ISSN 2675-3359
Revista Mineira de Recursos Hídricos
v. 7, 2026
RMRH
Instituto Mineiro de Gestão das Águas - rmrh.igam.mg.gov.br
DOI: https://doi.org/10.59824/rmrh.v7.358

PREVISÃO MENSAL DE VAZÃO PARA O RIO JEQUITINHONHA: uma abordagem por aprendizado de máquina

Monthly Streamflow Forecasting in the Jequitinhonha River using the CatBoost machine learning algorithm

Welson de Avelar Soares Filho[1]*
Paula Roberta Souza Carvalho[2]
Celso Bandeira de Melo Ribeiro[3]
Leonardo Goliatt da Fonseca[4]

RESUMO

A previsão de vazões é decisiva para mitigar riscos socioeconômicos e otimizar a operação de recursos hídricos, cuja variabilidade afeta o fornecimento de energia e abastecimento. Tendo em vista a lacuna de estudos nacionais sobre uso de aprendizado de máquina, este estudo utilizou o modelo *CatBoost* para gerar previsões mensais confiáveis para o rio Jequitinhonha. Foram empregados dados diários de vazão para treinamento do modelo e realizada a previsão para 31 dias (1 mês), sendo os resultados avaliados segundo as métricas MAPE, RMSE e KGE, além de intervalos de previsão nos percentis de 5 ° e 95°. O modelo obteve MAPE de aproximadamente 15,7 %, RMSE de 26,7 m³ /s e KGE 0,36. Foi melhorada a KGE *a posteriori* para 0,54 ao otimizar-se a quantidade de observações defasadas (*lags*) para treinamento, e o

[1] Mestrado em Modelagem Computacional pela Universidade Federal de Juiz de Fora (UFJF). Graduação em Ciência da Computação pela UFJF. Analista em Tecnologia da Informação no Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de MG - IF Sudeste MG/Campus JF. Pesquisador em aprendizado de máquina (machine learning) e aprendizado profundo (deep learning) com aplicação em Séries Temporais e Recursos Hídricos – e-mail: welson.avelar@ifsudestemg.edu.br * Autor correspondente
[2] Mestrado em Engenharia Civil pela UFJF. Especialista em Engenharia de Estruturas pela Pontifícia Universidade Católica de Minas Gerais (PUCMINAS). Graduação em Engenharia Ambiental e Engenharia Civil pela Universidade FUMEC. Atualmente é discente do doutorado em Engenharia Civil da UFJF – e-mail: paulasouza.carvalho@estudante.ufjf.br
[3] Pós-doutorado no Texas AM University. Doutorado pelo DEA/UFV. Mestre pela COPPE/UFRJ. Graduado em Engenheira Civil pela Universidade Federal de Juiz de Fora (Fac.Eng.UFJF). Professor Titular atuando no Departamento de Engenharia Sanitária e Ambiental (ESA) da UFJF. Professor permanente e orientador de mestrado junto ao Programa de Pós-graduação em Engenharia Civil (PEC) da UFJF e membro do Colegiado do Curso PEC/UFJF. Coordenador da Regional Sudeste da Associação Brasileira de Recursos Hídricos ABRHidro. Representante titular da UFJF no Comitê de Integração da Bacia Hidrográfica do Rio Paraíba do Sul (CEIVAP) – e-mail: celso.bandeira@ufjf.br
[4] Doutorado em Modelagem Computacional pelo Laboratório Nacional de Computação Científica pela UFJF. Graduação em Engenharia Civil pela UFJF – e-mail: leonardo.goliatt@ufjf.br

intervalo de previsão proposto, de 90%, que capturou os valores reais observados. O *CatBoost* demonstrou aptidão para representar o regime de vazão do rio Jequitinhonha, mas ganhos adicionais exigem séries mais longas e otimização de hiperparâmetros (profundidade da árvore criada, quantidade de nós por nível, entre outros) para elevar correlação e reduzir viés, consolidando-se como ferramenta promissora no apoio à gestão hídrica regional.

Palavras-chave: previsão de vazão; aprendizado de máquina; recursos hídricos.

ABSTRACT

Streamflow forecasting plays a crucial role in mitigating socioeconomic risks and optimizing water resource operations, as hydrological variability directly affects energy generation and water supply. Addressing the limited number of national studies employing machine learning techniques for this purpose, this study applies the CatBoost algorithm to produce reliable monthly streamflow forecasts for the Jequitinhonha River. The model was trained using daily discharge data and used to generate a 31-day ahead forecast. Model performance was assessed using MAPE, RMSE, and KGE metrics, along with prediction intervals at the 5th and 95th percentiles. The initial configuration yielded a MAPE of approximately 15.7%, an RMSE of 26.7 m³/s, and a KGE of 0.36. Post-hoc optimization of lagged input variables improved the KGE to 0.54. The proposed 90% prediction interval successfully encompassed the observed streamflow values. CatBoost proved capable of representing the hydrological regime of the Jequitinhonha River; however, further improvements depend on the availability of longer time series and systematic hyperparameter tuning (e.g., tree depth and number of nodes per level). These adjustments may enhance correlation, reduce bias, and strengthen the potential of CatBoost as a promising tool to support regional water-resource management.

***Keywords*:** *streamflow forecasting; machine learning; water resources.*

Data de submissão: 07/05/2025

Data de aprovação: 03/12/2025

1 INTRODUÇÃO

A gestão dos recursos hídricos desempenha um papel crucial nas políticas públicas. Nos âmbitos socioeconômico, cultural e de saúde pública, conhecer a dinâmica dos recursos hídricos e entender como fatores externos impactam seu comportamento é de grande importância para os gestores públicos. A compreensão desses aspectos permite uma melhor tomada de decisões, garantindo a sustentabilidade dos recursos, a segurança hídrica e o bem-

estar da população (Agência Nacional de Águas e Saneamento Básico - ANA, 2024; Ballarin *et al*., 2023; Gesualdo, 2021).

Neste sentido, prever a vazão de rios é um componente essencial na gestão de recursos hídricos e em diversos setores da sociedade, tais como: a operação de reservatórios, mitigação de desastres naturais, irrigação, abastecimento público, dentre outros (Silva *et al.*, 2020). Segundo o Balanço Energético Nacional (2023), ano-base 2022, divulgado pelo Ministério de Minas e Energia, a matriz hidrelétrica representa cerca de 64% da oferta interna total de geração de energia elétrica (Rio de Janeiro, 2023). Desta forma, a previsão da vazão de rios que abastecem os reservatórios das hidrelétricas tem importância econômica.

Na temática do bem-estar populacional é necessário considerar os eventos climáticos extremos. De acordo com os últimos desastres ocorridos nos estados de Minas Gerais, Rio de Janeiro, Espírito Santo, São Paulo, Bahia e mais recentemente, em maio de 2024, a tragédia no Rio Grande do Sul, estes eventos trouxeram não apenas perda econômica como também perda de vidas. Estes fatos apontam para a necessidade de mais estudos de previsão de vazão (Laforé; Figueiredo; Malmann, 2023; G1, 2022; Lopes, 2024; G1, 2023; BBC News Brasil, 2021, 2024).

O Relatório Diagnóstico dos Afluentes do Alto Jequitinhonha - JQ1, relata diversos problemas relacionados aos recursos hídricos na bacia hidrográfica do Jequitinhonha, a saber: escassez de água, degradação ambiental, assoreamento, queimadas, exploração mineral clandestina, monocultura de eucalipto, falta de controles ambientais, dentre outros (Gama Engenharia Recursos Hídricos, 2013).

Neste sentido, os modelos de aprendizagem de máquina se apresentam com grande potencial para contribuir para o comportamento hidrológico nessa importante bacia hidrográfica mineira.

As pesquisas relacionadas ao aprendizado de máquina compartilham um interesse convergente em aplicar redes neurais e metodologias inteligentes a problemas de previsão de vazão, cada qual em cenários ambientais específicos. A aplicação de redes neurais ou variações híbridas tende a trazer ganhos de desempenho, sobretudo quando parâmetros são devidamente otimizados e os modelos são treinados com séries históricas consistentes. Ainda que cada abordagem apresente limitações pontuais, como a dificuldade em capturar picos de vazão diária

ou em explicar processos hidrológicos de forma detalhada, os resultados mostram-se encorajadores.

Este trabalho teve como objetivo o desenvolvimento e a aplicação de um modelo de aprendizado de máquina voltado à previsão de vazão no rio Jequitinhonha, com o propósito de contribuir para a redução da lacuna existente na literatura quanto ao uso dessas técnicas nessa área específica.

2 REVISÃO DE LITERATURA

Atualmente, a utilização de linguagens de programação vem trazendo novas perspectivas e mostrando um grande potencial, com as técnicas de aprendizado de máquina, para aplicações no gerenciamento dos recursos hídricos. Diversos trabalhos vêm sendo desenvolvidos no Brasil e no mundo, trazendo contribuições e aplicações em planejamento e gestão dos recursos hídricos, dentre os quais trouxemos alguns que possuem relação com este trabalho.

Vilanova, Zanetti e Cecílio (2020) se empenharam em avaliar a calibração e a transferência de redes neurais artificiais para simular vazões em sub-bacias da Mata Atlântica brasileira, com ênfase na bacia do rio Itapemirim. Eles estruturaram redes do tipo *Multilayer Perceptron* (MLP), com o intuito de estimar vazões diárias em três sub-bacias de diferentes dimensões, analisando simultaneamente a possibilidade de aplicar esse modelo a outras regiões do mesmo sistema hidrográfico. Durante 32 anos, reuniram-se dados diários de chuva e vazão em 12 sub-bacias, evidenciando a capacidade das redes em reproduzir as vazões com acurácia satisfatória, embora as vazões máximas apresentassem limitações, possivelmente vinculadas à escala temporal e à heterogeneidade hidrológica. O estudo ressaltou que sub-bacias pequenas e grandes, por suas características específicas, requerem ajustes particulares, indicando que modelos localmente calibrados podem oferecer maior solidez.

No trabalho desenvolvido por Gomaa *et al.* (2023), foi investigado o emprego de algoritmos híbridos de aprendizagem de máquina para prever a vazão diária no reservatório de Três Marias, dentro da bacia hidrográfica do rio São Francisco. A disponibilidade de dados de chuva do *Tropical Rainfall Measuring Mission* (TRMM), aliada às observações de vazão ao

longo de 22 anos, possibilitou o teste de diferentes abordagens, como *Gaussian Radial Basis Function Neural Network* (GRNN), *Gaussian Process Regression* (GPR) e MLP otimizado por *Particle Swarm Optimization* (PSO). Adicionalmente, a integração da Decomposição Empírica de Modo (EMD em inglês) e Transformada de Hilbert-Huang (HHT em inglês), combinada à MLP-PSO, gerou melhorias significativas na previsão de vazões, com destaque para a fase de testes e elevados índices de *Nash–Sutcliffe Efficiency* (NSE). O modelo resultante, batizado de MLP-PSO-EMD, atingiu resultados precisos, revelando a robustez de técnicas de decomposição de sinal e de otimização por enxame de partículas.

O estudo de Nogueira Filho *et al.* (2022) centrou-se na modelagem de fluxos em bacias hidrográficas não monitoradas do Ceará, empregando redes recorrentes *Long Short-Term Memory* (LSTM) para comparar seu desempenho com redes *Feedforward Neural Network* (FFNN) tradicionais e com o modelo *Soil Moisture Accounting Procedure* (SMAP). A possibilidade de incluir atrasos (*lags*) de chuva e vazão nas arquiteturas FFNN mostrou-se promissora, chegando a superar até mesmo a LSTM em alguns cenários, ao apresentar índices de eficiência superiores. Essa descoberta realçou a relevância de explorar a memória de curto prazo em regiões semiáridas, onde a resposta hídrica é rápida e a acumulação de água no solo ocorre de maneira limitada. Apesar de a LSTM ter oferecido bons resultados, a FFNN com *lags* de vazão, denominada FFNN-2, obteve desempenho superior, indicando que a complexidade de uma rede recorrente nem sempre é indispensável. Para as condições semiáridas, a inclusão de múltiplos atrasos de entrada mostrou-se crucial, ressaltando a importância de se considerar a natureza efêmera dos regimes de chuva e escoamento.

Rodrigues *et al.* (2021) dedicou-se à análise hidrológica na bacia do rio Manuel Alves da Natividade, no bioma Cerrado, com vistas à gestão dos recursos hídricos. Foram aplicados dois modelos distintos, o *Soil and Water Assessment Tool* (SWAT), orientado por processos físicos, e uma Rede Neural Artificial (RNA), que se baseia unicamente em dados, ambos calibrados e validados com dados de 1986 a 2005. A RNA, ao incorporar também *lags* de um dia de vazão, mostrou-se mais eficaz em termos de NSE, obtendo índice de 0,91, superando ligeiramente o SWAT. Apesar disso, o modelo SWAT ofereceu maior compreensão dos processos hidrológicos subjacentes, o que pode ser fundamental em investigações sobre mudanças de uso do solo ou variações climáticas futuras. Desta forma, a decisão sobre qual

estratégia adotar, se um modelo baseado em dados ou um baseado nos processos físicos, depende do equilíbrio entre a necessidade de maior interpretabilidade do sistema e a busca pela precisão do ajuste quantitativo.

Por fim, Ribeiro, Reynoso-Meza e Siqueira (2020) propuseram o emprego de *Extreme Learning Machine* (ELM) e redes do tipo *Echo State Networks* (ESN) para prever vazões em cinco usinas hidrelétricas, examinando janelas de 1935 a 2010. A abordagem incluiu um novo método de treinamento para ESN, permitindo a utilização de *bagging*, além de adotar otimização multi-objetivo (MOB em inglês) para ajustar pesos de modelos-base em um *ensemble*. Com essa metodologia, foi possível equilibrar viés e variância, resultando em previsões mais robustas e adaptadas às dinâmicas de cada sistema hídrico analisado. Os testes abrangeram horizontes de previsão de 1, 3, 6 e 12 meses, comparando os modelos individuais (ELM, ESN) com as versões em *ensemble*, além de modelos estatísticos como ARIMA e SARIMA (ARIMA Sazonal). Verificou-se que as formas otimizadas, denominadas ESN-MOB e ELM-MOB, obtiveram melhor desempenho geral, notadamente em projeções de maior prazo, em que a incerteza tende a se ampliar. Essa constatação reforçou a utilidade de técnicas de *ensemble* e otimização multi-objetivo para contornar a complexidade intrínseca à previsão hidrológica.

Há um consenso de que as redes neurais podem constituir um ferramental bastante útil para aprimorar o planejamento e gestão de recursos hídricos, principalmente em regiões onde as estações de monitoramento são escassas.

Estes estudos também sublinham a importância de se combinar dados observacionais de boa qualidade com métodos capazes de lidar com a variabilidade hidrológica. Na bacia do rio Itapemirim (Vilanova; Zanetti; Cecílio, 2020), por exemplo, a calibragem específica para sub-bacias de tamanhos diversos envolveu lidar com diferentes comportamentos de deflúvio, intensificados em períodos de chuva intensa. Em Três Marias (Gomaa *et al.*, 2023), a adoção de dados de chuva via satélite potencializou a modelagem, mas demandou integração de técnicas avançadas, como EMD-HHT, para aperfeiçoar a previsibilidade. Nas bacias semiáridas do Ceará, a busca por estratégias simples, mas eficientes, ilustra a necessidade de processos de memória de curto prazo, haja vista o regime de chuvas irregular da região. Já no Cerrado, a possibilidade de comparar o SWAT com RNA mostra como modelos baseados em

dados podem suprir lacunas de conhecimento, mesmo que não elucidem todas as interações hidrológicas.

As limitações mais frequentes referem-se à representação de eventos extremos e à extrapolação para cenários distintos. Em contrapartida, a implementação de algoritmos com memória interna, como LSTM, ou de híbridos com processamento de sinais, como EMD-HHT, apresentam caminhos promissores na redução de erros sistemáticos.

3 MATERIAIS E MÉTODOS

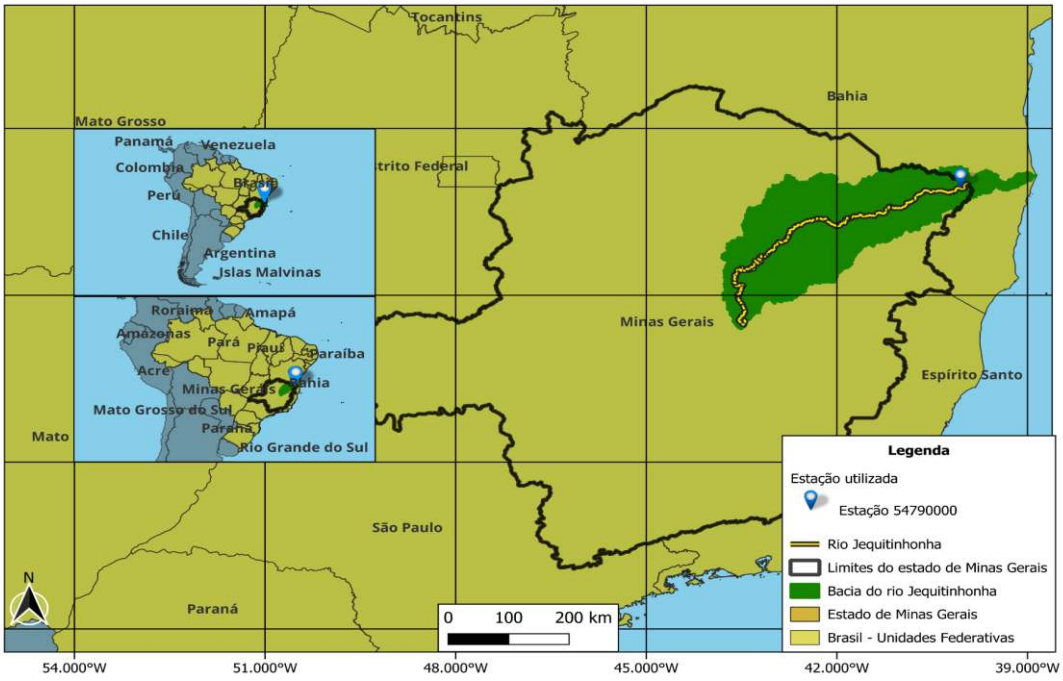
3.1 Área de estudo

O rio Jequitinhonha, com sua bacia hidrográfica abrangendo cerca de 6 milhões e 500 mil hectares, é uma peça fundamental na paisagem e na vida socioeconômica de Minas Gerais. Seu curso percorre 82 municípios mineiros, abrigando uma população de aproximadamente 939 mil habitantes que dependem diretamente de suas águas (Minas Gerais, 2025 a, b, c).

A Mata Atlântica é o bioma predominante e a atividade econômica principal na região é a agropecuária, ocupando uma área considerável de 2,6 milhões de hectares, em contraste com os 3,8 milhões de hectares de floresta. Essa dinâmica evidencia a pressão exercida sobre o ecossistema, demandando um olhar atento para a gestão sustentável dos recursos naturais (Brasil, 2025).

O Mapa 1 apresenta a localização da estação fluviométrica selecionada (UHE Itapebi montante 1) na bacia hidrográfica do rio Jequitinhonha - área de estudo - e a Tabela 1 os seus respectivos dados.

Mapa 1– Localização da estação fluviométrica cód. 54790000 (UHE Itabepbi montante 1), na bacia hidrográfica do rio Jequitinhonha: área de estudo



Fonte: Elaborado pelos autores (2025)

Tabela 1 – Estação selecionada na bacia hidrográfica do rio Jequitinhonha

Código	Nome	Município	Tipo	Latitude	Longitude
54790000	**UHE Itapebi montante 1**	**Salto da Divisa**	**Fluviométri-ca**	**-16,08**	**-40,0521**

Fonte: Elaborado pelos autores (2025)

3.2 Modelo *CatBoost*

Neste trabalho foi aplicado o modelo *Categorical Boosting*, usualmente referido por *CatBoost*. Este, tem o funcionamento baseado no princípio da construção de modelos fracos (*weak learners* em inglês) de árvores de decisão (*decision tree* em inglês) de forma sequencial, onde cada árvore sucessiva é treinada para corrigir os erros da árvore anterior. A estratégia do

CatBoost, no entanto, tem um nome: *Ordered Boosting*. A construção das árvores se dá de maneira sequencial, porém não se usa todos os dados disponíveis para isso. Os dados de treinamento são ordenados de maneira aleatória e apenas partições destes dados são utilizados no processo. Por trabalhar sempre com uma amostra dos dados de treinamento, e a apresentação aleatória destes dados ao modelo, o *CatBoost* tem resiliência a sobreajuste (*overfitting* em inglês). Contudo, parâmetros que realizam ajustes nas árvores de decisão também estão presentes e o programador tem controle sobre eles.

Os hiperparâmetros para execução do modelo foram escolhidos (Tabela 2) para que o mesmo executasse sem escrever (*allow_writing_files*) cada passo de treinamento em disco durante a execução, para salvar espaço em disco e evitar *overhead* dessa escrita desnecessária. O *thread_count* é para executar em paralelo, acelerando o treinamento do modelo. O hiperparâmetro *verbose* é para não escrever cada passo de treinamento na saída de texto do sistema e *random_seed* é para garantir reprodutibilidade dos resultados em todas as execuções do modelo. Por fim, *has_time* usa a ordem temporal dos dados de entrada, não executando permutações aleatórias durante os estágios de transformação de variáveis categóricas em variáveis numéricas e durante a escolha da estrutura da árvore (Catboost, 2025).

Tabela 2 – Hiperparâmetros do modelo *CatBoost*

Hiperparâmetro	**Valor atribuído**
random_seed	1989
verbose	False
allow_writing_files	False
has_time	True
thread_count	**8**

Fonte: Elaborado pelos autores (2025)

3.3 Métricas utilizadas

Para avaliar o desempenho do modelo de previsão utilizado, foram adotadas três métricas: MAPE, RMSE e KGE. A escolha dessas métricas baseia-se na necessidade de uma avaliação abrangente que considere diferentes aspectos da qualidade das previsões, como precisão, erro médio, correlação, variabilidade e viés. Nas formulações que se seguem, o termo "O_i" é o valor real observado e o termo "P_i" é o valor previsto pelo modelo.

- MAPE: Em inglês, *Mean Absolute Percentage Error* (Erro Percentual Absoluto Médio) é uma métrica amplamente utilizada para medir a precisão das previsões em termos percentuais. O algoritmo calcula a média das diferenças absolutas entre os valores observados e previstos, normalizadas pelos valores observados. O bom da métrica MAPE é a sua facilidade de interpretação, já que expressa o erro em uma faixa de 0 a 1 (sendo o valor 1 correspondente a 100%), e por ser "livre de escala", ou seja, independente da escala dos dados, tornando os resultados comparáveis entre diferentes séries temporais e modelos. O que se deseja aqui é minimizar, portanto, quanto mais próximo de 0 melhor (Hyndman; Athanasopoulos, 2021).

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{O_i - P_i}{O_i} \right|$$

- RMSE: O *Root Mean Squared Error* (Raiz do Erro Quadrático Médio) é uma métrica que quantifica a precisão de um modelo de previsão ao calcular, para cada ponto da série, a diferença entre o valor previsto e o valor observado, elevá-la ao quadrado, obter a média desses quadrados e extrair a raiz quadrada do resultado. Por estar expresso na mesma unidade da variável analisada, o RMSE oferece uma interpretação direta sobre o erro médio do modelo; entretanto, comparações entre diferentes séries só são válidas quando elas compartilham a mesma escala. Também o que se busca aqui é minimizar esta métrica, por isso quanto mais próximo de 0 melhor (Hyndman; Athanasopoulos, 2021).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}$$

- KGE: A *Kling-Gupta Efficiency* (Eficiência de Kling-Gupta, em tradução livre) fornece uma avaliação integrada do desempenho do modelo, considerando simultaneamente três componentes: correlação (r), variabilidade relativa (α) e viés (β) entre os valores previstos e observados. A KGE é uma métrica robusta que combina esses três fatores de forma equilibrada, fornecendo um entendimento geral da qualidade das previsões. Essa métrica é especialmente útil em estudos hidrológicos, pois tem capacidade de capturar a complexidade das relações entre variáveis hidrológicas de maneira mais eficaz do que métricas tradicionais focadas em um único aspecto. Quanto mais próximo de 1, melhor o desempenho do modelo (Gupta *et al.*, 2009).

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

3.3 Base de dados utilizados

A aquisição dos dados utilizados nesta pesquisa foi realizada com uso da biblioteca *hydrobr* (Carvalho, 2020). Esta biblioteca permitiu a listagem de todas as estações hidrométricas disponíveis de vazão, telemétrica ou convencional. Após a identificação e seleção da estação de interesse, cujo código estava disponível na base de dados da Agência Nacional de Águas e Saneamento Básico - ANA, desenvolveu-se um conjunto próprio de funções para automatizar o processo de extração. Essas funções permitiram o *download* dos dados referentes ao período especificado diretamente do *webservice* fornecido pela ANA.

A estação selecionada para estudo foi a de código 54790000, tipo telemétrica, e o período pesquisado compreende 1º de janeiro de 2020 a 31 de outubro de 2024, o que corresponde a 4 anos e 10 meses. Como a granularidade dos dados estava horária e o que se almejava era previsão diária, foi feito um primeiro ajuste quanto a este aspecto. Calculou-se a média de vazão para o dia e isso gerou uma base de dados com 1766 registros. Cabe destacar

que todos os dados de vazão obtidos no *webservice* da ANA já estavam na escala de metros cúbicos por segundo (m³/s). A Tabela 1 apresenta todas as informações acerca da estação.

3.4 Procedimentos metodológicos

Após a análise inicial, verificou-se a presença de dados nulos na série temporal e para preencher estes dados faltantes realizou-se imputação por interpolação. Para isso, usou-se informações dos dados mais próximos onde estava faltando a fim de que a série temporal ficasse totalmente completa. Para tal procedimento (interpolação) foi empregada a biblioteca *sktime* (Löning *et al.*, 2019).

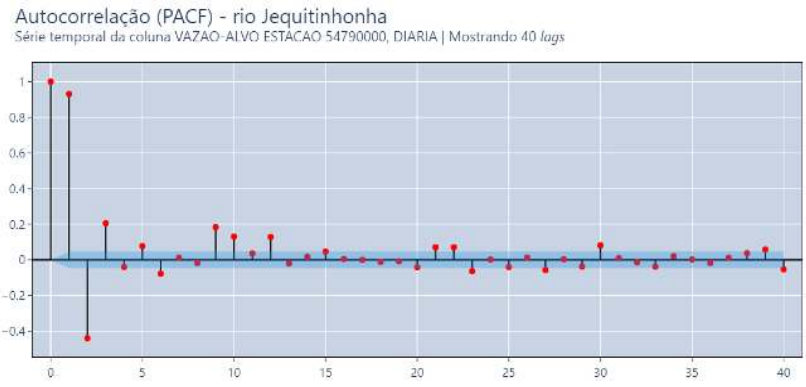
Com a série temporal sem lacunas, procedeu-se com a análise de autocorrelação parcial (PACF no inglês) para determinar quantas observações passadas (*lags*) utilizar como dados de entrada para realizar as previsões (GRÁFICO 1). Este procedimento é normalmente aplicado para modelos estatísticos, tais como *Holt-Winters* e ARIMA, mas esta informação pode ser útil mesmo para um modelo tipo *ensemble* como o *CatBoost*. Baseando-se no gráfico 2, chegou-se à conclusão de que 3 *lags* de vazões bastariam para tal procedimento.

Ainda no Gráfico 1, pode-se depreender que a amplitude da série é elevada, passando de 3500 m³/s em janeiro de 2022 e valores mínimos próximos a 30 m³/s. A Tabela 3 apresenta as estatísticas.

Gráfico 1– Série Temporal da estação, sem lacunas

Fonte: Elaborado pelos autores (2025)

Gráfico 2 – Gráfico PACF da série temporal



Fonte: Elaborado pelos autores (2025)

Tabela 3 – Estatísticas da massa de dados utilizados

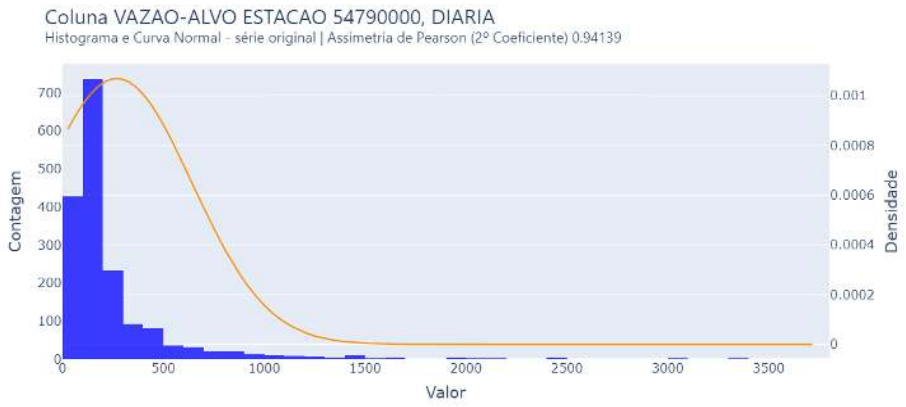
Nº. registros	Média	Desvio-Padrão	Mín.	Mediana	Máx.
1766	**269,76 m³/s**	**373,20 m³/s**	**26,87 m³/s**	**152,68 m³/s**	**3716,65 m³/s**

Fonte: Elaborado pelos autores (2025)

Foi realizada também análise quanto à distribuição dos dados de vazão (GRÁFICO 3). Calculando o Segundo Coeficiente de Assimetria de Pearson e seguindo o disposto por Crespo (2009), o valor da assimetria dos dados de vazão (0,94139) estaria na faixa “moderado”,

cuja faixa compreende 0,15 a 1. Este valor está bem próximo do limite que o autor dispõe como “forte”, ou seja, assimetria elevada. Valores elevados de assimetria podem indicar a presença de vazões atípicas (*outliers*) que distorcem a média. Esta informação é importante, pois foi a partir deste entendimento que se optou por utilizar como função de perda para o modelo CB a métrica Erro Absoluto Médio (*Mean Absolute Error*, MAE). Como o modelo busca minimizar a função de perda, o emprego de tal métrica obteve os melhores resultados por esta penalizar os erros elevados, forçando um comportamento mais próximo do centro da distribuição dos dados.

Gráfico 3 – Distribuição dos dados



Fonte: Elaborado pelos autores (2025)

Foi realizada previsão em modo direto. A previsão direta em várias etapas é uma estratégia de previsão de séries temporais na qual um modelo separado é treinado para prever cada etapa do horizonte de previsão. Isso contrasta com a previsão recursiva em várias etapas, na qual um único modelo é usado para fazer previsões para todas as etapas futuras, utilizando recursivamente sua própria saída como entrada (Hyndman *et al.*, 2025). Como o horizonte proposto neste trabalho foi de 31 dias (mês de outubro) isso significa que 31 modelos foram treinados para prever cada um dos dias. A escolha por adotar esta estratégia deveu-se a um problema muito comum na previsão recursiva, que é a acumulação do erro de previsão, e em

testes para estes dados, neste cenário, a previsão direta apresentou melhores resultados. Cabe destacar que esta decisão é muito particular para cada problema.

Como forma de apoiar a compreensão do fenômeno de previsão de vazão para o rio Jequitinhonha, foram geradas previsões quantílicas usando o 5° percentil e o 95° percentil como alvos. Uma possível interpretação do intervalo de previsão é: calculados os valores inferior e superior do intervalo de previsão, o valor real observado futuro tem 90% (fora os 5% inferiores e os 5% superiores) de probabilidade de estar dentro destes limites. Apesar do foco em capturar um intervalo de previsão de 90%, o que pode ser considerado um intervalo bastante amplo, nem sempre o valor real é capturado. Existe uma incerteza quanto à previsão e esta incerteza advém da própria natureza dos dados hidrológicos e dos erros de ajuste do modelo.

Todos os experimentos foram realizados em um computador AMD Ryzen 7 5700 de 8 núcleos físicos e 16 processadores lógicos e 32GB de memória RAM. Além das bibliotecas Python *catboost*, *hydrobr* e *sktime*, mencionadas anteriormente, também foram utilizadas as bibliotecas *pandas* (The Pandas Development Team, 2020), *mlforecast* (Nixtla inc., 2022) e *plotly* (Plotly Technologies Inc., 2015). A versão da linguagem Python utilizada foi a 3.11.0 em um sistema Linux (*kernel* 6).

4 RESULTADOS

Os resultados mostraram-se bons, conforme Gráfico 4, considerando que havia uma quantidade de dados que pode ser entendida como reduzida, de 4 anos e 10 meses, conforme mencionado anteriormente. Em todas as imagens, a letra "y" representa a variável alvo, ou seja, os valores observados de vazão do rio Jequitinhonha.

Analisando apenas o modelo médio, a linha em verde, o mesmo obteve um MAPE de 15,66%. O comportamento médio das previsões esteve bem próximo do comportamento observado para o período. Um erro médio de vazão (RMSE) de 26,74 m³/s para uma série temporal que, conforme apresentado anteriormente, possui uma amplitude elevada, denota boa estabilidade do modelo *CatBoost*. Para efeitos de comparação, este valor está abaixo do que

Gomaa *et al.* (2023) encontrou em sua análise, de 60,03 m³/s, considerando a fase de teste do modelo mais complexo desenvolvido no estudo (MLP-PSO-EMD).

Gráfico 4 – Resultado com todos os modelos

Série	MAPE	RMSE	KGE
CatBoost	0.15658	26.74389	0.36064
CatBoost-q05	0.30651	49.65857	-0.06274
CatBoost-q95	1.29879	208.84858	-3.65755

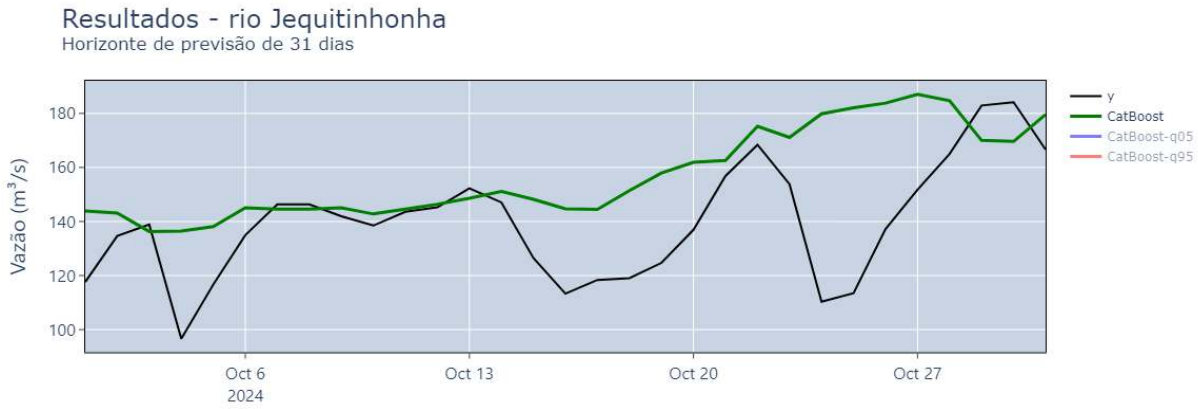
Fonte: Elaborado pelos autores (2025)

Se observarmos o trabalho de Nogueira Filho *et al.* (2022) podemos perceber que a escala do erro aqui ficou distante das faixas de valores apresentados pelos autores, que oscilou de 9,706 m³/s para o modelo LSTM-ic a 14,670 m³/s para o modelo SMAP-rg. Entretanto, um modelo era uma rede neural (LSTM), mais complexa em termos de treinamento e demanda de recursos computacionais, e um modelo físico (SMAP), o qual possui uma formulação distinta da apresentada aqui.

A KGE foi de 0,36, ou seja, distante do valor ótimo, que é 1. Contudo, a métrica agrega 3 informações ao mesmo tempo. Destrinchando esta métrica temos o seguinte: correlação de 0,41, variabilidade de 0,77 e viés de 1,12. As previsões do modelo médio representaram 77% da variabilidade real dos dados e um viés 12% acima. A correlação ficou pobre, de fato, mas com esta informação em mãos é possível direcionar esforços para buscar melhorar este marcador. Para uma boa visualização dos gráficos, foram geradas imagens com

apenas o modelo médio (Gráfico 5) e com os modelos para o 5° e 95° percentil, (respectivamente, em azul e vermelho, Gráfico 6).

Gráfico 5 –Resultado apenas com o modelo médio



Fonte: Elaborado pelos autores (2025)

Gráfico 6 – Resultado apenas com os percentis

Fonte: Elaborado pelos autores (2025)

Pelo Gráfico 6 é possível verificar que o intervalo de previsão proposto, de 90%, capturou satisfatoriamente o comportamento real observado para o período. Excetuando-se alguns momentos em que o valor observado (em preto) aproximou-se do previsto para o 5° percentil, quase ficando abaixo, os valores reais de vazão estiveram dentro do intervalo de

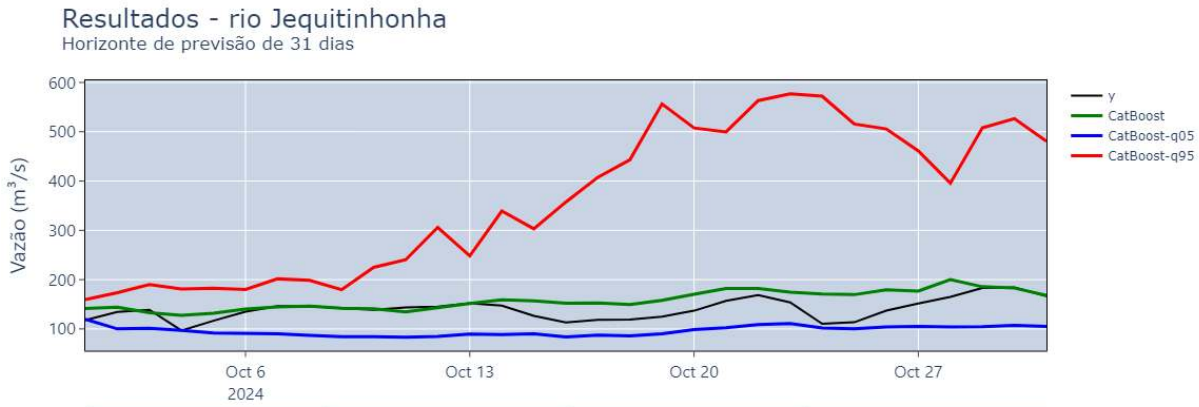
previsão. Como pode ser visto pela linha vermelha do Gráfico 6, o previsto para o 95° percentil descolou-se muito dos valores observados,

O modelo captou um cenário em que a vazão observada poderia chegar a 500 m³/s. Apesar de parecer não contribuir muito, o uso de intervalos de previsão, especialmente quando há um descolamento tão elevado como visto aqui, podem servir de apoio ao tomador de decisão. Entretanto, vale frisar, o trabalho não se prestou a analisar, especificamente, cenários extremos. O modelo previu um cenário deste, mas é certo que se faz necessário o uso de outras ferramentas (outros resultados oriundos de modelagem física, sensoriamento remoto, radar, etc.) que apoiem e possam embasar uma tomada de decisão.

Retomando sobre possíveis melhorias nos resultados. Cabe lembrar que este trabalho não se prestou a realizar otimização de hiperparâmetros do modelo *CatBoost*. Por isso, uma possível estratégia, sem necessariamente empregar esta otimização, seria buscar melhorar o marcador de correlação utilizando mais *lags* de vazão como dados de entrada para o treinamento do modelo. Para tanto, retome o gráfico PACF visto anteriormente no Gráfico 2. A próxima *lag* com autocorrelação parcial com valor significativo (fora da área enevoada azul clara) é a *lag* 5.

Perceba como houve uma melhora dos modelos médio e de 5° percentil (Gráfico 7) utilizando 5 *lags* para as previsões. O modelo que realizou previsão para o 95° percentil, no entanto, piorou em relação ao anterior. Uma explicação pode ser que empregando mais *lags* para treinamento o modelo captou mais incertezas acerca das previsões e encontrou outro cenário, descolando mais dos valores observados para o período analisado.

Gráfico 7 – Resultado melhorado com uso de 5 *lags*



Série	MAPE	RMSE	KGE
CatBoost	0.15381	25.7621	0.54341
CatBoost-q05	0.29935	47.91372	0.0628
CatBoost-q95	1.60224	263.50785	-5.3349

Fonte: Elaborado pelos autores (2025)

Destrinchando a métrica KGE, percebe-se uma considerável melhora na correlação. Veja que para uma KGE de 0,54, as componentes foram correlação de 0,57, variabilidade de 0,89 e viés de 1,12. O viés continuou 12% acima dos dados reais, mas a correlação saltou de 0,41 para 0,57 e a variabilidade de 0,77 para 0,89.

5 CONCLUSÃO

O modelo *CatBoost* demonstrou um comportamento médio bom na previsão de vazão com os dados utilizados neste trabalho, apresentando MAPE de aproximadamente 15%. Reduzir o erro médio é interessante, mas a métrica "perseguida" nesta pesquisa, a KGE, por sua complexidade, foi a que demonstrou melhores direcionamentos quanto à qualidade do modelo, bem como de melhorias. Com uma KGE de 0,54 o modelo, quando bem ajustado, pode extrair mais conhecimento a partir dos dados de entrada e melhorar a qualidade das previsões. O direcionamento apontado por esta métrica indica a necessidade de melhorar a correlação entre os dados previstos e os dados observados, mantendo-se em patamares próximos a variabilidade e o viés presentes. Isso pode ser alcançado com mais dados para treinamento, estendendo a base

de dados, bem como por ajustes nos hiperparâmetros do modelo *CatBoost*, o que pode ser realizado por busca em grade (*grid search* em inglês), busca aleatória (*random search* em inglês) ou otimização bayesiana.

Este trabalho teve como objetivo principal desenvolver e aplicar um modelo de aprendizado de máquina para previsão de vazão no rio Jequitinhonha, visando preencher uma lacuna no conhecimento sobre o emprego de modelos para previsão. Não apenas isso, mas servir também de base para estudos futuros que porventura queiram aplicar aprendizado de máquina para previsão de vazão de rio. É possível avançar mais sobre a compreensão das dinâmicas dos rios a partir desta ótica. Modelos de aprendizado de máquina podem prover de modo bastante eficiente dados em regiões onde há falta de informações e isso pode ser utilizado por modelos físicos, que realizam grandes simulações e de longo prazo. E dados simulados obtidos a partir de modelos físicos podem alimentar modelos de aprendizado de máquina para análise de cenários mais contidos, de curto prazo.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO (Brasil). **HIDROWEB**: sistemas de informações hidrológicas. 2005. Brasília, DF. Disponível em: https://www.snirh.gov.br/hidroweb/apresentacao. Acesso em: 15 mar. 2025.

AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO (Brasil). **Conjuntura dos recursos hídricos no Brasil 2024:** informe anual**.** Brasília: ANA, 2024. Disponível em: https://biblioteca.ana.gov.br/sophia_web/Acervo/Detalhe/106160?returnUrl=/sophia_web/Home/Index&guid=1734307203948. Acesso em: 15 mar. 2025.

BALLARIN, A. S. *et al*. CLIMBra - Climate Change Dataset for Brazil. **Scientific Data**, London, v. 10, n. 47, 2023. Disponível em: https://www.nature.com/articles/s41597-023-01956-z. Acesso em: 15mar. 2025.

BBC NEWS BRASIL. **A cronologia da tragédia no Rio Grande do Sul**. 2024. Disponível em: https://g1.globo.com/rs/rio-grande-do-sul/noticia/2024/05/12/a-cronologia-da-tragedia-no-rio-grande-do-sul.ghtml. Acesso em: 20 mar. 2025.

BBC NEWS BRASIL. **Chuvas na Bahia:** os fenômenos extremos que causam a tragédia no Estado. 2021. Disponível em: https://www.bbc.com/portuguese/brasil-59804297. Acesso em: 20 mar. 2025.

BRASIL. MAPBIOMAS. **Plataforma MapBiomas**. 2025. Disponível em: https://plataforma.brasil.mapbiomas.org. Acesso em: 25 mar. 2025.

BRASIL. Ministério de Minas e Energia. **Balanço Energético Nacional 2023**: ano base 2022. Rio de Janeiro: Empresa de Pesquisa Energética, 2023. Disponível em: https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-748/topico-687/BEN2023.pdf. Acesso em: 10 mar. 2025.

CARVALHO, W. M. de. **HydroBr**: a python package to work with brazilian hydrometeorological time series.2020. Disponível em: https://zenodo.org/records/3931027. Acesso em: 15 mar. 2025.

CATBOOST. **Common parameters**. 2025. Disponível em: https://catboost.ai/docs/en/references/training-parameters/common. Acesso em: 02 abr. 2025.

CRESPO, A. A. **Estatística fácil**. São Paulo: Saraiva, 2009.

GAMA ENGENHARIA E RECURSOS HÍDRICOS. **Relatório Diagnóstico dos Afluentes do Alto Jequitinhonha (JQ1) RT2**. Maceió: Gama, 2013. Disponível em: http://repositorioigam.meioambiente.mg.gov.br/handle/123456789/52. Acesso em: 20 mar. 2025.

GESUALDO, G. C. *et al*. Unveiling water security in Brazil: current challenges and future perspectives. **Hydrological Sciences Journal**, United Kingdom, v. 66, n. 5, p. 759–768, 2021. DOI: https://www.tandfonline.com/doi/full/10.1080/02626667.2021.1899182. Disponível em: https://www.tandfonline.com/doi/full/10.1080/02626667.2021.1899182. Acesso em: 20 mar. 2025.

GOMAA, E. *et al*. Assessment of hybrid machine learning algorithms using TRMM rainfall data for daily inflow forecasting in Três Marias Reservoir, eastern Brazil. **Heliyon**, London, v. 9, n. 8, ago. 2023. DOI: http://dx.doi.org/10.1016/j.heliyon.2023.e18819. Disponível em: https://www.sciencedirect.com/science/article/pii/S2405844023060279. Acesso em: 10 mar. 2025.

GUPTA, H. V. *et al*. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. **Journal Of Hydrology**, Netherlands, v. 377, p. 80-91. out. 2009. DOI: https://doi.org/10.1016/j.jhydrol.2009.08.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169409004843?via%3Dihub. Acesso em: 10 mar. 2025.

HYNDMAN, R. J. *et al*. **Forecasting:** Principles and Practice, the Pythonic Way. 2025. 14.3 Modern neural network architectures. Disponível em: https://otexts.com/fpppy/nbs/14-neural-networks.html#sec-modern-architectures. Acesso em: 24 jul. 2025.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. **International Journal Of Forecasting**. Amsterdam, v. 22, n.4 , out./dez. 2006. DOI: https://doi.org/10.1016/j.ijforecast.2006.03.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169207006000239?via%3Dihub. Acesso em: 10 mar. 2025.

HYNDMAN, R.; ATHANASOPOULOS, G. **Forecasting**: principles and practice. 2021. Disponível em: https://otexts.com/fpp3/. Acesso em: 10 mar. 2025.

INSTITUTO MINEIRO DE GESTÃO DAS ÁGUAS (IGAM). **Comitê da Bacia Hidrográfica dos Afluentes Mineiros do Alto Jequitinhonha - JQ1**. Disponível em: https://comites.igam.mg.gov.br/web/comites/jq1. Acesso em: 25 mar. 2025.

INSTITUTO MINEIRO DE GESTÃO DAS ÁGUAS (IGAM). **Comitê da Bacia Hidrográfica do Rio Araçuaí - JQ2**. Disponível em: https://comites.igam.mg.gov.br/web/comites/jq2. Acesso em: 25 mar. 2025.

INSTITUTO MINEIRO DE GESTÃO DAS ÁGUAS (IGAM). **Comitê da Bacia Hidrográfica dos Afluentes Mineiros do Médio e Baixo Rio Jequitinhonha - JQ3**. Disponível em: https://comites.igam.mg.gov.br/web/comites/jq3. Acesso em: 25 mar. 2025.

LAFORÉ, B.; FIGUEIREDO, C.; MALLMANN, D. **Temporais causam estragos em Minas Gerais e deixam desabrigados e desalojados.** 2023. CNN Brasil. Disponível em: https://www.cnnbrasil.com.br/nacional/temporais-causam-estragos-em-minas-gerais-e-deixam-desabrigados-e-desalojados/. Acesso em: 20 mar. 2025.

LOPES, V. **Entenda o que causou temporal na Região Sul do ES e o que pode ser feito para evitar novas tragédias**. 2024. G1 ES. Disponível em: https://g1.globo.com/es/espirito-santo/noticia/2024/03/27/entenda-o-que-causou-temporal-na-regiao-sul-do-es-e-o-que-pode-ser-feito-para-evitar-novas-tragedias.ghtml. Acesso em: 20 mar. 2025.

LÖNING, M. *et al*. Sktime: a unified interface for machine learning with time series. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS), 33., 2019, Vancouver. **Proceedings of the 33rd International Conference on Neural Information Processing Systems**. Red Hook, NY: Curran Associates Inc., 2019. Disponível em: http://learningsys.org/neurips19/assets/papers/sktime_ml_systems_neurips2019.pdf. Acesso em: 15 mar. 2025.

NIXTLA INC. **Machine Learning Forecast**: scalable machine learning for time series forecasting. Scalable machine learning for time series forecasting. 2022. Disponível em: https://nixtlaverse.nixtla.io/mlforecast/index.html. Acesso em: 02 mar. 2025.

NOGUEIRA FILHO, F. J. M. *et al*. Deep Learning for Streamflow Regionalization for Ungauged Basins: application of long-short-term-memory cells in semiarid regions. **Water**, Basel, v. 14, n. 9, abr. 2022. DOI: http://dx.doi.org/10.3390/w14091318. Disponível em: https://www.scopus.com/record/display.uri?eid=2-s2.0-85129356156&doi=10.3390%2fw14091318&origin=inward&txGid=a589086331ca528805e791b4bf0f2b09. Acesso em: 10 mar. 2025.

PLOTLY TECHNOLOGIES INC. **Collaborative data science**. 2015. Disponível em: https://plot.ly. Acesso em: 02 mar. 2025.

RIBEIRO, V. H. A.; REYNOSO-MEZA, G.; SIQUEIRA, H. V. Multi-objective ensembles of echo state networks and extreme learning machines for streamflow series forecasting. **Engineering Applications Of Artificial Intelligence**, England, v. 95, out. 2020. DOI: http://dx.doi.org/10.1016/j.engappai.2020.103910. Disponível em: https://www.scopus.com/record/display.uri?eid=2-s2.0-85089817082&doi=10.1016%2fj.engappai.2020.103910&origin=inward&txGid=0d8bd9dc1573fa1928c642bf144189fe. Acesso em: 10 mar. 2025.

RODRIGUES, J. A. M. *et al*. Hydrological modeling in a basin of the Brazilian Cerrado biome. **Ambiente e Agua,** Taubaté, v. 16, n. 1, jan. 2021. DOI: http://dx.doi.org/10.4136/ambi-agua.2639. Disponível em: https://www.scopus.com/record/display.uri?eid=2-s2.0-85100339826&doi=10.4136%2fambi-agua.2639&origin=inward&txGid=a7fa59f27c343f6537b326781821b154. Acesso em: 10 mar. 2025.

SILVA, F. B.; ALMEIDA, L. T. de; VIEIRA, E. de O. *et al*. Pluviometric and fluviometric trends in association with future projections in areas of conflict for water use. **Journal of Environmental Management**, England, v. 271, out. 2020. DOI: https://doi.org/10.1016/j.jenvman.2020.110991. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0301479720309191. Acesso em: 10 mar. 2025.

TEMPORAL devastador no Litoral Norte de SP completa um mês: confira um resumo da tragédia. **Reportagem G1 Globo**. 2023. Disponível em: https://g1.globo.com/sp/vale-do-paraiba-regiao/noticia/2023/03/19/temporal-devastador-no-litoral-norte-de-sp-completa-um-mes-confira-um-resumo-da-tragedia.ghtml. Acesso em: 20 mar. 2025.

TEMPORAL em Petrópolis: entenda o que provocou as chuvas intensas que causaram destruição na cidade. **Reportagem G1 Globo**. 2022. Disponível em: https://g1.globo.com/meio-ambiente/noticia/2022/02/15/temporal-em-petropolis-entenda-o-

que-provocou-as-chuvas-intensas-que-causaram-destruicao-na-cidade.ghtml. Acesso em: 20 mar. 2025.

THE PANDAS DEVELOPMENT TEAM. **Pandas-dev/pandas**: pandas. 2020. Versão 2.2.3. DOI: https://doi.org/10.5281/zenodo.3509134. Disponível em: https://zenodo.org/records/18328522. Acesso em: 02 mar. 2025.

VILANOVA, R. S.; ZANETTI, S. S.; CECÍLIO, R. A. Transferência de parâmetros de modelos baseada em redes neurais artificiais na simulação de vazão em bacias hidrográficas da Mata Atlântica brasileira**. Journal of Hydrologic Engineering**, Reston, v. 25, n. 7, maio 2020. DOI: http://dx.doi.org/10.1061/(asce)he.1943-5584.0001947. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084632633&doi=10.1061%2f%28ASCE%29HE.1943-5584.0001947&partnerID=40&md5=811109b65a634772ca7366f5db87c1fb. Acesso em: 10 mar. 2025.

AGRADECIMENTOS

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG e ao Instituto Mineiro de Gestão das Águas - IGAM pelo apoio financeiro, processo APQ01226-22.