \tag{7}$$

onde h representa o horizonte de previsão (Bontempi; Taieb; Borgne, 2013).

No contexto do aprendizado de máquina, é comum reformular o problema temporal como uma tarefa supervisionada, construindo vetores de entrada a partir de janelas passadas:

$$X_t = [y_{t-p+1}, \ldots, y_t,\ x_{t-p+1}, \ldots, x_t] \tag{8}$$

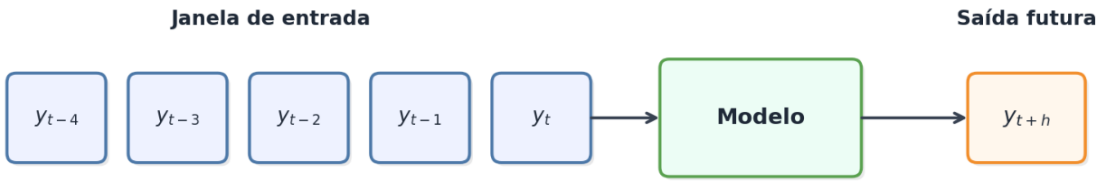
Esse vetor é então associado ao alvo futuro y_{t+h} (Bontempi; Taieb; Borgne, 2013). Essa reformulação permite a aplicação de modelos gerais de regressão, desde que a separação entre treino, validação e teste preserve a ordem temporal.

Para previsão multi-passo, a literatura destaca três estratégias principais (Bontempi; Taieb; Borgne, 2013): (i) estratégia recursiva, em que um modelo de um passo à frente é aplicado iterativamente; (ii) estratégia direta, em que se treina um modelo específico para cada horizonte;

(iii) estratégia de saída múltipla, em que uma única arquitetura produz simultaneamente vários valores futuros.

Em aplicações hidrológicas e meteorológicas, o uso de aprendizado de máquina tem crescido devido à sua capacidade de capturar relações não lineares entre precipitação, temperatura, umidade, radiação e demais variáveis ambientais (Tyralis; Papacharalampous, 2022). No caso da precipitação, o problema é agravado por assimetria, heterocedasticidade, grande frequência de valores nulos e ocorrência esporádica de extremos, o que reforça a importância de combinar acurácia pontual e quantificação de incerteza. A transformação da série temporal em um problema supervisionado é ilustrada na Figura 4.

Figura 4 – Transformação de uma série temporal em problema supervisionado



Legenda: Esquema ilustrando a transformação de uma série temporal em um problema de aprendizagem supervisionada, no qual uma janela de observações passadas é utilizada como entrada do modelo para estimar um valor futuro em um horizonte de previsão h.

2.3 REDES NEURAIS RECORRENTES, LSTM E GRU

Redes neurais recorrentes (*Recurrent Neural Networks* – RNN) são arquiteturas desenvolvidas para modelar dados sequenciais por meio de um estado oculto que evolui ao longo do tempo (Hewamalage; Bergmeir; Bandara, 2021). Em sua forma básica,

$$h_t = \phi(W_x x_t + W_h h_{t-1} + b) \tag{9}$$

$$\hat{y}_t = \psi(W_y h_t + c) \tag{10}$$

Nessas expressões, x_t representa a entrada no instante t, h_t é o estado oculto atualizado recursivamente, e $\hat{y}_t$ é a saída prevista. As matrizes W_x, W_h e W_y correspondem aos pesos aprendidos pela rede, enquanto b e c são termos de viés. Essa estrutura permite modelar dependência temporal ao combinar a informação atual com o histórico resumido em h_{t-1}. Entretanto, RNNs simples sofrem com o problema de desaparecimento de gradientes, o que dificulta o aprendizado de dependências de longo prazo (Hochreiter; Schmidhuber, 1997).

A arquitetura LSTM (*Long Short-Term Memory*) introduz uma célula de memória e portas de controle, favorecendo o aprendizado de dependências mais longas (Hochreiter; Schmidhuber,

1997). De forma resumida,

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \qquad i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{11}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \qquad h_t = o_t \odot \tanh(c_t) \tag{12}$$

Nessas equações, f_t é a porta de esquecimento, que controla quanta informação da memória anterior c_{t-1} deve ser mantida; i_t é a porta de entrada, que regula a incorporação de nova informação; c_t é o novo estado da célula de memória; e h_t é o estado oculto de saída. O símbolo σ representa a função sigmoide, que produz valores entre 0 e 1, e $\odot$ indica multiplicação elemento a elemento. Essa formulação tornou a LSTM uma das arquiteturas recorrentes mais utilizadas em *forecasting* temporal (Lim; Zohren, 2021).

A GRU (*Gated Recurrent Unit*) é uma alternativa mais compacta, com menor número de parâmetros (Cho et al., 2014). Sua dinâmica pode ser resumida por

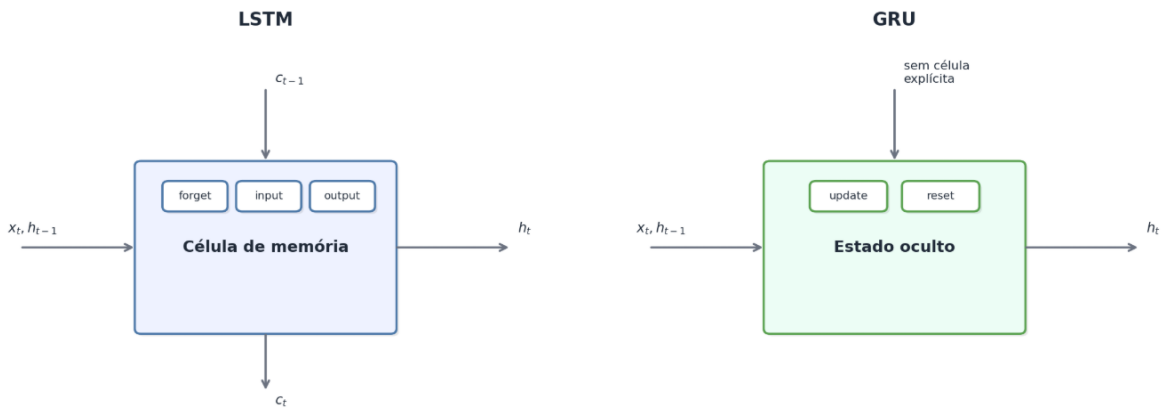
$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z), \qquad r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \tag{13}$$

$$\tilde{h}_t = \tanh(W_h[r_t \odot h_{t-1}, x_t] + b_h), \qquad h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{14}$$

Nesse caso, z_t é a porta de atualização, que controla quanto do estado anterior deve ser preservado, enquanto r_t é a porta de reinicialização, responsável por regular a influência do passado na nova informação candidata $\tilde{h}_t$. O novo estado oculto h_t resulta de uma combinação entre o estado anterior e a informação atualizada. Por sua simplicidade relativa, a GRU costuma oferecer bom compromisso entre custo computacional e desempenho.

Na hidrologia, LSTM e GRU vêm sendo utilizadas em tarefas como transformação chuva–vazão e previsão de vazão, com resultados competitivos em relação a abordagens tradicionais (Kratzert et al., 2018). Neste trabalho, ambas são utilizadas como representantes de arquiteturas recorrentes modernas para previsão pluviométrica. Como mostra a Figura 5, embora LSTM e GRU compartilhem a ideia de controlar o fluxo de informação ao longo do tempo, diferem na complexidade estrutural de suas células.

Figura 5 – Representação esquemática das arquiteturas LSTM e GRU



Legenda: Comparação esquemática entre as arquiteturas recorrentes LSTM e GRU. A LSTM utiliza uma célula de memória explícita e múltiplas portas de controle para regular o fluxo de informação, enquanto a GRU adota uma estrutura mais compacta, com menor número de parâmetros e sem estado de memória separado.

2.4 RANDOM FOREST E SEU USO NA PREVISÃO DE SÉRIES TEMPORAIS

A *Random Forest* (RF) é um método baseado na agregação de múltiplas árvores de decisão ajustadas sobre amostras *bootstrap* (Breiman, 2001). Em regressão, a previsão final é dada por

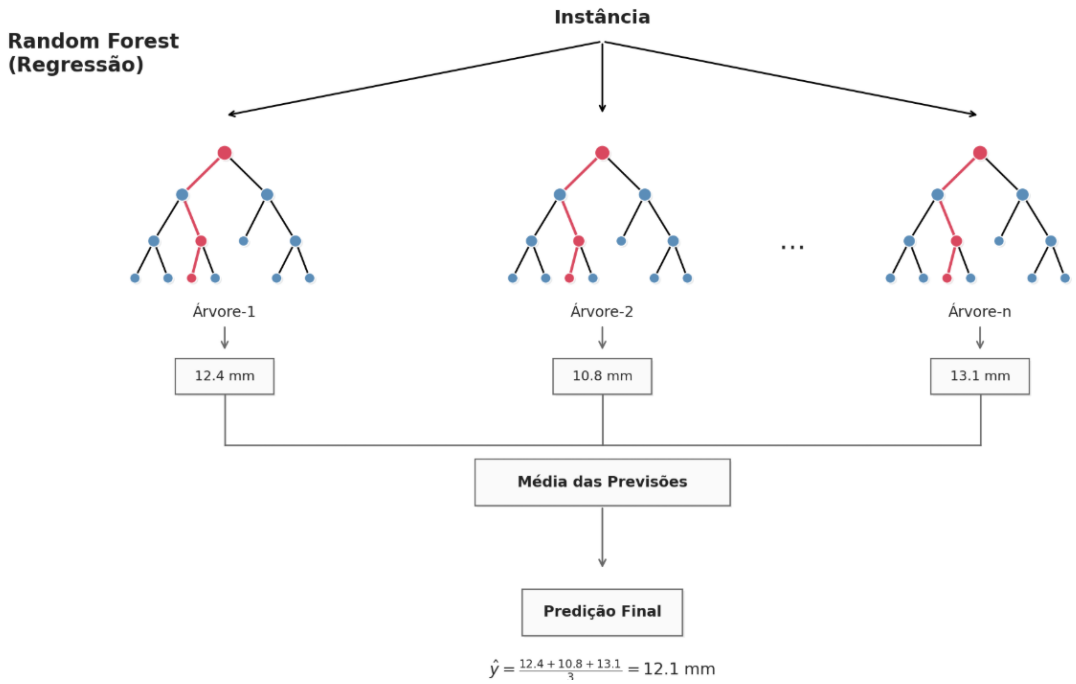
$$\hat{f}_{RF}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{15}$$

onde $T_b(x)$ é a previsão da b-ésima árvore (Breiman, 2001). Essa agregação reduz variância e aumenta robustez.

Em séries temporais, a RF é aplicada após a transformação da série em um problema supervisionado, utilizando defasagens e covariáveis como atributos (Goehry et al., 2023). O método é particularmente interessante quando há relações não lineares e interações complexas entre variáveis (Tyralis; Papacharalampous; Langousis, 2019).

Uma limitação importante é que a RF padrão fornece apenas previsões pontuais. Uma extensão relevante é a *Quantile Regression Forest* (QRF), proposta por Meinshausen (Meinshausen, 2006), que permite estimar quantis condicionais. Ainda assim, a obtenção de intervalos com garantias formais de cobertura continua sendo desafiadora, o que motiva sua combinação com CP (Angelopoulos; Bates, 2021). A lógica de agregação do método é resumida na Figura 6.

Figura 6 – Esquema conceitual do modelo *Random Forest* em regressão



Legenda: Representação esquemática do funcionamento da *Random Forest* em regressão. Múltiplas árvores são ajustadas a partir de amostras *bootstrap* dos dados e subconjuntos aleatórios de atributos, produzindo previsões individuais que são posteriormente agregadas por média para formar a saída final do modelo.

Neste trabalho, a RF é empregada como linha de base não linear baseada em *ensemble*, servindo como contraponto às arquiteturas recorrentes e ao *benchmark* estatístico clássico.

2.5 MODELO SARIMAX

Os modelos ARIMA constituem uma classe clássica de métodos para previsão temporal, combinando componentes autoregressivos, de diferenciação e de médias móveis (Box et al., 2015). Sua extensão sazonal, SARIMA, pode ser representada genericamente por

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t \tag{16}$$

onde B é o operador defasagem e s é o período sazonal (Box et al., 2015).

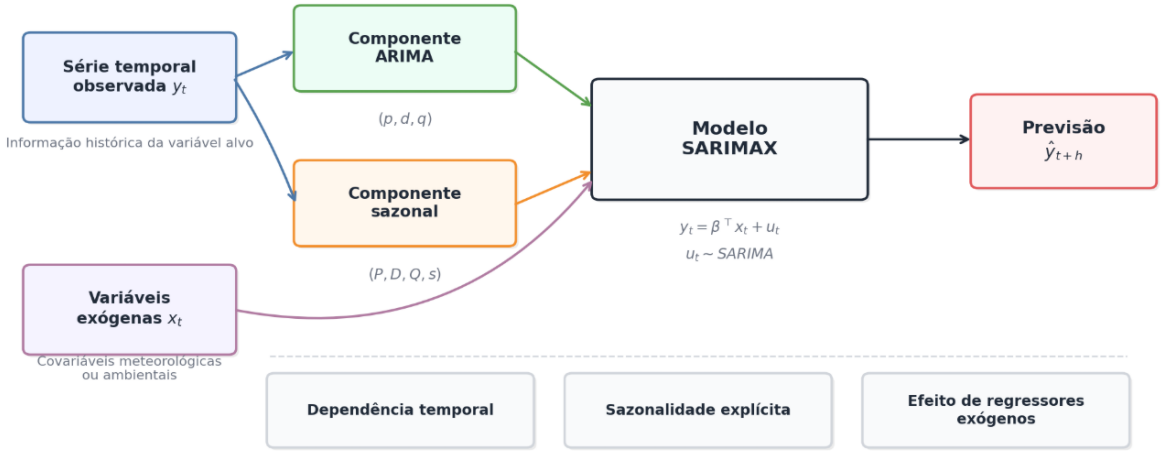
O modelo SARIMAX amplia essa estrutura ao incorporar regressores exógenos:

$$y_t = \beta^\top x_t + u_t \tag{17}$$

onde u_t segue uma dinâmica SARIMA (Alharbi; Csala, 2022). Em implementações modernas, como a classe `SARIMAX` do `statsmodels`, essa formulação é tratada no contexto de modelos em espaço de estados (Statsmodels Development Team, 2024).

SARIMAX é particularmente útil quando a série apresenta componentes lineares e sazonais bem definidos, além da possível influência de covariáveis externas (Mulla; Pande; Singh, 2024). Entretanto, trata-se de um modelo fundamentalmente linear, o que pode limitar sua capacidade de representar relações altamente não lineares entre precipitação e variáveis atmosféricas. Por essa razão, neste trabalho ele é adotado como *benchmark* estatístico clássico. Como ilustra a Figura 7, a previsão final decorre da combinação entre estrutura autoregressiva, componente sazonal e regressores exógenos.

Figura 7 – Esquema conceitual do modelo SARIMAX



Legenda: Representação esquemática do modelo SARIMAX, destacando a combinação entre componente ARIMA, componente sazonal e variáveis exógenas para geração da previsão final.

2.6 TRABALHOS RECENTES QUE COMBINAM APRENDIZADO DE MÁQUINA E CONFORMAL PREDICTION (2020–2025)

Entre 2020 e 2025, observa-se um crescimento do uso de *Conformal Prediction* como camada de quantificação de incerteza para modelos de aprendizado de máquina aplicados a séries temporais (Stankeviciūtė; Alaa; Schaar, 2021). Em termos gerais, essa linha de pesquisa busca preservar a capacidade preditiva do modelo de base e acrescentar intervalos calibrados por meio de uma etapa conforme (Romano; Patterson; Candès, 2019).

No contexto de séries temporais, Stankeviciute et al. (Stankeviciūtė; Alaa; Schaar, 2021) discutem estratégias conformes para previsão multi-horizonte em redes recorrentes. Em outra direção, Xu e Xie (Xu; Xie, 2021) propõem o EnbPI, baseado em *ensembles bootstrap* para construção sequencial de intervalos em séries dinâmicas. Em arquiteturas mais recentes, Auer et al. (Auer et al., 2023) exploram o uso de CP com *Modern Hopfield Networks*, enquanto Lee, Xu e Xie (Lee; Xu; Xie, 2024) tratam da integração entre *transformers* e *Conformal Prediction* para *forecasting* temporal.

Na linha de regressão quantílica, Romano et al. (Romano; Patterson; Candès, 2019) introduzem o CQR, que se tornou uma das principais referências para construção de intervalos

adaptativos com validade marginal. A partir dessa ideia, Jensen, Bianchi e Anfinsen (Jensen; Bianchi; Anfinsen, 2022) propõem uma estratégia baseada em *ensemble* e regressão quantílica, buscando intervalos mais informativos em cenários heterocedásticos. De forma semelhante, Hu et al. (Hu et al., 2022) aplicam modelos quantílicos com conformalização à previsão intervalar em energia eólica.

Em aplicações hídricas e meteorológicas, Iwakin e Moazeni (Iwakin; Moazeni, 2024) exploram *Conformal Prediction* em um modelo híbrido para previsão de demanda hídrica urbana. No campo da previsão meteorológica, Mortier et al. (Mortier et al., 2025) ilustram o interesse recente pelo uso de variantes conformes em modelos de *weather forecasting*. Em conjunto, esses trabalhos indicam que CP vem sendo empregada como uma camada modelo-agnóstica de quantificação de incerteza sobre preditores avançados (Angelopoulos; Bates, 2021).

Nesse cenário, o presente trabalho se insere ao comparar, de forma sistemática, diferentes modelos de previsão pluviométrica, *Random Forest*, LSTM, GRU e SARIMAX, combinados a estratégias como ICP, Mondrian e CQR, em múltiplos horizontes de previsão. Assim, a contribuição do estudo é principalmente comparativa e aplicada, concentrando-se na relação entre acurácia pontual, validade dos intervalos e comportamento dos modelos em um problema pluviométrico real.

3 METODOLOGIA

Este capítulo apresenta a abordagem metodológica adotada para investigar o uso de *Conformal Prediction* na quantificação de incerteza em previsões pluviométricas. Partindo dos fundamentos teóricos revisados no Capítulo 2, discute-se inicialmente uma questão central que distingue este trabalho de aplicações clássicas do método: a violação da hipótese de permutabilidade (*exchangeability*) imposta pela natureza temporal dos dados. Em seguida, apresenta-se a abordagem geral adotada para contornar essa limitação de forma metodologicamente coerente, descrevendo como a validade dos intervalos preditivos é verificada empiricamente em vez de assumida por construção teórica. A compreensão dessa distinção é fundamental para a correta interpretação dos resultados apresentados no Capítulo 4.

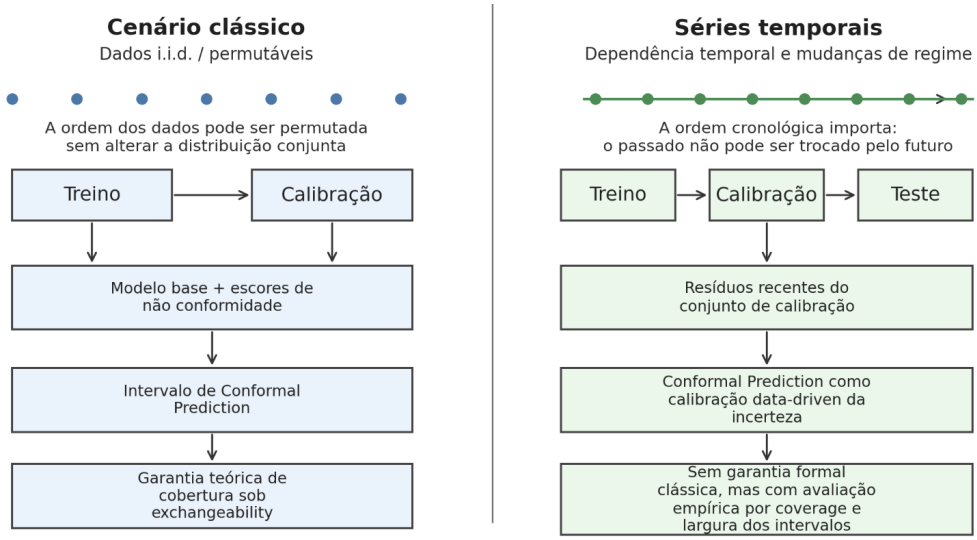
3.1 CONFORMAL PREDICTION EM SÉRIES TEMPORAIS: ENTRE A GARANTIA TEÓRICA E A VALIDADE EMPÍRICA

A aplicação de *Conformal Prediction* a séries temporais exige uma discussão preliminar importante, pois o arcabouço teórico clássico do método repousa sobre uma hipótese que séries temporais, por natureza, violam. Como apresentado no Capítulo 2, a garantia formal de cobertura do CP é estabelecida sob a hipótese de permutabilidade (*exchangeability*) das observações, que pressupõe que a ordem dos dados não altera sua distribuição conjunta (Shafer; Vovk, 2008). Em termos práticos, isso significa que qualquer permutação entre os pontos da amostra deveria ser igualmente plausível.

Séries temporais violam essa hipótese de forma estrutural. A ordem temporal das observações é essencial: o passado não pode ser trocado pelo futuro sem alterar fundamentalmente a natureza do problema. Além disso, séries como a precipitação diária apresentam dependência temporal, mudanças de regime entre períodos secos e chuvosos e não estacionariedade, características que tornam os dados intrinsecamente não permutáveis (Stankeviciūtė; Alaa; Schaar, 2021). Diante disso, uma questão central se impõe: é possível aplicar *Conformal Prediction* a séries temporais de forma útil, mesmo sem a garantia formal que a teoria estabelece?

A Figura 8 ilustra essa questão. Enquanto, no caso ideal, o *Conformal Prediction* se apoia na hipótese de *exchangeability*, em séries temporais essa hipótese é quebrada pela própria estrutura cronológica dos dados. Ainda assim, os resíduos obtidos em um conjunto de calibração temporalmente coerente podem ser utilizados como base empírica para calibrar a incerteza das previsões futuras.

Figura 8 – Uso de *Conformal Prediction* em séries temporais sem *exchangeability*



Legenda: Representação conceitual da diferença entre o cenário clássico de *Conformal Prediction*, baseado em *exchangeability*, e o caso de séries temporais, em que a ordem cronológica impede a permutabilidade estrita das observações. Ainda assim, os erros observados em um conjunto de calibração temporalmente coerente podem ser utilizados para calibrar empiricamente a incerteza.

A hipótese central deste trabalho é que sim. O argumento não é que a garantia teórica se mantém, pois ela formalmente não se mantém na ausência de *exchangeability*. O argumento é que, mesmo sem essa garantia, o *Conformal Prediction* continua sendo capaz de produzir intervalos empiricamente calibrados, desde que o conjunto de calibração seja construído de forma temporalmente coerente e que os resíduos nele observados sejam suficientemente representativos do comportamento futuro do modelo. Nesse sentido, o CP passa a ser tratado não como um método com garantia formal de cobertura, mas como um método *data-driven* de calibração de incerteza, cuja validade é de natureza empírica e não teórica (Angelopoulos; Bates, 2021).

Essa distinção é fundamental para a interpretação dos resultados deste trabalho. Em vez de assumir que o *Conformal Prediction* é válido por construção, o que seria incorreto no contexto de séries temporais, este trabalho mede o potencial de validade do método por meio da cobertura empírica observada no conjunto de teste. A cobertura empírica funciona, assim, como evidência prática em substituição à garantia teórica: se os intervalos cobrem aproximadamente $1 - \alpha$ das observações avaliadas, isso indica que o método foi capaz de capturar, a partir dos resíduos de calibração, uma distribuição empírica dos erros suficientemente representativa para produzir intervalos úteis (Xu; Xie, 2021). Essa validação empírica não substitui a prova formal, mas oferece uma base interpretável e metodologicamente coerente para o uso do CP em problemas temporais reais. Vale notar, contudo, que a violação de *exchangeability* em séries temporais não é necessariamente homogênea ao longo do tempo. Em séries com sazonalidade marcada, como a precipitação diária de Brasília, é possível argumentar que observações pertencentes ao mesmo regime climático, por exemplo, dois dias distintos dentro da estação chuvosa , são mais permutáveis entre si do que observações de regimes opostos. Essa ideia sugere a existência de

estruturas de *quasi-exchangeability* condicional ao regime ou à estação do ano, que poderiam ser exploradas por abordagens conformes condicionais ou segmentadas (Boström; Johansson; Löfström, 2021). Embora esse refinamento esteja além do escopo deste trabalho, ele representa uma direção conceitualmente interessante para investigações futuras sobre o uso de *Conformal Prediction* em séries climáticas.

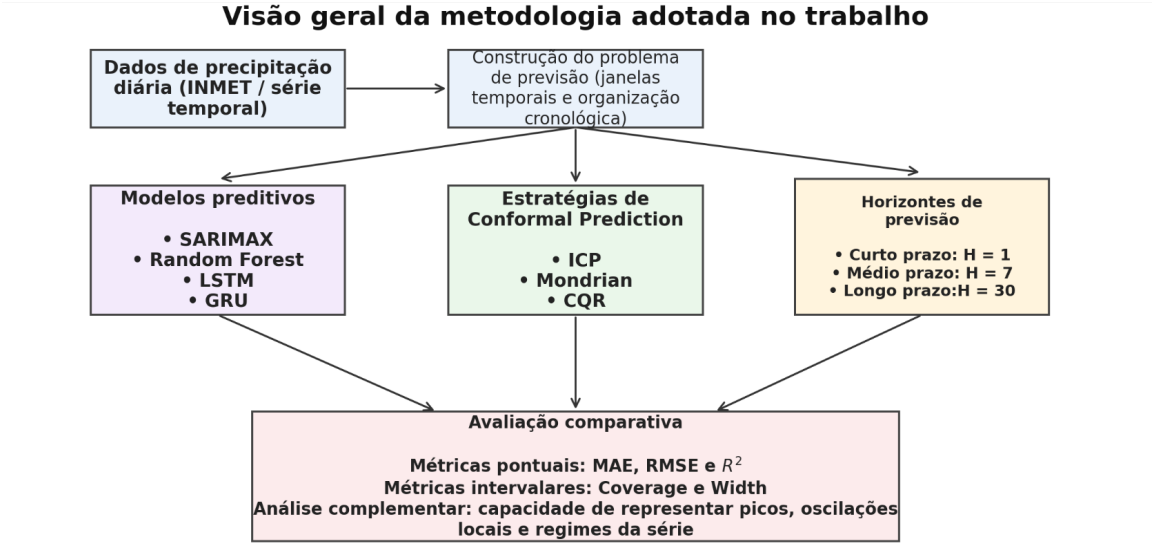
3.2 ABORDAGEM METODOLÓGICA GERAL

A metodologia deste trabalho foi estruturada em torno de uma comparação sistemática entre diferentes modelos preditivos e diferentes estratégias de *Conformal Prediction*, aplicados a um problema real de previsão pluviométrica diária. O objetivo central não foi propor um novo método, mas investigar, de forma controlada, como diferentes combinações entre preditor base e estratégia conforme se comportam em termos de acurácia pontual, cobertura empírica e informatividade dos intervalos, em múltiplos horizontes de previsão.

Para isso, foram considerados modelos preditivos pertencentes a famílias distintas, de modo a incluir tanto abordagens estatísticas clássicas quanto métodos de aprendizado de máquina e arquiteturas neurais recorrentes. Esses modelos foram combinados a três estratégias de *Conformal Prediction*: *Inductive Conformal Prediction* (ICP), *Mondrian Conformal Prediction* e *Conformalized Quantile Regression* (CQR). A comparação foi realizada em três horizontes de previsão, correspondentes a curto, médio e longo prazo, com o objetivo de avaliar como a utilidade empírica dos intervalos conformes se altera à medida que aumenta a dificuldade do problema preditivo.

A estrutura geral dessa abordagem é resumida na Figura 9. Como pode ser observado, o estudo foi organizado a partir da combinação entre diferentes modelos base, diferentes estratégias conformes e diferentes horizontes de previsão, sendo a comparação final realizada com base em métricas pontuais e intervalares.

Figura 9 – Visão geral da metodologia adotada no trabalho



Legenda: Estrutura geral da metodologia empregada no trabalho. Diferentes modelos preditivos foram combinados a diferentes estratégias de *Conformal Prediction* e avaliados em múltiplos horizontes de previsão, permitindo analisar o comportamento dos intervalos em termos de acurácia pontual, cobertura empírica e informatividade.

Essa escolha metodológica foi motivada pelo interesse em analisar, dentro de um mesmo problema, diferentes níveis de complexidade tanto no modelo base quanto no método conforme. Em particular, buscou-se verificar se estratégias conformes mais sofisticadas necessariamente produzem melhores resultados ou se, em alguns cenários, métodos mais simples podem ser igualmente ou até mais eficientes quando combinados a preditores base adequados. Essa questão é especialmente relevante no contexto de séries temporais, em que a validade dos intervalos depende menos de garantias formais e mais da qualidade empírica da calibração.

Sob essa perspectiva, a avaliação adotada neste trabalho não se limita à verificação da cobertura empírica isoladamente. Como discutido ao longo do estudo, a utilidade prática dos intervalos preditivos depende também de sua largura e de sua capacidade de representar adequadamente a dinâmica da série, incluindo picos, oscilações locais e mudanças de regime. Dessa forma, a metodologia foi concebida para permitir uma análise comparativa que contemplasse não apenas a validade empírica dos intervalos, mas também o tipo de compromisso que cada combinação estabelece entre cobertura, nitidez e adaptatividade.

Os detalhes relativos ao conjunto de dados utilizado, ao pré-processamento, à construção do problema supervisionado, ao particionamento temporal, à configuração dos modelos, à implementação das estratégias conformes e às métricas de avaliação são apresentados no capítulo seguinte, dedicado ao experimento.

4 EXPERIMENTO

Este capítulo descreve o desenho experimental utilizado para comparar as diferentes combinações entre modelos preditivos e estratégias de *Conformal Prediction* propostas no Capítulo 3. São apresentados o conjunto de dados utilizado, o *pipeline* de pré-processamento e transformação da série temporal em problema supervisionado, o particionamento cronológico dos dados, a configuração detalhada de cada modelo e de cada método conforme, e as métricas de avaliação adotadas para medir tanto a qualidade das previsões pontuais quanto a qualidade dos intervalos preditivos. Por fim, são apresentados e discutidos os resultados obtidos para os três horizontes de previsão avaliados. A leitura deste capítulo pressupõe familiaridade com os conceitos introduzidos nos Capítulos 2 e 3.

4.1 SETUP EXPERIMENTAL

Os experimentos foram realizados em ambiente local, por meio de *notebooks* Jupyter (`.ipynb`) executados no *Visual Studio Code* (VS Code), em sistema operacional *Windows*, com uso de um ambiente virtual dedicado (`.venv`) baseado em *Python* 3.11.9. A máquina utilizada contava com processador *Intel Core i5-14600K* e 32 GB de memória RAM DDR5. Não foi empregada unidade de processamento gráfico (*GPU*); assim, todas as etapas de treinamento, inferência e avaliação foram conduzidas exclusivamente em CPU.

Ao longo dos experimentos, foram utilizadas bibliotecas voltadas à manipulação de dados, modelagem, avaliação e visualização. As bibliotecas *NumPy* e *Pandas* foram empregadas na organização, transformação e processamento dos dados. A geração de gráficos e a inspeção visual dos resultados foram realizadas com *Matplotlib*. A biblioteca *scikit-learn* foi utilizada nas etapas de pré-processamento, cálculo de métricas e implementação do modelo *Random Forest*. Para os modelos neurais recorrentes, utilizou-se *TensorFlow/Keras*, enquanto o modelo *SARIMAX* foi implementado com *statsmodels*. Também foram empregadas bibliotecas auxiliares, como *Pathlib* e *warnings*, principalmente para organização do fluxo de arquivos e controle de mensagens durante a execução.

Com o intuito de favorecer a reprodutibilidade, adotou-se semente aleatória igual a 42 em todos os componentes estocásticos do *pipeline* onde essa configuração era aplicável: inicialização dos pesos das redes neurais, amostragem *bootstrap* do *Random Forest* e embaralhamento interno de lotes durante o treinamento. A semente foi fixada simultaneamente em *NumPy*, *Python* nativo e *TensorFlow* para garantir consistência entre execuções. Além disso, os experimentos foram organizados de forma modular, com *notebooks* independentes para cada família de modelo. Essa separação facilitou o controle do *pipeline* e tornou mais clara a distinção entre treinamento, geração de previsões e aplicação dos métodos de *Conformal Prediction*.

Os modelos neurais recorrentes foram construídos com níveis de complexidade comparáveis,

de modo a permitir uma comparação mais equilibrada entre as arquiteturas *LSTM* e *GRU*. Em ambos os casos, a arquitetura foi composta por uma camada recorrente com 64 unidades, seguida de uma camada de *Dropout* com taxa de 0,2 e de uma camada densa intermediária com 32 neurônios e ativação *ReLU*. O treinamento foi realizado com o otimizador *Adam*, taxa de aprendizado de 10^{-3}, número máximo de 30 *epochs*, *batch size* de 128 e mecanismo de *Early Stopping* com paciência de 5 épocas e restauração dos melhores pesos (`restore_best_weights=True`), monitorando a perda no conjunto de validação (`val_loss`). Na prática, o *Early Stopping* interrompeu o treinamento antecipadamente na maioria dos casos antes de atingir o limite de 30 épocas, reduzindo o tempo de treinamento e evitando sobreajuste.

No caso do modelo *LSTM*, a versão voltada à previsão pontual utilizou uma camada de saída com um único neurônio e ativação *softplus*. Essa escolha foi adotada para manter previsões estritamente não negativas, compatível com a natureza física da variável alvo. A função de perda utilizada foi a *Tweedie deviance*, com parâmetro $p = 1{,}5$, por sua adequação a distribuições assimétricas com massa de probabilidade em zero, como a precipitação diária. A variável alvo foi mantida em escala original de milímetros durante o treinamento com *Tweedie*. Já na versão quantílica empregada pelo método *Conformalized Quantile Regression* (CQR), a mesma estrutura base foi mantida, substituindo-se a saída pontual por duas saídas densas com ativação linear, responsáveis pelos quantis $\tau_{lo} = 0{,}10$ e $\tau_{hi} = 0{,}90$, treinadas com *pinball loss*. Nessa versão, a variável alvo foi transformada por $\log(1 + y)$ antes do treinamento, e as saídas foram reconvertidas por $\exp(\cdot) - 1$ após a inferência.

De forma semelhante, o modelo *GRU* foi construído substituindo-se apenas a célula recorrente *LSTM* por uma célula *GRU*, também com 64 unidades. Na versão pontual, adotou-se uma camada de saída linear com um único neurônio, treinada com erro quadrático médio (*Mean Squared Error* – MSE), com a variável alvo transformada por $\log(1 + y)$ antes do treinamento e reconvertida por $\exp(\cdot) - 1$ após a inferência. Para a versão quantílica associada ao CQR, foram novamente utilizadas duas saídas treinadas com *pinball loss*, preservando-se a mesma estrutura intermediária e a mesma transformação logarítmica da variável alvo. É importante registrar que as diferenças entre as configurações de *LSTM* e *GRU*, especialmente a função de perda e a ativação de saída na versão pontual, foram escolhas pragmáticas feitas ao longo do desenvolvimento e reduzem a comparabilidade direta entre as duas arquiteturas, conforme discutido na Seção 4.8.

Para o modelo *Random Forest*, foi utilizada a implementação `RandomForestRegressor` da biblioteca *scikit-learn*. A configuração adotada incluiu 400 árvores (`n_estimators = 400`), profundidade máxima não limitada (`max_depth = None`), mínimo de 2 amostras por folha (`min_samples_leaf = 2`), *bootstrap* habilitado (`bootstrap = True`), paralelização com uso de todos os núcleos disponíveis do processador (`n_jobs = -1`) e semente aleatória fixada em 42 (`random_state = 42`). Essa configuração foi mantida como base tanto para as previsões pontuais quanto para a etapa posterior de construção dos intervalos conformes. Ao contrário dos modelos neurais, o *Random Forest* não requer uma etapa de validação explícita durante o treinamento. O mecanismo de *bootstrap* habilitado por `bootstrap=True`

introduz uma forma implícita de avaliação por meio das amostras não selecionadas em cada reamostragem, denominadas *out-of-bag* (OOB) (Breiman, 2001). Neste experimento, o escore OOB não foi monitorado formalmente; a avaliação do modelo foi conduzida exclusivamente sobre os conjuntos de calibração e teste, em consonância com o particionamento temporal adotado para os demais modelos. A variável alvo foi mantida em escala original de milímetros, sem transformação logarítmica, e as variáveis de entrada também não foram padronizadas, pois o *Random Forest* opera com partições baseadas em limiares que são invariantes a transformações monotônicas das variáveis.

O modelo *SARIMAX*, por sua vez, foi implementado por meio da classe `SARIMAX` da biblioteca *statsmodels*. A especificação adotada foi $(p, d, q) = (1, 0, 1)$ para a componente não sazonal e $(P, D, Q, s) = (1, 0, 1, 7)$ para a componente sazonal, com período sazonal $s = 7$ correspondente ao ciclo semanal identificado na série. Também foram desabilitadas as restrições de estacionariedade e invertibilidade, por meio dos parâmetros `enforce_stationarity = False` e `enforce_invertibility = False`, de forma a permitir maior flexibilidade no ajuste do modelo à série observada. Os parâmetros foram estimados por máxima verossimilhança via algoritmo de espaço de estados, utilizando os dados do conjunto de treinamento. Durante os períodos de calibração e teste, o modelo foi atualizado sequencialmente por meio do método `extend`, que incorpora cada nova observação ao estado interno do modelo sem reestimar os parâmetros. Esse modo de operação ponto a ponto tornou o *SARIMAX* o modelo com maior tempo total de execução entre os avaliados, conforme discutido ao final desta seção.

No que se refere aos métodos de *Conformal Prediction*, foram avaliadas três estratégias principais, cada uma com seus próprios hiperparâmetros de calibração. Em todos os casos, o nível de significância adotado foi $\alpha = 0{,}10$, correspondendo a cobertura nominal de $1 - \alpha = 0{,}90$.

O *Inductive Conformal Prediction* (ICP) é parametrizado essencialmente pelo nível de significância α. O único procedimento adicional é o modo de cálculo do quantil empírico: utilizou-se o quantil de ordem $\lceil (n+1)(1-\alpha) \rceil / n$, em que n é o número de amostras de calibração, o que garante cobertura marginal em amostras finitas (Angelopoulos; Bates, 2021). Os escores de não conformidade foram calculados como resíduos absolutos $s_i = |y_i - \hat{y}_i|$ sobre o conjunto de calibração, e o intervalo final para cada ponto do conjunto de teste foi construído como $[\hat{y} - \hat{q},\ \hat{y} + \hat{q}]$, com imposição de não negatividade no limite inferior por truncamento em zero. Não há outros hiperparâmetros livres no ICP.

O *Mondrian Conformal Prediction* possui hiperparâmetros adicionais relacionados ao esquema de agrupamento. Além do nível de significância $\alpha = 0{,}10$, foram definidos: número de grupos (`n_bins = 4`), correspondente a uma divisão por quartis da distribuição de $\hat{y}$ no conjunto de calibração; tamanho mínimo de grupo (`min_group = 50`), adotado para assegurar que cada grupo contenha amostras suficientes para uma estimativa estável do quantil empírico, evitando instabilidades em grupos esparsamente populados (Angelopoulos; Bates, 2021); e fator de suavização (`shrink = 0,35`), que interpola o quantil local de cada grupo com o quantil global, atenuando a variância da estimativa em grupos com poucas amostras sem anular a

diferenciação entre regimes (Boström; Johansson; Löfström, 2021). Para cada grupo g, o quantil efetivo aplicado foi $(1 - \texttt{shrink}) \cdot \hat{q}_g + \texttt{shrink} \cdot \hat{q}_{\text{global}}$, em que $\hat{q}_g$ é o quantil local calculado sobre os escores de calibração pertencentes ao grupo, e $\hat{q}_{\text{global}}$ é o quantil calculado sobre todos os escores de calibração. Grupos com menos de `min_group` amostras utilizaram diretamente o quantil global. Esses hiperparâmetros foram definidos empiricamente com base em critérios de estabilidade da calibração, conforme discutido ao final desta seção.

O *Conformalized Quantile Regression* (CQR) depende, além de α, dos quantis estimados pelo modelo preditivo quantílico. Foram treinados modelos quantílicos específicos para estimar os limites inferior e superior das previsões, com $\tau_{lo} = 0{,}10$ e $\tau_{hi} = 0{,}90$, utilizando *pinball loss* como função de perda. O procedimento de conformalização consiste em calcular, no conjunto de calibração, os escores de não conformidade definidos por $r_i = \max\{\hat{q}_{lo}(x_i) - y_i,\ y_i - \hat{q}_{hi}(x_i),\ 0\}$, e em seguida estimar o quantil empírico $\hat{\delta}_{1-\alpha}$ desses escores. O intervalo final é então obtido como $[\max(0,\ \hat{q}_{lo} - \hat{\delta}),\ \hat{q}_{hi} + \hat{\delta}]$, com imposição de não negatividade no limite inferior (Romano; Patterson; Candès, 2019). Os hiperparâmetros livres do CQR são α, τ_{lo} e τ_{hi}, sendo os dois últimos determinados diretamente por α como $\tau_{lo} = \alpha$ e $\tau_{hi} = 1 - \alpha$.

Cabe registrar que a escolha dos hiperparâmetros de todos os modelos, neurais, estatísticos e conformes, não foi realizada por meio de otimização sistemática, como busca em grade (*grid search*) ou otimização bayesiana. As configurações adotadas foram definidas com base em quatro critérios complementares: coerência com valores amplamente utilizados na literatura para problemas similares; estabilidade observada durante testes exploratórios preliminares conduzidos ao longo do desenvolvimento; viabilidade computacional dentro do escopo de um Projeto Final de Curso, conduzido exclusivamente em CPU sem acesso a unidades de processamento gráfico; e ausência de sinais de instabilidade ou divergência durante o treinamento dos modelos neurais, monitorada por meio do mecanismo de *Early Stopping*. Para os modelos neurais *LSTM* e *GRU*, os principais hiperparâmetros definidos dessa forma foram o número de unidades da camada recorrente (64), a taxa de *Dropout* (0,2), o número de neurônios da camada densa intermediária (32), a taxa de aprendizado (10^{-3}), o *batch size* (128) e a paciência do *Early Stopping* (5 épocas). Para o *Random Forest*, os principais parâmetros foram o número de árvores (400) e o mínimo de amostras por folha (2). Para o *SARIMAX*, a especificação das ordens $(p, d, q)(P, D, Q)_s$ foi baseada nas características sazonais conhecidas da série e em práticas consolidadas da literatura para séries diárias de precipitação (Mulla; Pande; Singh, 2024). Para os métodos de *Conformal Prediction*, os hiperparâmetros do *Mondrian* foram definidos empiricamente com foco na estabilidade da calibração, conforme descrito anteriormente. É importante destacar que configurações alternativas poderiam produzir resultados diferentes, e que uma análise de sensibilidade sistemática aos hiperparâmetros constituiria um refinamento natural deste estudo. No entanto, dado o caráter comparativo e aplicado do trabalho, cujo objetivo central é investigar o comportamento relativo de diferentes combinações entre modelo base e estratégia conforme, e não otimizar cada modelo individualmente, essa limitação não compromete as conclusões gerais obtidas.

Em termos de custo computacional, observou-se variação relevante entre as famílias de modelos avaliadas. Os modelos neurais *LSTM* e *GRU* apresentaram tempo de treinamento moderado por horizonte, com duração influenciada principalmente pelo mecanismo de *Early Stopping*, que interrompeu o treinamento antecipadamente na maioria dos casos antes de atingir o limite de 30 épocas. O *Random Forest* foi o modelo com menor tempo de treinamento, beneficiado pela paralelização em todos os núcleos disponíveis do processador (`n_jobs = -1`) e pela ausência de etapa iterativa de otimização por gradiente. O *SARIMAX* foi o modelo com maior custo computacional total, devido ao seu modo de operação sequencial no período de calibração e teste: diferentemente dos modelos de aprendizado de máquina, que são treinados uma única vez e depois aplicados em lote sobre os conjuntos de calibração e teste, o *SARIMAX* atualiza seu estado interno a cada nova observação por meio do método `extend`, o que exige processamento ponto a ponto ao longo de todo o período de avaliação. Esse comportamento, inerente à natureza sequencial dos modelos em espaço de estados, tornou o *SARIMAX* significativamente mais lento do que os demais modelos em termos de tempo total de execução por horizonte, ainda que seu custo de estimação inicial dos parâmetros seja baixo. Em um ambiente com aceleração por GPU, o tempo de treinamento dos modelos neurais seria substancialmente reduzido, alterando a relação de custo computacional entre as arquiteturas. No contexto deste trabalho, o custo computacional não foi um critério de seleção ou comparação entre modelos, mas constitui uma informação relevante para quem considerar replicar ou estender estes experimentos.

Os detalhes sobre organização temporal dos dados, divisão entre conjuntos experimentais, definição dos horizontes de previsão e encadeamento completo do experimento são apresentados na Seção 4.3.

4.2 DATASET

O conjunto de dados utilizado neste trabalho é composto por registros meteorológicos da estação de Brasília, disponibilizados pelo Instituto Nacional de Meteorologia (INMET). A base original contém observações em resolução horária no período de 1 de janeiro de 2016 a 30 de setembro de 2025. Em sua forma bruta, o conjunto apresenta 85.464 registros distribuídos em 23 colunas, correspondendo a 24 observações por dia ao longo de praticamente toda a série analisada.

Entre as variáveis disponíveis, destaca-se a precipitação horária em milímetros, adotada neste estudo como variável de interesse principal. Além dela, a base inclui atributos meteorológicos potencialmente relevantes para a modelagem preditiva, como radiação global, pressão atmosférica, temperatura do ar, temperatura do ponto de orvalho, umidade relativa, velocidade do vento, rajada máxima do vento e direção do vento. Trata-se, portanto, de um conjunto multivariado em que a variável alvo pode ser analisada em conjunto com covariáveis capazes de refletir condições atmosféricas associadas à ocorrência de chuva.

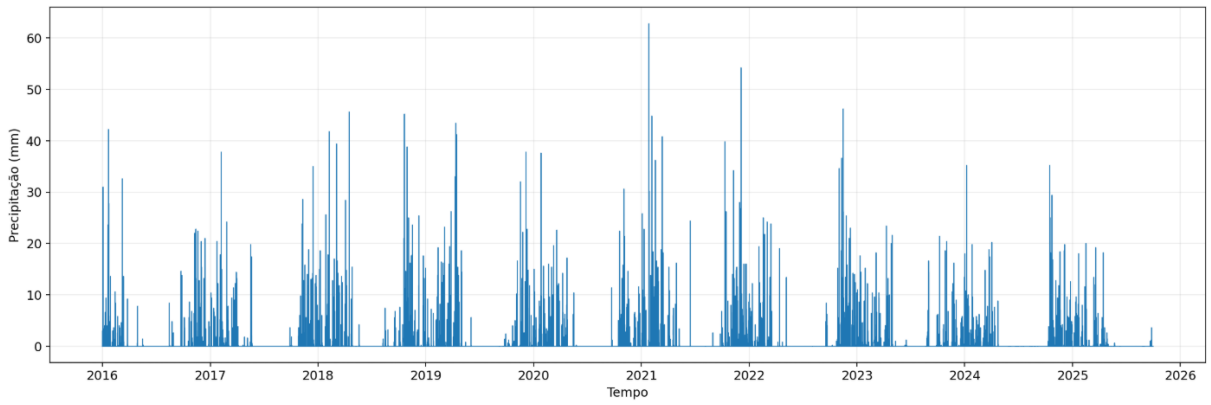
A escolha desse *dataset* se justifica tanto por razões aplicadas quanto metodológicas. Do

ponto de vista prático, a previsão de precipitação é relevante em contextos como planejamento urbano, gestão de recursos hídricos, monitoramento ambiental e mitigação de impactos associados a eventos extremos. Em situações desse tipo, não basta apenas prever chuva ou ausência de chuva: é importante também quantificar a incerteza envolvida, sobretudo quando erros podem levar a decisões inadequadas.

Sob a perspectiva metodológica, o conjunto também é bastante adequado ao problema proposto. Séries de precipitação costumam apresentar comportamento irregular, forte variabilidade temporal, muitos períodos sem chuva e ocorrência esporádica de eventos intensos. Essas características tornam o problema particularmente desafiador, especialmente quando o objetivo não se limita à previsão pontual, mas inclui também a construção de intervalos preditivos confiáveis. Nesse sentido, o *dataset* de Brasília oferece um cenário realista para avaliar tanto a capacidade preditiva dos modelos quanto a utilidade dos métodos de *Conformal Prediction* na quantificação de incerteza.

Outro aspecto relevante é a extensão temporal da base. Por abranger quase dez anos de observações meteorológicas, o conjunto permite explorar padrões sazonais, flutuações interanuais e diferentes regimes de precipitação ao longo do tempo. Essa abrangência é útil porque fornece volume amostral suficiente para treinamento, calibração e teste em diferentes horizontes preditivos, além de expor os métodos a períodos mais estáveis e a trechos de maior variabilidade. Como ilustra a Figura 10, a série apresenta alternância clara entre períodos secos e chuvosos, além de episódios pontuais de precipitação mais intensa.

Figura 10 – Série temporal da precipitação no período analisado



Legenda: Série temporal da precipitação observada na estação de Brasília ao longo do período de janeiro de 2016 a setembro de 2025, evidenciando a variabilidade temporal, a sazonalidade anual e a ocorrência pontual de eventos mais intensos.

Em síntese, o *dataset* adotado neste trabalho reúne características que o tornam especialmente pertinente para o problema estudado: trata-se de uma base real, multivariada, com extensão temporal suficiente e associada a um fenômeno naturalmente incerto e não linear. Essas propriedades justificam sua utilização como base experimental para comparar modelos de previsão e estratégias de quantificação de incerteza.

4.3 PIPELINE DOS EXPERIMENTOS

O *pipeline* experimental deste trabalho foi estruturado em duas etapas principais. A primeira correspondeu ao pré-processamento do arquivo bruto obtido junto ao INMET, com o objetivo de convertê-lo em uma base diária consistente para modelagem. A segunda etapa foi realizada já no ambiente dos *notebooks* e envolveu seleção e transformação de atributos, construção das amostras supervisionadas, particionamento temporal e preparação específica para cada família de modelo. Para tornar esse processo mais claro, o *pipeline* foi representado em quatro figuras complementares, correspondentes às etapas de preparação do dataset, construção supervisionada e divisão temporal, modelagem preditiva, e quantificação de incerteza com avaliação final.

Inicialmente, o arquivo original em resolução horária foi transformado em uma base diária. Nessa etapa, valores sentinela iguais a `-9999`, utilizados no arquivo bruto para indicar ausência de observação, foram convertidos em valores ausentes. Em seguida, as variáveis meteorológicas foram agregadas por dia de acordo com a natureza física de cada atributo. A variável alvo `PRECIP_MM` foi obtida por soma diária da precipitação horária, já que o interesse do trabalho recai sobre o volume total acumulado em cada dia. A variável `RADIACAO_GLOBAL_(KJ_M)` também foi agregada por soma diária. As variáveis de pressão atmosférica, temperatura do ar, temperatura do ponto de orvalho, umidade relativa e velocidade do vento foram agregadas por média diária. A variável `VENTO_RAJADA_MAXIMA_(M_S)` foi sintetizada por máximo diário, enquanto a direção do vento foi agregada por média circular, o que evita distorções decorrentes de sua natureza angular. Além disso, colunas auxiliares presentes no arquivo bruto, como `DATE`, `TIME_UTC`, `UNNAMED_19` e `__SOURCE_FILE`, foram removidas da versão final utilizada nos experimentos. Essa transformação inicial é resumida na Figura 11.

Figura 11 – Preparação do conjunto de dados

Legenda: Conversão da base horária em uma série diária consistente para modelagem. Etapas iniciais do *pipeline* experimental, desde o *dataset* bruto do INMET até a construção da base diária final utilizada nos experimentos.

Ainda nessa etapa, foram incorporadas variáveis de calendário derivadas da informação temporal. O *dataset* diário final passou a conter os atributos `year`, `month`, `day`, `weekday` e `dayofyear`, além das variáveis cíclicas `sin_doy` e `cos_doy`, calculadas a partir do dia do ano. A inclusão dessas variáveis teve como objetivo explicitar padrões sazonais, em especial os associados ao ciclo anual da precipitação.

Após a obtenção do *dataset* diário, o arquivo processado foi carregado nos *notebooks* por meio de uma rotina de leitura robusta, preparada para diferentes codificações de caracteres. Em seguida, a coluna temporal foi detectada automaticamente, convertida para o tipo *datetime* e utilizada para ordenar cronologicamente toda a série. Registros com data inválida foram removidos. A variável alvo `PRECIP_MM` também foi convertida explicitamente para formato numérico, e observações remanescentes com valor ausente nessa variável foram descartadas antes das etapas posteriores de modelagem. Eventuais valores ausentes residuais nas covariáveis foram preenchidos por propagação temporal para frente e para trás (*forward fill* seguido de *backward fill*), como salvaguarda contra lacunas pontuais que poderiam comprometer a construção das janelas supervisionadas.

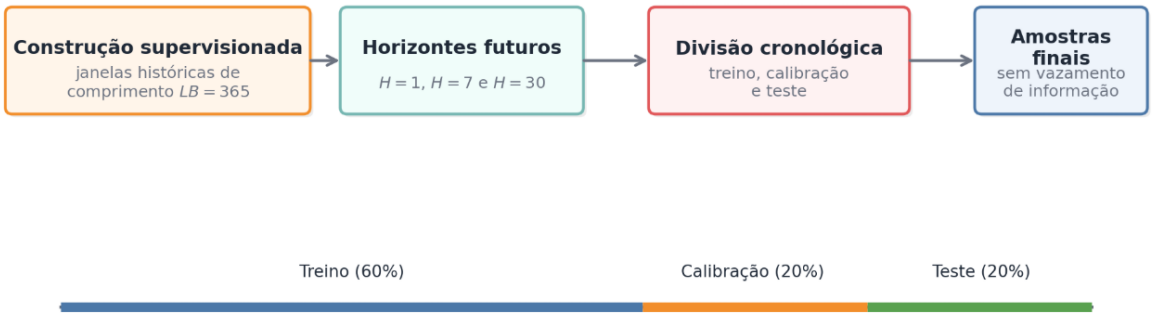
Na construção dos atributos preditores, foram adotadas transformações adicionais para tornar o conjunto mais adequado aos modelos. A direção do vento, originalmente representada em graus, foi convertida para uma representação cíclica por meio das variáveis seno e cosseno do ângulo, evitando a descontinuidade artificial entre 0° e 360°. De forma semelhante, nos *notebooks* dos modelos *LSTM* e *Random Forest*, a sazonalidade anual foi representada pelas variáveis `month`, `doy_sin` e `doy_cos`, derivadas diretamente da coluna temporal. No notebook da *GRU*, foram aproveitadas as variáveis de calendário já presentes no *dataset* diário, com reconstrução de `sin_doy` e `cos_doy` quando necessário.

A seleção das variáveis de entrada foi realizada de maneira distinta conforme a arquitetura considerada. Nos modelos *LSTM* e *Random Forest*, foi adotado um conjunto de atributos composto pelo histórico da própria precipitação, radiação global, pressão atmosférica ao nível da estação, temperatura do ar, temperatura do ponto de orvalho, umidade relativa, velocidade do vento, rajada máxima do vento, codificação cíclica da direção do vento e atributos sazonais de calendário. No modelo *GRU*, foi empregada uma configuração mais ampla, incluindo, além da precipitação passada, variáveis meteorológicas adicionais de máximos e mínimos horários previamente agregados para a escala diária, bem como os atributos de calendário disponíveis no *dataset* processado. Já no caso do *SARIMAX*, a modelagem foi conduzida de forma univariada, utilizando exclusivamente a série temporal da precipitação.

Para transformar a série em um problema de aprendizado supervisionado, foram construídas janelas temporais deslizantes contendo apenas informações passadas. Em todos os modelos baseados em aprendizado de máquina, utilizou-se uma janela histórica de comprimento $LB = 365$ dias, de modo que cada amostra foi formada pelos 365 dias imediatamente anteriores ao instante de previsão. Essa escolha foi motivada pela necessidade de capturar um ciclo anual completo de sazonalidade, especialmente relevante para séries de precipitação com padrão sazonal marcado como a de Brasília. A variável resposta foi definida como a precipitação observada em um horizonte futuro H, sendo considerados os horizontes $H = 1$, $H = 7$ e $H = 30$, correspondentes, respectivamente, às previsões de curto, médio e longo prazo. Essa construção garante que cada predição seja gerada exclusivamente a partir de informações disponíveis até o instante corrente, evitando vazamento de informação futura. A organização

dessas janelas e sua relação com os horizontes de previsão são ilustradas na Figura 12.

Figura 12 – Construção supervisionada e divisão temporal



Legenda: As amostras são construídas apenas com informações passadas, preservando a ordem temporal da série. Transformação da série temporal em amostras supervisionadas, definição dos horizontes preditivos e particionamento cronológico em treino, calibração e teste.

A forma de representar essas janelas variou conforme o modelo. Nos modelos recorrentes *LSTM* e *GRU*, cada amostra foi mantida no formato matricial $(LB \times F)$, preservando explicitamente a estrutura sequencial temporal, em que F é o número de variáveis de entrada. No modelo *Random Forest*, por outro lado, cada janela foi achatada em um vetor unidimensional de dimensão $LB \times F$, convertendo o problema para uma representação tabular mais compatível com estimadores baseados em árvores. Para o *SARIMAX*, não houve construção de janelas multivariadas; nesse caso, a série foi mantida em sua forma temporal original, utilizando-se os primeiros 365 dias como histórico mínimo inicial para geração das previsões.

Após a construção das amostras supervisionadas, realizou-se o particionamento temporal dos dados em três subconjuntos: treinamento, calibração e teste, nas proporções de 60%, 20% e 20%, respectivamente. Esse particionamento preservou integralmente a ordem cronológica da série, sem embaralhamento, de modo a manter coerência com o problema real de previsão temporal. O conjunto de calibração foi utilizado exclusivamente para o cálculo dos escores de não conformidade de cada método de *Conformal Prediction*, e o conjunto de teste foi reservado para avaliação final, não sendo acessado em nenhuma etapa anterior. A proporção de 20% para o conjunto de calibração foi escolhida para garantir volume amostral suficiente para a estimação estável dos quantis empíricos dos métodos conformes, especialmente no caso do *Mondrian*, que realiza calibração separada por grupos. No caso do *SARIMAX*, o conjunto de avaliação passou a ser considerado apenas após a disponibilidade do histórico mínimo de 365 observações, e as previsões foram produzidas de forma sequencial, com atualização do estado do modelo ao longo da calibração e do teste, mas sem reestimação completa dos parâmetros a cada novo ponto observado.

A etapa de modelagem propriamente dita envolveu quatro famílias de modelos preditivos: *SARIMAX*, *Random Forest*, *LSTM* e *GRU*. Essa escolha permitiu comparar, em um mesmo

problema, uma abordagem estatística clássica, um método baseado em árvores e duas arquiteturas neurais recorrentes. A Figura 13 sintetiza essa etapa, destacando os modelos avaliados no estudo.

Figura 13 – Modelos preditivos avaliados

Legenda: Quatro famílias de modelos foram comparadas sob os mesmos horizontes de previsão.
Conjunto de modelos preditivos comparados no trabalho, incluindo uma abordagem estatística clássica, um *ensemble* baseado em árvores e duas arquiteturas neurais recorrentes.

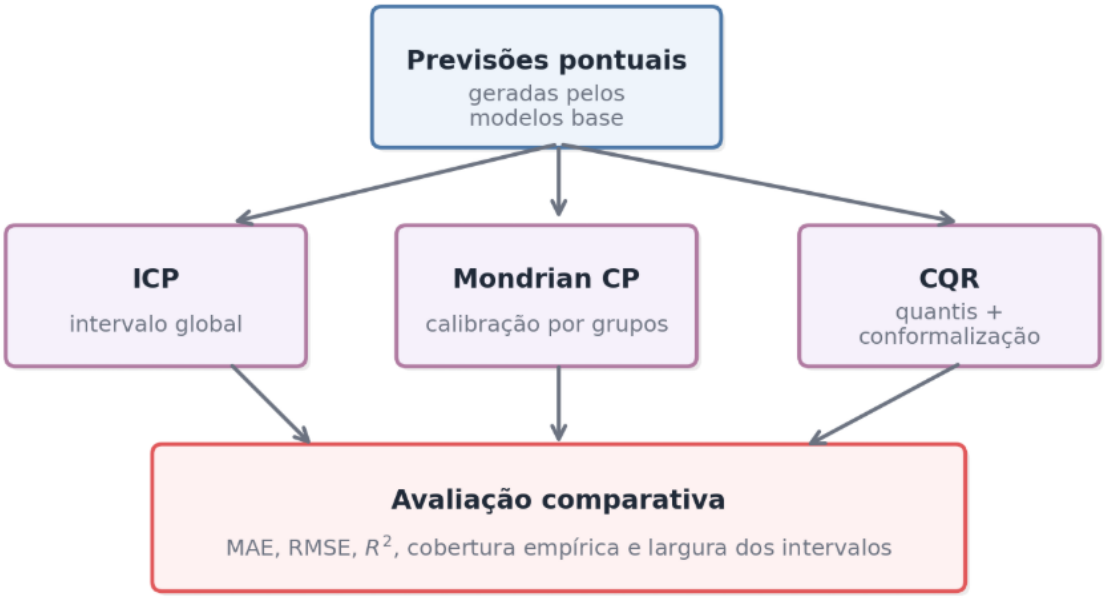
Etapas adicionais de pré-processamento foram aplicadas especificamente aos modelos neurais. Para *LSTM* e *GRU*, as variáveis de entrada foram padronizadas com `StandardScaler` da biblioteca *scikit-learn*, ajustado exclusivamente sobre o conjunto de treinamento e posteriormente aplicado, sem novo ajuste, aos conjuntos de calibração e teste. Essa escolha evita vazamento de informação estatística entre as partições. A necessidade da padronização foi verificada empiricamente durante o desenvolvimento: testes exploratórios preliminares conduzidos sem normalização das variáveis de entrada produziram resultados consistentemente inferiores em termos de MAE e RMSE, além de instabilidade no treinamento dos modelos neurais, manifestada por oscilações na curva de perda de validação e ativação prematura do *Early Stopping*. Esses resultados motivaram a adoção da padronização como etapa obrigatória do *pipeline* para as arquiteturas recorrentes. Para os modelos *Random Forest* e *SARIMAX*, a padronização das entradas não foi aplicada, pois esses métodos não são sensíveis à escala das variáveis da forma que redes neurais são: o *Random Forest* opera com partições baseadas em limiares que são invariantes a transformações monotônicas, e o *SARIMAX* estima seus parâmetros por máxima verossimilhança de forma independente da escala das covariáveis.

Ainda nos modelos neurais, a variável alvo foi transformada por $\log(1 + y)$, após imposição de não negatividade, com o objetivo de reduzir a assimetria da distribuição da precipitação e estabilizar o treinamento. Essa transformação foi aplicada tanto às versões pontuais quanto às versões quantílicas da *GRU*, e à versão quantílica do *LSTM*. Na versão pontual do *LSTM*, a variável alvo foi mantida em escala original de milímetros, compatível com a função de perda *Tweedie* utilizada. Após a inferência, as previsões transformadas foram reconvertidas à escala original por meio da transformação inversa $\exp(\cdot) - 1$. Nos modelos *Random Forest* e *SARIMAX*, a variável alvo foi mantida em sua escala original de milímetros em todas as etapas.

Por fim, os métodos de *Conformal Prediction* foram aplicados sobre as previsões obtidas em cada horizonte. O conjunto de calibração foi utilizado para calcular os escores de não conformidade que serviram de base para a construção dos intervalos preditivos nos métodos *Inductive*

Conformal Prediction (ICP) e *Mondrian Conformal Prediction*. No caso do *Conformalized Quantile Regression* (CQR), os modelos quantílicos forneceram inicialmente limites inferiores e superiores, que foram posteriormente ajustados pela etapa de conformalização com base nos resíduos observados no conjunto de calibração. Em todos os casos, os quantis empíricos foram estimados exclusivamente sobre o conjunto de calibração, e os intervalos finais foram construídos e avaliados sobre o conjunto de teste. Em seguida, as previsões pontuais e os intervalos gerados foram avaliados por métricas de acurácia, cobertura empírica e largura dos intervalos, todas calculadas na escala original de milímetros após reconversão das previsões transformadas. Essa etapa final do *pipeline* é resumida na Figura 14.

Figura 14 – Quantificação de incerteza e avaliação final



Legenda: As previsões pontuais foram transformadas em intervalos preditivos e avaliadas segundo critérios de acurácia, validade e informatividade. Etapa final do *pipeline* experimental, envolvendo a aplicação dos métodos de *Conformal Prediction* e a avaliação comparativa das previsões pontuais e dos intervalos preditivos.

Em conjunto, esse *pipeline* articulou preparação do conjunto de dados, construção supervisionada, modelagem preditiva e quantificação de incerteza em uma estrutura única e coerente, permitindo comparar diferentes famílias de modelos e diferentes estratégias conformes sob os mesmos horizontes de previsão.

4.3.1 Reprodutibilidade

Para garantir a reprodutibilidade completa dos experimentos, registram-se aqui as condições necessárias para reexecução. O ambiente utilizado foi Python 3.11.9 com semente aleatória 42 fixada simultaneamente em *NumPy*, *Python* nativo e *TensorFlow*. Os dados podem ser obtidos diretamente do portal público do INMET para a estação de Brasília, período de janeiro de 2016

a setembro de 2025. Qualquer diferença de hardware ou versão de biblioteca pode produzir pequenas variações numéricas nos resultados, especialmente nos modelos neurais, devido a não determinismo residual em operações de ponto flutuante em CPU mesmo com semente fixada.

4.4 MÉTRICAS

A avaliação dos experimentos foi conduzida sob duas perspectivas complementares: o desempenho das previsões pontuais e a qualidade dos intervalos preditivos. Assim, buscou-se analisar não apenas o quão próximas as previsões ficaram dos valores observados, mas também se os intervalos produzidos foram, ao mesmo tempo, confiáveis e informativos.

Para avaliar as previsões pontuais, foram utilizadas as métricas *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE) e coeficiente de determinação (R^2). O MAE mede o erro absoluto médio entre os valores observados e previstos, sendo definido por

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{18}$$

em que $y_i \in R_{\geq 0}$ representa o valor observado de precipitação no instante i, $\hat{y}_i \in R_{\geq 0}$ é a previsão pontual correspondente produzida pelo modelo, e n é o número total de observações avaliadas no conjunto de teste. Como essa métrica é expressa na unidade original do problema, sua interpretação é direta, permitindo quantificar, em média, o desvio absoluto das previsões em milímetros de precipitação.

O RMSE atribui maior penalização a erros de maior magnitude, sendo dado por

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{19}$$

em que y_i, $\hat{y}_i$ e n têm as mesmas definições da equação 18. Essa métrica é particularmente relevante em séries pluviométricas, nas quais erros associados a eventos intensos tendem a ter maior impacto prático. Assim como o MAE, o RMSE é expresso na unidade original da variável alvo.

O coeficiente de determinação R^2 é calculado por

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{20}$$

em que $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ representa a média dos valores observados no conjunto de teste. O R^2 varia tipicamente entre $-\infty$ e 1: valores próximos de 1 indicam que o modelo explica bem a variabilidade da série; valores próximos de 0 indicam desempenho equivalente ao de uma previsão pela média; e valores negativos indicam que o modelo é pior do que simplesmente prever $\bar{y}$ para todas as observações.

No caso dos intervalos preditivos gerados pelos métodos de *Conformal Prediction*, é necessário antes definir formalmente os parâmetros que governam sua construção. O nível de significância $\alpha \in (0, 1)$ controla o grau de conservadorismo do intervalo: quanto menor α, maior a cobertura exigida e, consequentemente, mais largos tendem a ser os intervalos. O valor complementar $1 - \alpha$ é a cobertura nominal desejada, ou seja, a fração mínima de observações do conjunto de teste que se espera que o intervalo cubra. Neste trabalho, adotou-se $\alpha = 0{,}10$, correspondendo a uma cobertura nominal de $1 - \alpha = 0{,}90$, ou seja, espera-se que aproximadamente 90% das observações reais pertençam aos intervalos construídos.

A principal métrica de validade dos intervalos adotada foi a cobertura empírica (*empirical coverage*), que mede a fração de observações reais efetivamente contidas nos intervalos estimados, sendo calculada por

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^{n} I\left(y_i \in [L_i, U_i]\right) \tag{21}$$

em que $L_i \in R_{\geq 0}$ e $U_i \in R_{\geq 0}$ representam, respectivamente, os limites inferior e superior do intervalo preditivo associado à observação i, com $L_i \leq U_i$, e $I(\cdot)$ é a função indicadora, que assume valor 1 quando a condição entre parênteses é verdadeira e 0 caso contrário. A cobertura empírica deve ser interpretada em relação à cobertura nominal $1 - \alpha = 0{,}90$: valores próximos a $0{,}90$ indicam intervalos bem calibrados; valores muito superiores indicam intervalos excessivamente conservadores; e valores inferiores indicam subcalibração.

Entretanto, a cobertura isoladamente não é suficiente para avaliar a qualidade dos intervalos. Um método pode alcançar alta cobertura apenas por gerar intervalos muito largos, o que reduz sua utilidade prática. Por essa razão, também foram analisadas a largura pontual e a largura média dos intervalos. A largura pontual de cada intervalo é definida por

$$W_i = U_i - L_i \tag{22}$$

em que U_i e L_i têm as mesmas definições da equação 21. A largura média ao longo do conjunto de teste é então

$$\overline{W} = \frac{1}{n} \sum_{i=1}^{n} W_i \tag{23}$$

utilizada neste trabalho como principal medida de informatividade dos intervalos. Em termos práticos, quanto menor $\overline{W}$, mais informativo é o intervalo, desde que a cobertura empírica permaneça próxima do nível nominal $1 - \alpha$.

Dessa forma, a análise dos métodos conformes considerou conjuntamente dois critérios centrais: *validade*, representada pela cobertura empírica próxima a $1 - \alpha = 0{,}90$, e *informatividade*, representada pela menor largura média possível dado esse nível de cobertura. Essa combinação permite comparar não apenas qual modelo prevê melhor em média, mas também qual estratégia

produz intervalos mais confiáveis e mais úteis em um contexto de apoio à decisão em previsão pluviométrica.

4.5 RESULTADOS PARA O CURTO PRAZO

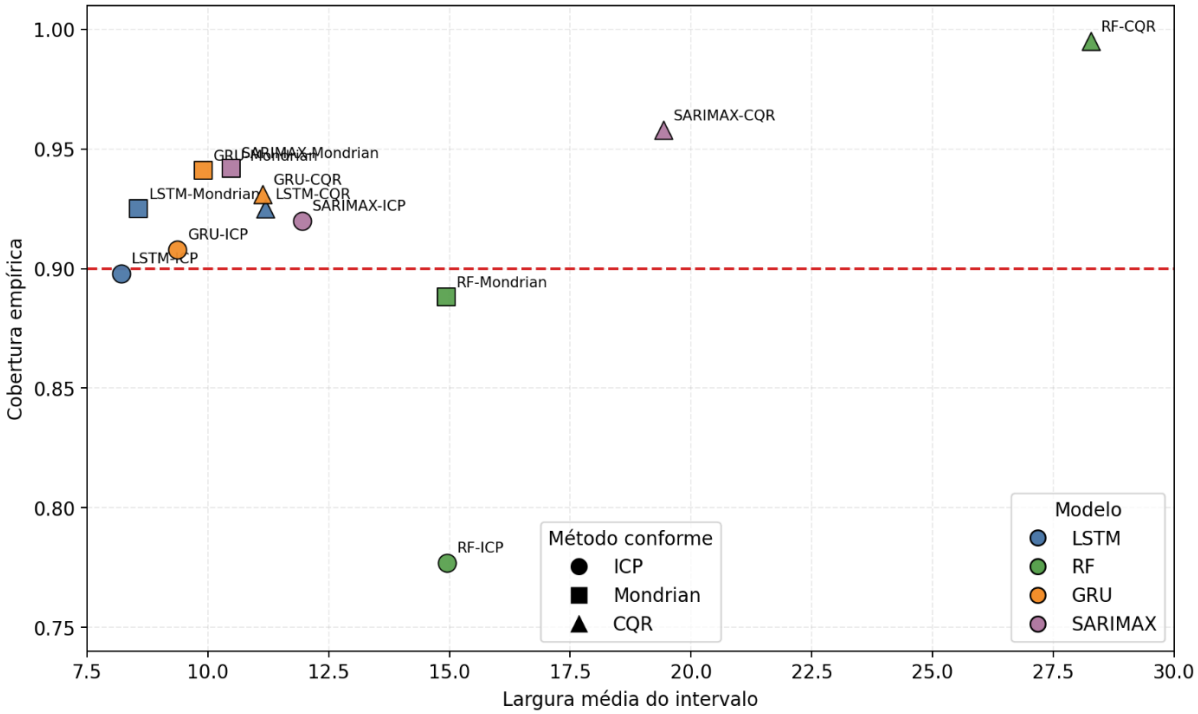
A Tabela 1 apresenta os resultados obtidos para o horizonte de curto prazo ($H = 1$). Como o foco deste trabalho está na qualidade dos intervalos conformes, a análise desse horizonte foi orientada principalmente por dois critérios: a proximidade da cobertura empírica em relação ao nível nominal de $0{,}90$ e a largura média dos intervalos. A capacidade de acompanhar picos e oscilações locais da série foi utilizada como critério qualitativo complementar, especialmente quando duas ou mais combinações apresentaram desempenho quantitativo semelhante. Assim, coberturas muito acima de $0{,}90$ não foram tratadas automaticamente como superiores, pois também podem refletir intervalos excessivamente conservadores.

Tabela 1 – Resultados obtidos para o horizonte de curto prazo ($H = 1$).

Modelo	**CP**	**MAE**	**RMSE**	**R^2**	**Coverage**	**Width**
LSTM	ICP	**2,984**	7,820	0,088	**0,898**	**8,208**
	Mondrian				0,925	8,568
	CQR				0,925	11,191
RF	ICP	5,938	8,686	-0,125	0,777	14,950
	Mondrian				0,888	14,930
	CQR				0,995	28,280
GRU	ICP	3,042	7,846	0,082	0,908	9,372
	Mondrian				0,941	9,898
	CQR				0,931	11,140
SARIMAX	ICP	3,517	**7,673**	**0,116**	0,920	11,948
	Mondrian				0,942	10,484
	CQR				0,958	19,427

A Figura 15 resume a relação entre cobertura empírica e largura média dos intervalos para $H = 1$. Nessa perspectiva, as combinações mais interessantes não são necessariamente aquelas com maior *coverage*, mas sim aquelas que permanecem próximas de $0{,}90$ sem ampliar excessivamente a largura dos intervalos. Sob esse critério, o *LSTM* com ICP se destaca por combinar cobertura de $0{,}898$, praticamente coincidente com o valor nominal, e a menor largura média entre as combinações competitivas, $8{,}208$. O *LSTM* com *Mondrian* também apresenta resultado bastante forte, com *coverage* de $0{,}925$ e largura média de $8{,}568$, enquanto a *GRU* com ICP atinge *coverage* de $0{,}908$ com largura de $9{,}372$, mantendo-se igualmente próxima do alvo.

Figura 15 – Cobertura empírica versus largura média dos intervalos no horizonte $H = 1$



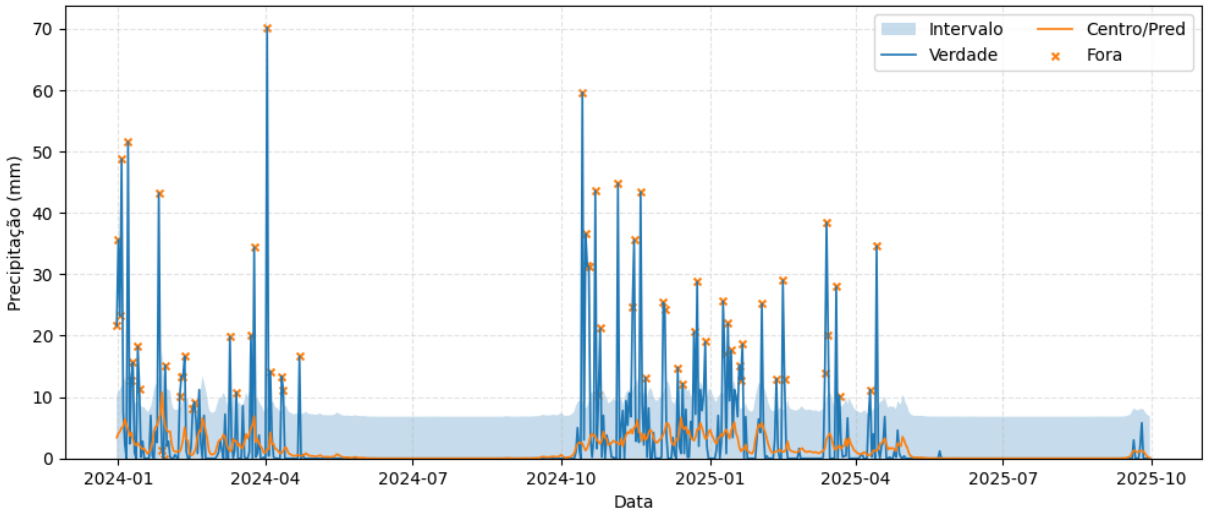
Legenda: Relação entre cobertura empírica e largura média dos intervalos no horizonte de curto prazo, considerando todas as combinações entre modelos e métodos de *Conformal Prediction*.

Embora combinações como *SARIMAX* + *Mondrian* e *SARIMAX* + CQR tenham apresentado coberturas mais altas, esse ganho veio acompanhado de maior afastamento em relação ao alvo de $0{,}90$ e, no caso do CQR, de intervalos substancialmente mais largos. Assim, quando se evita interpretar conservadorismo excessivo como vantagem automática, o destaque quantitativo do curto prazo passa a recair principalmente sobre combinações baseadas em *LSTM*.

Em particular, a combinação *LSTM* + ICP se sobressai nesse horizonte por reunir cobertura praticamente coincidente com o nível nominal e a menor largura média entre as combinações mais competitivas. Esse resultado indica que, no curto prazo, uma estratégia conforme mais simples pode ser suficiente para produzir intervalos bastante eficientes quando associada a um modelo base capaz de representar adequadamente a dinâmica recente da série. Ao mesmo tempo, outras combinações com *LSTM*, como *LSTM* + *Mondrian* e *LSTM* + CQR, também permanecem relevantes, embora privilegiem compromissos ligeiramente diferentes entre cobertura, largura e sensibilidade local.

Como caso representativo de bom desempenho nesse horizonte, a Figura 16 apresenta a combinação entre *LSTM* e ICP. Essa escolha se justifica não apenas pela cobertura muito próxima do nível nominal, mas também pela largura reduzida dos intervalos e pela boa aderência visual à dinâmica geral da série.

Figura 16 – Resultado representativo do *LSTM* com ICP no horizonte $H = 1$



Legenda: Exemplo representativo de resultado no horizonte de curto prazo, destacando a combinação entre *LSTM* e ICP, que apresentou cobertura muito próxima do nível nominal com baixa largura média dos intervalos e boa aderência visual à série.

Por outro lado, a inspeção visual dos gráficos sugere que o *LSTM* com CQR tende, em vários trechos, a acompanhar melhor os picos e oscilações locais da série. Esse comportamento não o torna automaticamente o melhor método do horizonte, já que seus intervalos são mais largos, mas o torna uma referência qualitativa importante quando se deseja observar maior sensibilidade a extremos. Essa leitura também é compatível com a discussão apresentada no Capítulo 2 sobre o CQR, cuja estrutura busca produzir intervalos mais adaptativos à incerteza condicional.

Além disso, ainda que não tenha sido realizada uma avaliação formal separada por regime seco e chuvoso, a boa consistência do *Mondrian* e do *LSTM* em parte dos resultados sugere comportamento compatível com a presença de diferentes regimes de variabilidade na série. Em termos pontuais, *LSTM*, *GRU* e *SARIMAX* apresentaram MAE e RMSE bem inferiores aos do *Random Forest*, além de valores positivos de R^2, enquanto o *Random Forest* obteve R^2 negativo. Como essas métricas dependem do modelo base, e não do método conforme em si, elas reforçam que arquiteturas temporalmente mais adequadas tendem a oferecer uma base mais favorável também para a construção dos intervalos preditivos.

4.6 RESULTADOS PARA O MÉDIO PRAZO

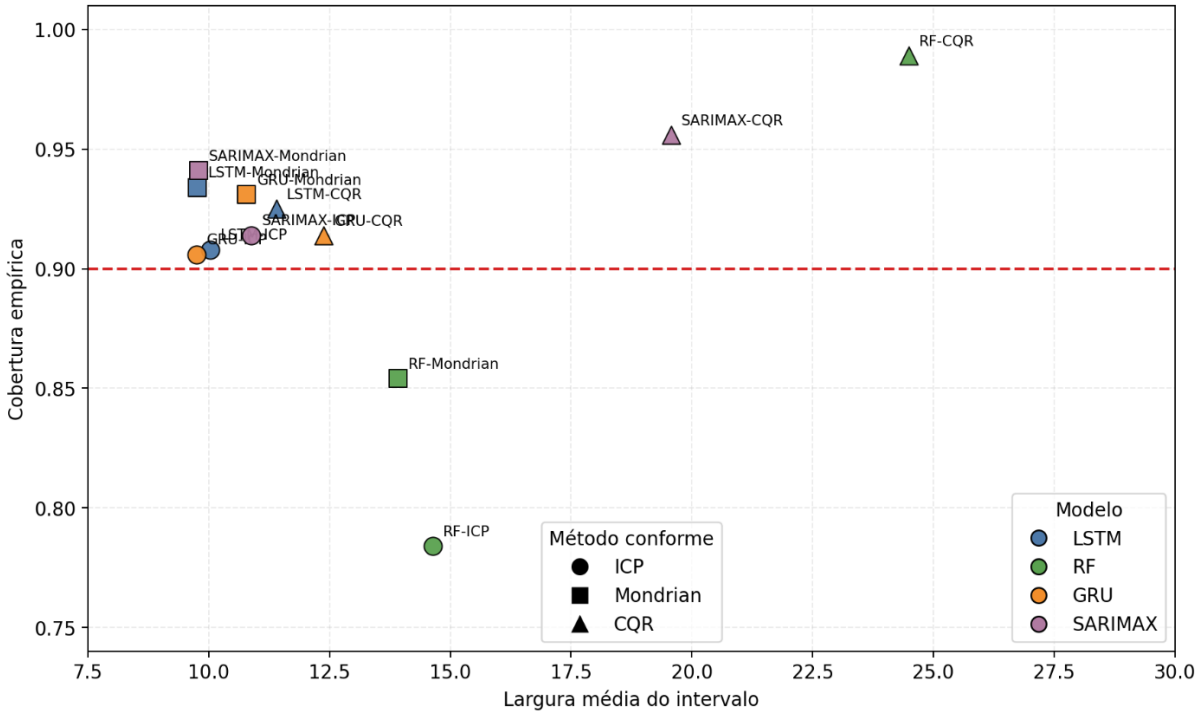
A Tabela 2 reúne os resultados obtidos para o horizonte de médio prazo ($H = 7$). Assim como no curto prazo, a análise foi orientada principalmente pela proximidade da cobertura em relação a $0{,}90$ e pela largura média dos intervalos, utilizando a representação de picos como critério qualitativo complementar. Nesse horizonte, o quadro é menos concentrado em uma única combinação, já que diferentes alternativas apresentaram desempenho bastante próximo.

Tabela 2 – Resultados obtidos para o horizonte de médio prazo ($H = 7$).

Modelo	CP	MAE	RMSE	R^2	Coverage	Width
LSTM	ICP	**3,091**	8,029	0,032	0,908	10,027
	Mondrian				0,934	9,769
	CQR				0,925	11,405
RF	ICP	5,686	9,084	-0,239	0,784	14,641
	Mondrian				0,854	13,917
	CQR				0,989	24,490
GRU	ICP	3,167	7,958	0,049	**0,906**	**9,750**
	Mondrian				0,931	10,785
	CQR				0,914	12,383
SARIMAX	ICP	3,454	**7,721**	**0,105**	0,914	10,883
	Mondrian				0,941	9,786
	CQR				0,956	19,570

A Figura 17 mostra que três combinações concentram os resultados mais competitivos nesse horizonte: *GRU* + ICP, *LSTM* + *Mondrian* e *SARIMAX* + *Mondrian*. Entre elas, o *GRU* com ICP apresenta a combinação quantitativamente mais eficiente entre proximidade ao alvo e menor largura média, com *coverage* de $0{,}906$ e width de $9{,}750$. O *LSTM* com *Mondrian* aparece logo em seguida, com *coverage* de $0{,}934$ e width de $9{,}769$, enquanto o *SARIMAX* com *Mondrian* registra *coverage* de $0{,}941$ e width de $9{,}786$. As diferenças entre essas duas últimas combinações são pequenas, e ambas permanecem competitivas do ponto de vista intervalar.

Figura 17 – Cobertura empírica versus largura média dos intervalos no horizonte $H = 7$



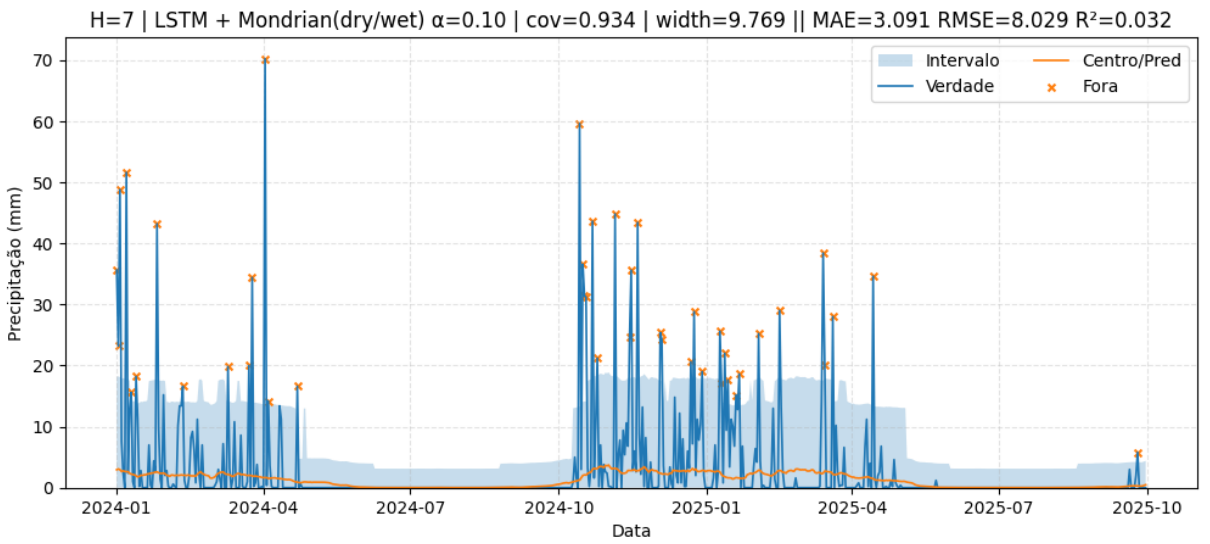
Legenda: Relação entre cobertura empírica e largura média dos intervalos no horizonte de médio prazo, considerando todas as combinações entre modelos e métodos de *Conformal Prediction*.

Nesse caso, a escolha entre as combinações mais próximas depende mais claramente do critério qualitativo complementar, isto é, da capacidade de acompanhar picos e oscilações locais da série. Sob esse aspecto, o *LSTM* com CQR e, em menor grau, o *LSTM* com *Mondrian* tendem a apresentar maior aderência visual à dinâmica local. Assim, embora o *GRU* + ICP tenha uma vantagem quantitativa ligeira em termos de *coverage* próximo ao alvo e menor largura, o *LSTM* com *Mondrian* permanece como alternativa particularmente forte quando se considera também a representação visual dos extremos.

Essa leitura é compatível com a discussão teórica do Capítulo 2. Por um lado, o bom desempenho do *GRU* + ICP mostra que uma estratégia global simples ainda pode ser bastante eficiente quando combinada a um modelo base competitivo. Por outro, o desempenho consistente do *Mondrian* e a boa resposta visual do *LSTM* em trechos mais irregulares sugerem um comportamento compatível com a existência de diferentes regimes de variabilidade ao longo da série, ainda que essa hipótese não tenha sido testada formalmente por uma análise separada de períodos secos e chuvosos.

Por essa razão, a Figura 18 apresenta o *LSTM* com *Mondrian* como caso representativo de bom resultado no horizonte intermediário. A escolha dessa combinação não decorre de superioridade isolada em todas as métricas, mas do fato de ela reunir desempenho intervalar competitivo e um comportamento visual convincente na representação dos picos.

Figura 18 – Resultado representativo do *LSTM* com *Mondrian* no horizonte $H = 7$



Legenda: Exemplo representativo de resultado no horizonte de médio prazo, destacando a combinação entre *LSTM* e *Mondrian Conformal Prediction*, que apresentou bom equilíbrio entre cobertura, largura dos intervalos e representação visual dos picos.

As métricas pontuais continuam consistentes com essa leitura. *LSTM*, *GRU* e *SARIMAX* mantiveram erros bem inferiores aos do *Random Forest*, além de R^2 positivo, o que reforça a importância do modelo base para a utilidade final dos intervalos conformes.

4.7 RESULTADOS PARA O LONGO PRAZO

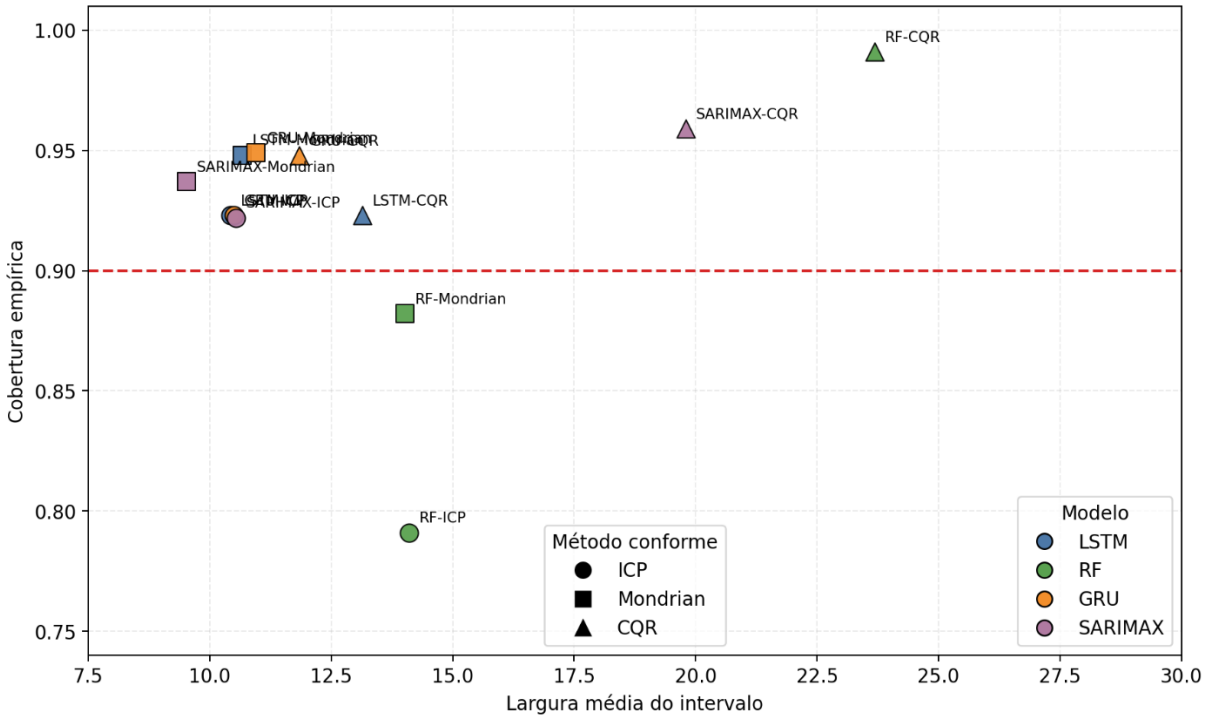
A Tabela 3 apresenta os resultados para o horizonte de longo prazo ($H = 30$). Nesse cenário, a análise passa a favorecer de forma mais clara o *Mondrian*, especialmente nas combinações com *SARIMAX* e *GRU*. Ainda assim, o mesmo critério foi mantido: prioridade para *coverage* próximo de $0{,}90$, largura média reduzida e, de forma complementar, boa representação dos picos.

Tabela 3 – Resultados obtidos para o horizonte de longo prazo ($H = 30$).

Modelo	**CP**	**MAE**	**RMSE**	**R^2**	**Coverage**	**Width**
LSTM	ICP	3,139	7,723	0,031	0,923	10,430
	Mondrian				0,948	10,672
	CQR				0,923	13,134
RF	ICP	5,347	8,610	-0,205	0,791	14,095
	Mondrian				**0,882**	14,012
	CQR				0,991	23,686
GRU	ICP	**2,994**	**7,678**	0,042	0,923	10,494
	Mondrian				0,949	10,954
	CQR				0,948	11,846
SARIMAX	ICP	3,216	7,737	**0,101**	0,922	10,542
	Mondrian				0,937	**9,518**
	CQR				0,959	19,794

Na Figura 19, a combinação *SARIMAX* + *Mondrian* assume posição particularmente forte: *coverage* de $0{,}937$, suficientemente próximo do alvo, e a menor largura média do horizonte entre os métodos competitivos, $9{,}518$. *LSTM* + ICP, *GRU* + ICP e *GRU* + *Mondrian* também apresentam resultados sólidos, mas com larguras ligeiramente superiores.

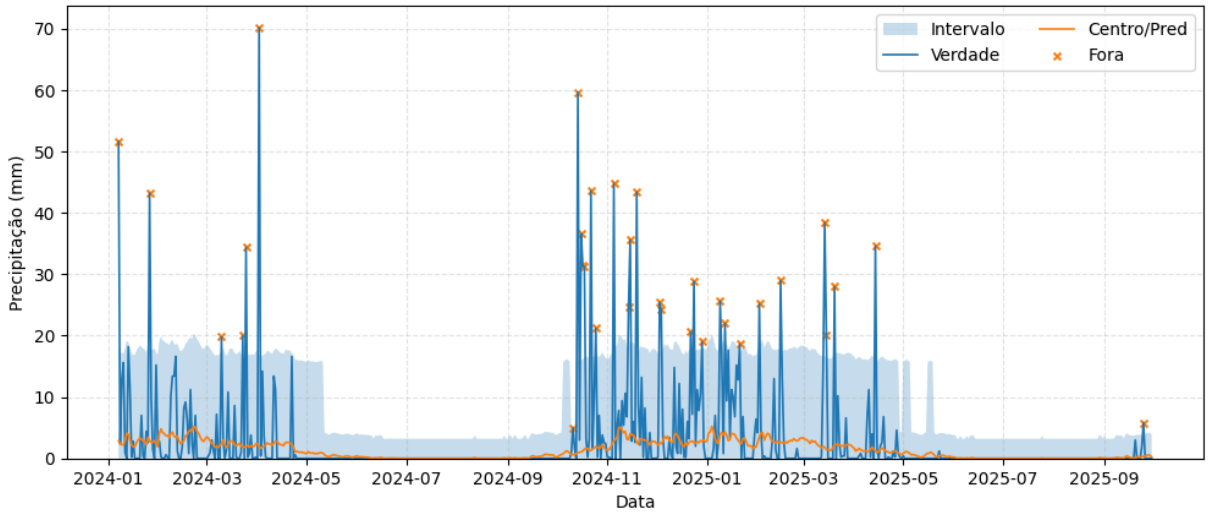
Figura 19 – Cobertura empírica versus largura média dos intervalos no horizonte $H = 30$



Legenda: Relação entre cobertura empírica e largura média dos intervalos no horizonte de longo prazo, considerando todas as combinações entre modelos e métodos de *Conformal Prediction*.

Ao mesmo tempo, o horizonte longo traz um resultado adicional importante: o modelo base *GRU* passa a apresentar o melhor desempenho pontual em MAE e RMSE. Entre os métodos conformes aplicados a esse modelo, a combinação com *Mondrian* produz *coverage* de $0{,}949$ com largura ainda moderada. Por esse motivo, a Figura 20 foi mantida como caso representativo. Embora *SARIMAX + Mondrian* tenha sido a combinação mais eficiente em termos estritamente intervalares, a escolha da *GRU + Mondrian* como figura representativa permite ilustrar melhor uma situação em que desempenho pontual competitivo e qualidade intervalar aparecem de forma conjunta.

Figura 20 – Resultado representativo da *GRU* com *Mondrian* no horizonte $H = 30$



Legenda: Exemplo representativo do comportamento da *GRU* combinada com *Mondrian Conformal Prediction* no longo prazo, destacando uma solução equilibrada entre cobertura adequada, largura moderada e boa qualidade preditiva.

A observação qualitativa sobre os picos permanece válida também nesse horizonte. Em alguns trechos, o *LSTM* com CQR ainda parece mais responsivo aos extremos. No entanto, como seus intervalos continuam mais largos, ele não se impõe como melhor combinação global do horizonte. No longo prazo, portanto, os resultados favorecem de forma mais clara o *SARIMAX* + *Mondrian* em termos de eficiência intervalar, com a *GRU* + *Mondrian* surgindo como alternativa particularmente forte quando também se considera o desempenho pontual do modelo base.

Esse resultado também ajuda a qualificar a discussão teórica do Capítulo 2. Embora o *SARIMAX* seja um modelo linear e clássico, seu bom desempenho mostra que maior complexidade não implica, necessariamente, superioridade sistemática. Em uma série com estrutura temporal e sazonal ainda relevante, um *benchmark* estatístico bem especificado pode permanecer competitivo, inclusive quando combinado a uma estratégia conforme adequada. Ao mesmo tempo, o desempenho do *Mondrian* continua compatível com a hipótese de que calibrações mais segmentadas podem ser vantajosas em séries com regimes distintos de variabilidade.

4.8 DISCUSSÃO

A interpretação dos resultados deste trabalho precisa partir de uma característica central do problema estudado: a previsão pluviométrica é, por natureza, uma tarefa difícil, e essa dificuldade cresce à medida que o horizonte de previsão se afasta do instante presente. Diferentemente de séries mais regulares, a precipitação diária apresenta alta variabilidade temporal, forte intermitência, grande frequência de valores nulos e ocorrência esporádica de eventos intensos. Em termos práticos, isso significa que o sistema a ser previsto combina longos períodos de baixa atividade com mudanças abruptas de comportamento, o que torna desafiadora tanto a estimação pontual quanto a construção de intervalos preditivos informativos.

Essa dificuldade estrutural ajuda a contextualizar os resultados obtidos. Mesmo as combinações que se mostraram mais competitivas não produziram um cenário de desempenho uniformemente alto em todos os horizontes. Em outras palavras, os resultados não devem ser lidos apenas como um *ranking* entre modelos e métodos conformes, mas também como expressão dos limites impostos por um fenômeno naturalmente incerto. Esse ponto se torna ainda mais importante nos horizontes mais distantes, nos quais a incerteza acumulada sobre a evolução futura da série passa a ser maior. À medida que o horizonte aumenta, torna-se mais difícil preservar simultaneamente boa acurácia pontual, cobertura próxima ao nível nominal e intervalos estreitos. Por isso, em vez de buscar um método "ótimo" em sentido absoluto, a análise precisou se concentrar no tipo de compromisso que cada combinação conseguiu estabelecer entre essas dimensões.

É justamente nesse contexto que a avaliação dos métodos de *Conformal Prediction* precisou ser ampliada. Em uma leitura mais tradicional, a qualidade intervalar poderia ser julgada principalmente pela cobertura empírica e pela largura média dos intervalos. No entanto, ao longo da análise, tornou-se claro que essa abordagem era insuficiente para descrever adequadamente o comportamento dos métodos diante de uma série tão irregular. Por essa razão, a discussão passou a considerar não apenas a proximidade da cobertura em relação ao nível nominal de $0{,}90$ e a largura dos intervalos, mas também a capacidade de acompanhar picos, oscilações locais e o comportamento da série em diferentes regimes, especialmente em períodos secos e molhados.

Essa mudança de critério foi importante porque evitou uma interpretação excessivamente simplificada dos resultados. Em vários casos, coberturas muito acima de $0{,}90$ estiveram associadas a intervalos demasiadamente largos, isto é, a soluções mais conservadoras, porém menos úteis do ponto de vista prático. Em um problema já marcado por alta incerteza, ampliar demais os intervalos pode até preservar a cobertura, mas ao custo de reduzir fortemente a informatividade da previsão. Assim, a qualidade do método conforme passou a ser interpretada de forma mais exigente: não basta cobrir mais; é preciso cobrir de forma próxima ao alvo, sem perda excessiva de nitidez e, idealmente, com sensibilidade à dinâmica real da série.

Sob essa perspectiva, os resultados sugerem que não existe um único método de *Conformal Prediction* superior em todos os sentidos. O *Mondrian* permaneceu muito competitivo quando a análise se concentrou naquilo que pode ser chamado de *qualidade intervalar agregada*, isto é, no equilíbrio global entre cobertura adequada e larguras moderadas ao longo do conjunto de teste. Em especial, suas combinações com *SARIMAX* e, em alguns cenários, com *GRU*, mostraram desempenho bastante consistente, sobretudo nos horizontes de médio e longo prazo. Esse comportamento é particularmente relevante em um contexto no qual previsões mais distantes tendem naturalmente a exigir intervalos mais amplos; mesmo assim, o *Mondrian* conseguiu, em várias situações, preservar um equilíbrio razoável entre validade e informatividade.

Entretanto, essa não foi a única dimensão relevante observada nos experimentos. Quando a análise incorporou critérios mais finos, como a capacidade de responder a picos, oscilações abruptas e transições entre diferentes regimes de variabilidade, o papel do *CQR* passou a ganhar

maior destaque. Em vários trechos, especialmente na combinação com *LSTM*, o *CQR* apresentou comportamento visual mais aderente à dinâmica local da série, acompanhando melhor extremos e variações rápidas. Esse resultado faz sentido em um problema como o da precipitação, em que parte da dificuldade está justamente na ocorrência irregular de eventos intensos. Nesses casos, um método que se mostre mais adaptativo localmente pode se tornar particularmente valioso, mesmo quando não apresenta a menor largura média em termos globais.

Dessa forma, os resultados apontam para a existência de pelo menos duas noções complementares de qualidade do *Conformal Prediction*. A primeira corresponde à *qualidade intervalar clássica ou agregada*, centrada na relação entre cobertura próxima ao nível nominal e largura reduzida. A segunda pode ser entendida como uma *qualidade intervalar adaptativa*, mais relacionada à capacidade de o método responder à variabilidade local da série, acompanhar picos e representar de forma mais convincente mudanças entre períodos secos e chuvosos. No presente estudo, a primeira dimensão favoreceu com frequência o *Mondrian*, enquanto a segunda atribuiu maior relevância ao *CQR*, principalmente quando combinado ao *LSTM*.

Essa distinção ajuda a interpretar de forma mais fiel o comportamento observado nos diferentes horizontes. No curto prazo, em que parte da estrutura recente da série ainda está mais próxima do instante de previsão, combinações com *LSTM* mostraram-se particularmente fortes, com destaque especial para a combinação *LSTM* + ICP, que constitui o resultado mais relevante deste trabalho e merece discussão aprofundada.

Uma primeira hipótese para explicar essa superioridade está relacionada à arquitetura da célula de memória do *LSTM*. A presença de portas de controle explícitas, em especial a porta de esquecimento, permite que o modelo selecione de forma adaptativa quais informações passadas são relevantes para cada instante de previsão (Hochreiter; Schmidhuber, 1997). Em uma série pluviométrica com forte intermitência, essa capacidade de reter seletivamente padrões recentes é particularmente vantajosa no curto prazo, horizonte em que a estrutura local da série ainda carrega informação preditiva relevante e em que a proximidade temporal entre treino e teste favorece a generalização do modelo. Diferentemente da *GRU*, cuja estrutura mais compacta reduz o número de parâmetros mas também limita a granularidade do controle sobre o fluxo de informação, o *LSTM* parece conseguir representar com mais fidelidade as transições abruptas entre períodos secos e chuvosos que caracterizam a série de Brasília (Greff et al., 2017).

Uma segunda hipótese diz respeito à interação entre o modelo base e o método conforme. O ICP, por ser uma estratégia global baseada em um único quantil empírico dos resíduos de calibração, tende a produzir intervalos mais estreitos quando o modelo base já apresenta resíduos com baixa dispersão e distribuição relativamente homogênea (Angelopoulos; Bates, 2021). Nesse sentido, a combinação entre um preditor base com boa acurácia pontual e uma estratégia conforme simples pode ser mais eficiente do que combinar um método conforme mais complexo a um modelo base menos preciso. Quando o *LSTM* já captura bem a estrutura da série, a etapa de calibração conforme precisa corrigir menos, o que se traduz em intervalos mais estreitos e, portanto, mais informativos. Esse raciocínio também ajuda a explicar por que

o *Mondrian* e o *CQR*, apesar de introduzirem maior sofisticação na calibração, não superaram o ICP no curto prazo quando combinados ao *LSTM*: a complexidade adicional do método conforme não compensou o que já havia sido capturado pelo modelo base. Em termos práticos, esse resultado sugere que, em problemas de previsão pluviométrica de curto prazo, investir na qualidade do preditor base pode ser tão ou mais importante do que escolher uma estratégia conforme sofisticada (Angelopoulos; Bates, 2021). O *CQR* apareceu, ainda assim, como alternativa qualitativamente relevante pela melhor sensibilidade aos extremos, mesmo quando não apresentava a menor largura média. No médio prazo, o cenário se tornou mais equilibrado, com diferentes combinações competitivas a depender do peso atribuído aos critérios quantitativos e qualitativos. Já no longo prazo, a dificuldade do problema se torna ainda mais evidente: o aumento da incerteza futura favorece métodos que consigam preservar maior estabilidade intervalar, o que ajuda a explicar por que o *Mondrian*, especialmente com *SARIMAX*, voltou a apresentar vantagem mais clara em termos de eficiência intervalar global, sem que isso anulasse a relevância do *CQR* em critérios de adaptatividade.

Os resultados também podem ser interpretados à luz da discussão do Capítulo 2 sobre heterocedasticidade e presença de regimes distintos na série. A alternância entre períodos secos e chuvosos tende a produzir mudanças na dispersão dos erros ao longo do tempo, o que torna limitada uma avaliação baseada apenas em um único padrão global de comportamento. Ainda que este trabalho não tenha realizado uma análise formal separando explicitamente esses regimes, a diferença de comportamento entre *Mondrian* e *CQR* parece compatível com essa estrutura: o primeiro se mostrou forte na calibração intervalar globalmente equilibrada, enquanto o segundo apresentou maior sensibilidade a variações locais da incerteza. Em um problema cuja dificuldade decorre justamente da coexistência entre regularidade sazonal e eventos irregulares, essa diferença entre os métodos se torna ainda mais significativa.

As métricas pontuais ajudam a sustentar essa leitura, ainda que não constituam o foco principal do trabalho. *LSTM*, *GRU* e *SARIMAX* apresentaram, de forma consistente, MAE e RMSE inferiores aos do *Random Forest*, além de R^2 positivo, enquanto o *Random Forest* obteve R^2 negativo em todos os horizontes. Esse resultado merece atenção específica: ele indica que o modelo performou pior do que uma previsão ingênua baseada na média dos valores observados, o que é particularmente expressivo em séries com alta frequência de zeros e eventos esporádicos de precipitação intensa. Uma hipótese plausível é que a representação tabular das janelas temporais, na qual a estrutura sequencial é achatada em um vetor unidimensional, compromete a capacidade do modelo de capturar dependências temporais de longo alcance presentes na série. Além disso, o *Random Forest* não possui mecanismo interno para modelar a ordem temporal das observações, o que pode limitar sua expressividade diante de fenômenos com forte sazonalidade e dinâmica não estacionária como a precipitação diária (Bontempi; Taieb; Borgne, 2013). Nesse sentido, os resultados reforçam que a escolha da arquitetura preditiva tem impacto direto não apenas na acurácia pontual, mas também na qualidade final dos intervalos conformes gerados a partir dela, uma vez que um preditor base fraco tende a comprometer toda a cadeia de calibração conforme

subsequente.

O bom desempenho do *SARIMAX* em parte dos experimentos também merece atenção. Embora o Capítulo 2 tenha discutido suas limitações diante de relações fortemente não lineares, os resultados mostraram que um modelo estatístico clássico ainda pode ser bastante competitivo quando a série preserva estrutura temporal e sazonal relevante. Em especial nos horizontes mais distantes, em que a previsibilidade local se enfraquece e a incerteza cresce, a capacidade de capturar componentes estruturais mais estáveis da série pode continuar sendo vantajosa. Ao mesmo tempo, o destaque qualitativo do *LSTM* com *CQR* mostra que modelos mais flexíveis podem se tornar particularmente relevantes quando a avaliação valoriza a resposta a extremos e a adaptatividade local dos intervalos.

Em síntese, a principal conclusão desta seção não é a existência de um único método conforme superior em todos os critérios, mas sim a constatação de que a própria dificuldade da previsão pluviométrica exige uma leitura mais cuidadosa da qualidade intervalar. O resultado mais relevante do trabalho, a combinação *LSTM* + ICP no horizonte de curto prazo, ilustra que um preditor base de alta qualidade combinado a uma estratégia conforme simples pode superar alternativas mais sofisticadas, desde que o modelo base já capture bem a estrutura temporal da série. Em horizontes mais distantes, em que a previsibilidade se deteriora, o *Mondrian* se mostrou a estratégia mais consistente em termos de equilíbrio entre cobertura e largura dos intervalos, enquanto o *CQR* ganhou relevância nos critérios de adaptatividade local. Assim, a contribuição mais importante dos resultados não está em apontar um único vencedor absoluto, mas em mostrar que a avaliação de métodos de *Conformal Prediction* para precipitação precisa levar em conta, de forma explícita, a complexidade do fenômeno, a qualidade do preditor base e a deterioração natural da previsibilidade em horizontes mais distantes.

Cabe registrar, ainda, uma limitação do desenho experimental adotado. Embora a comparação realizada tenha permitido identificar tendências relevantes entre modelos base e estratégias de *Conformal Prediction*, o *setup* não foi completamente homogêneo entre todas as arquiteturas. Em particular, houve diferenças em escolhas como função de perda, função de ativação na camada de saída e pré-processamento da variável alvo, além de transformações aplicadas apenas a parte dos modelos. Essas diferenças não invalidam os resultados obtidos, mas indicam que comparações ainda mais rigorosas poderiam ser conduzidas em estudos futuros a partir de um protocolo experimental mais uniforme, controlando de forma sistemática aspectos como perdas, ativações, regularizações e transformações de entrada e saída.

5 CONCLUSÃO

Este trabalho investigou a aplicação de métodos de *Conformal Prediction* na construção de intervalos preditivos para previsão pluviométrica diária, considerando diferentes horizontes de previsão e diferentes famílias de modelos base. O problema estudado é particularmente relevante devido à natureza irregular, sazonal e altamente incerta da precipitação, em que a simples obtenção de previsões pontuais não é suficiente para apoiar decisões de forma confiável. Além disso, a própria dificuldade de prever eventos pluviométricos aumenta à medida que o horizonte de previsão se afasta do instante presente, o que torna especialmente desafiadora a obtenção simultânea de boa acurácia pontual, cobertura adequada e intervalos informativos. Nesse contexto, a proposta central do estudo foi comparar o comportamento de três estratégias de *Conformal Prediction*, ICP, *Mondrian Conformal Prediction* e CQR, quando combinadas com quatro modelos de previsão distintos: *SARIMAX*, *Random Forest*, *LSTM* e *GRU*.

Os experimentos foram conduzidos com base em uma série temporal real de precipitação da estação de Brasília, disponibilizada pelo INMET, após pré-processamento e transformação em problema supervisionado. A análise foi realizada para os horizontes $H = 1$, $H = 7$ e $H = 30$, avaliando tanto métricas pontuais quanto métricas associadas à qualidade dos intervalos preditivos. De forma geral, os resultados mostraram que *LSTM*, *GRU* e *SARIMAX* apresentaram desempenho superior ao *Random Forest*, tanto em previsão pontual quanto na construção dos intervalos conformes. Ao mesmo tempo, os resultados também evidenciaram que, em um problema tão incerto quanto a previsão pluviométrica, especialmente em horizontes mais distantes, nenhum método se impõe de forma absoluta em todos os critérios.

Uma das principais conclusões do trabalho é que a avaliação da qualidade de métodos conformes depende diretamente do critério adotado. Quando a análise se concentra no equilíbrio clássico entre cobertura empírica próxima ao nível nominal e largura moderada dos intervalos, o *Mondrian Conformal Prediction* se mostrou uma estratégia bastante consistente no conjunto dos experimentos, com destaque para combinações como *SARIMAX* + *Mondrian* e, em alguns cenários, *GRU* + *Mondrian*. Por outro lado, quando a avaliação incorpora critérios mais refinados, como a capacidade de acompanhar picos, oscilações locais e mudanças entre regimes de variabilidade, o *Conformalized Quantile Regression* passa a ganhar maior relevância, especialmente na combinação com *LSTM*. Em vários trechos, essa combinação apresentou maior aderência visual à dinâmica local da série, ainda que nem sempre produzisse os intervalos mais estreitos em termos globais.

Nesse sentido, o trabalho mostrou que há uma distinção importante entre, pelo menos, duas dimensões da qualidade intervalar. A primeira pode ser entendida como *qualidade intervalar agregada*, centrada no compromisso entre validade e informatividade ao longo de todo o conjunto de teste. A segunda corresponde a uma *qualidade intervalar adaptativa*, mais relacionada à capacidade do método de responder a variações locais da série, acompanhar extremos e refletir

melhor a incerteza em diferentes regimes, como períodos secos e chuvosos. No presente estudo, a primeira dimensão favoreceu com maior frequência o *Mondrian*, enquanto a segunda atribuiu maior protagonismo ao *CQR*, principalmente quando combinado ao *LSTM*. Assim, a principal contribuição dos resultados não está em apontar um único vencedor absoluto, mas em mostrar que diferentes métodos de *Conformal Prediction* favorecem aspectos distintos da qualidade intervalar em previsão pluviométrica.

Entre as contribuições do trabalho, destaca-se a realização de uma análise comparativa aplicada entre diferentes modelos base e diferentes estratégias conformes em um problema pluviométrico real, sob múltiplos horizontes de previsão. O estudo também evidenciou que a escolha do método de *Conformal Prediction* não pode ser dissociada da escolha do modelo base, já que a qualidade da previsão pontual influencia diretamente a utilidade final dos intervalos produzidos. Além disso, os resultados reforçaram que a crescente dificuldade de prever precipitação em horizontes mais longos deve ser incorporada à interpretação dos experimentos: em vez de se esperar desempenho elevado de forma uniforme, é mais realista analisar como cada combinação lida com o aumento da incerteza e com a deterioração natural da previsibilidade ao longo do tempo.

O bom desempenho do *SARIMAX* em parte dos experimentos também merece destaque. Embora se trate de um modelo estatístico clássico e linear, seus resultados mostraram que maior complexidade arquitetural não implica, necessariamente, melhor desempenho, sobretudo quando a série ainda preserva estrutura temporal e sazonal relevante. Por outro lado, o papel qualitativamente forte do *LSTM* com *CQR* mostra que modelos mais flexíveis podem se tornar particularmente relevantes quando a análise valoriza a adaptatividade local dos intervalos e a representação de extremos. Dessa forma, os achados do trabalho apontam menos para uma hierarquia rígida entre métodos e mais para uma relação entre tipo de modelo, tipo de método conforme e tipo de qualidade intervalar buscada.

Como limitação, destaca-se que os experimentos foram conduzidos com dados de uma única estação meteorológica e com um conjunto restrito de métodos conformes e arquiteturas preditivas, o que naturalmente limita a generalização dos resultados para outros contextos climáticos e operacionais. Além disso, o *setup* experimental não foi completamente homogêneo entre todos os modelos, havendo diferenças em aspectos como função de perda, função de ativação e transformações aplicadas à variável alvo, o que pode introduzir algum viés na comparação entre arquiteturas. Embora isso não invalide os resultados obtidos, indica que estudos futuros podem tornar essa comparação mais rigorosa por meio de protocolos experimentais mais uniformes. Por fim, embora a discussão tenha considerado a relevância de regimes secos e chuvosos, não foi realizada uma avaliação formal separando explicitamente esses regimes, o que poderia aprofundar a interpretação sobre o comportamento adaptativo dos métodos conformes.

Como perspectivas para trabalhos futuros, sugere-se ampliar a análise para outras localidades e séries pluviométricas com características climáticas distintas, permitindo verificar a robustez dos achados em cenários mais diversos. Também seria interessante investigar outras variantes de

Conformal Prediction voltadas a séries temporais, bem como explorar estratégias condicionais ou adaptativas mais refinadas para lidar com regimes distintos de precipitação. Também seria relevante adotar, em estudos futuros, protocolos experimentais mais homogêneos entre arquiteturas, controlando de forma mais sistemática escolhas como funções de perda, funções de ativação, regularizações e transformações de entrada e saída. Do ponto de vista da modelagem, futuras pesquisas podem incorporar arquiteturas mais recentes, como modelos baseados em atenção e *transformers*, além de abordagens híbridas que combinem modelos estatísticos e neurais. Outra direção promissora consiste em aprofundar a análise por regime hidrometeorológico, avaliando separadamente períodos secos, chuvosos e eventos extremos, o que permitiria examinar com mais rigor a diferença entre qualidade intervalar agregada e qualidade intervalar adaptativa. Por fim, também se mostra relevante investigar não apenas a qualidade estatística dos intervalos, mas sua utilidade operacional em aplicações concretas de gestão hídrica, monitoramento ambiental e apoio à decisão diante de eventos extremos. Uma direção conceitualmente relevante diz respeito à própria hipótese de *exchangeability* e à forma como ela é violada em séries com sazonalidade marcada. Em séries pluviométricas, observações pertencentes ao mesmo regime climático, como dois dias distintos dentro da estação chuvosa, podem ser consideradas mais permutáveis entre si do que observações de regimes opostos. Essa estrutura de *quasi-exchangeability* condicional ao regime sugere que abordagens conformes segmentadas por estação ou por regime hidrometeorológico poderiam produzir intervalos com validade empírica mais robusta do que abordagens globais, ao calibrar separadamente os escores de não conformidade dentro de cada regime. Investigar essa hipótese formalmente, combinando segmentação sazonal com estratégias como o *Mondrian Conformal Prediction* aplicado por regime em vez de por nível de $\hat{y}$, constitui uma extensão natural e metodologicamente rica deste trabalho.

REFERÊNCIAS

ALHARBI, F. R.; CSALA, D. A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) forecasting model-based time series approach. **Inventions**, MDPI, Basel, v. 7, n. 4, p. 94, 2022.

ANGELOPOULOS, A. N.; BATES, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. **arXiv preprint**, 2021. ArXiv:2107.07511. Disponível em: https://arxiv.org/abs/2107.07511. Acesso em: 16 nov. 2025.

AUER, A. et al. Conformal prediction for time series with modern hopfield networks. In: **Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)**. [S.l.: s.n.], 2023. ArXiv:2303.12783.

BONTEMPI, G.; TAIEB, S. B.; BORGNE, Y.-A. L. Machine learning strategies for time series forecasting. In: AUFAURE, M.; ZIMÁNYI, E. (Ed.). **Business Intelligence**: Second european summer school, ebiss 2012, brussels, belgium, july 15–21, 2012, tutorial lectures. Berlin: Springer, 2013, (Lecture Notes in Business Information Processing, v. 138). p. 62–77. DOI: 10.1007/978-3-642-36318-4_3.

BOSTRÖM, H.; JOHANSSON, U.; LÖFSTRÖM, T. Mondrian conformal predictive distributions. In: CONFORMAL AND PROBABILISTIC PREDICTION AND APPLICATIONS. **Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications**. Online: PMLR, 2021. (Proceedings of Machine Learning Research, v. 152), p. 24–38. Disponível em: https://proceedings.mlr.press/v152/bostrom21a.html. Acesso em: 16 nov. 2025.

BOX, G. E. P. et al. **Time Series Analysis: Forecasting and Control**. 5. ed. Hoboken: John Wiley & Sons, 2015.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CHO, K. et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha: Association for Computational Linguistics, 2014. p. 1724–1734.

GOEHRY, B. et al. Random forests for time series: A statistical perspective. **REVSTAT – Statistical Journal**, v. 21, n. 2, p. 283–302, 2023.

GREFF, K. et al. LSTM: A search space odyssey. **IEEE Transactions on Neural Networks and Learning Systems**, v. 28, n. 10, p. 2222–2232, 2017.

HEWAMALAGE, H.; BERGMEIR, C.; BANDARA, K. Recurrent neural networks for time series forecasting: Current status and future directions. **International Journal of Forecasting**, v. 37, n. 1, p. 388–427, 2021.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997.

HU, X. et al. Conformalized temporal convolutional quantile regression networks for wind power interval forecasting. **Energy**, 2022.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. 3. ed. Melbourne: OTexts, 2021. ISBN: 9780987507136. Disponível em: https://otexts.com/fpp3/. Acesso em: 1 nov. 2025.

IWAKIN, O.; MOAZENI, F. Improving urban water demand forecast using conformal prediction-based hybrid machine learning models. **Journal of Water Process Engineering**, v. 58, p. 104721, 2024.

JENSEN, V.; BIANCHI, F. M.; ANFINSEN, S. N. Ensemble conformalized quantile regression for probabilistic time series forecasting. **IEEE Transactions on Neural Networks and Learning Systems**, PP, n. 99, p. 1–12, 2022.

KRATZERT, F. et al. Rainfall–runoff modelling using long short-term memory (LSTM) networks. **Hydrology and Earth System Sciences**, v. 22, n. 11, p. 6005–6022, 2018.

LEE, J.; XU, C.; XIE, Y. **Transformer Conformal Prediction for Time Series**. 2024. ArXiv preprint arXiv:2406.05332.

LIM, B.; ZOHREN, S. Time-series forecasting with deep learning: A survey. **Philosophical Transactions of the Royal Society A**, v. 379, n. 2194, p. 20200209, 2021.

MEINSHAUSEN, N. Quantile regression forests. **Journal of Machine Learning Research**, v. 7, p. 983–999, 2006.

MORTIER, T. et al. **Valid Prediction Intervals for Weather Forecasting with Conformal Prediction**. 2025. Preprint apresentado na EGU General Assembly 2025. Doi:10.5194/egusphere-egu25-17783.

MULLA, S.; PANDE, C. B.; SINGH, S. K. Times series forecasting of monthly rainfall using Seasonal Auto Regressive Integrated Moving Average with EXogenous Variables (SARIMAX) model. **Water Resources Management**, Cham, v. 38, n. 6, p. 1825–1846, 2024.

PAPACHARALAMPOUS, G.; TYRALIS, H. A review of machine learning concepts and methods for addressing challenges in probabilistic hydrological post-processing and forecasting. **Frontiers in Water**, Lausanne, v. 4, p. 961954, 2022. Disponível em: https://www.frontiersin.org/articles/10.3389/frwa.2022.961954. Acesso em: 16 nov. 2025.

ROMANO, Y.; PATTERSON, E.; CANDÈS, E. J. Conformalized quantile regression. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS. **Advances in Neural Information Processing Systems**. Vancouver, 2019. p. 3538–3548. NeurIPS 2019. Disponível em: https://papers.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html. Acesso em: 16 nov. 2025.

SHAFER, G.; VOVK, V. A tutorial on conformal prediction. **Journal of Machine Learning Research**, Cambridge, MA, v. 9, p. 371–421, 2008. Disponível em: https://jmlr.org/papers/v9/shafer08a.html. Acesso em: 16 nov. 2025.

STANKEVIČIŪTĖ, K.; ALAA, A. M.; SCHAAR, M. van der. Conformal time-series forecasting. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS.

Advances in Neural Information Processing Systems. Red Hook, NY, 2021. p. 6216–6228. NeurIPS 2021. Disponível em: https://proceedings.neurips.cc/paper/2021/hash/312f1ba2a72318edaaa995a67835fad5-Abstract.html. Acesso em: 16 nov. 2025.

Statsmodels Development Team. **statsmodels.tsa.statespace.sarimax.SARIMAX: Documentation**. 2024. Documentação online. Disponível em: https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html. Acesso em: 16 nov. 2025.

TYRALIS, H.; PAPACHARALAMPOUS, G.; LANGOUSIS, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. **Water**, v. 11, n. 5, p. 910, 2019.

TYRALIS, H.; PAPACHARALAMPOUS, G. A. **A review of probabilistic forecasting and prediction with machine learning**. 2022. ArXiv preprint arXiv:2209.08307. DOI: 10.48550/arXiv.2209.08307. Disponível em: https://arxiv.org/abs/2209.08307. Acesso em: 1 nov. 2025.

VOVK, V.; GAMMERMAN, A.; SHAFER, G. **Algorithmic learning in a random world**. Berlin: Springer-Verlag, 2005. DOI: 10.1007/b106715.

XU, C.; XIE, Y. Conformal prediction interval for dynamic time-series. In: **Proceedings of the 38th International Conference on Machine Learning**. [S.l.]: PMLR, 2021. (Proceedings of Machine Learning Research, v. 139), p. 11559–11569.