Avaliação de algoritmos de *Machine Learning* para estimativa da condutividade elétrica de águas subterrâneas em vale aluvial do semiárido brasileiro

Eduardo Matheus Costa da Fonte[1] ; Oto Barbosa de Andrade[2] ; Thayná Alice Brito Almeida[3] & Abelardo Antônio de Assunção Montenegro[4]

RESUMO – A salinização é um grande desafio para a segurança hídrica no semiárido, e a gestão eficaz exige bancos de dados consistentes. Este estudo avalia algoritmos de aprendizado de máquina para estimar a condutividade elétrica da água subterrânea no Vale Aluvial do Mimoso, visando preencher falhas em séries históricas de monitoramento. A pesquisa utilizou dados pluviométricos e espaço-temporais da salinidade em poços vizinhos, testando os modelos K-Nearest Neighbors (KNN), Random Forest, Regressão Linear Múltipla e Cubist. A análise estatística revelou alta heterogeneidade espacial e temporal na resposta da salinidade à recarga. A precipitação do mês anterior destacou-se como um importante preditor para todos os modelos, demonstrando que a memória hidrológica recente contribui para a dinâmica salina. Dentre os modelos avaliados, o modelo KNN apresentou o melhor desempenho geral na predição e no preenchimento de dados ausentes, superando a Regressão Linear e o Random Forest, que falharam em capturar as relações não-lineares do transporte de sais no aquífero. Desta forma, conclui-se que algoritmos preditivos, baseados em variáveis acessíveis como a pluviometria, são ferramentas promissoras para reconstruir falhas de monitoramento, garantindo suporte essencial à gestão sustentável das águas subterrâneas no semiárido.

ABSTRACT– Salinization is a major challenge for water security in semi-arid regions, and effective management requires consistent and reliable databases. This study evaluates machine learning algorithms for estimating groundwater electrical conductivity in the Mimoso Alluvial Valley, aiming to fill gaps in historical monitoring series. The research used rainfall data and spatiotemporal salinity information from neighboring wells, testing the K-Nearest Neighbors (KNN), Random Forest, Multiple Linear Regression, and Cubist models. Statistical analyses revealed high spatial and temporal heterogeneity in the salinity response to rainfall inputs. Rainfall from the previous month emerged as an important predictor for the models, demonstrating that recent hydrological memory contributes to salinity dynamics. Among the evaluated approaches, the KNN model showed the best overall performance for prediction and gap-filling of missing data, outperforming Multiple Linear Regression and Random Forest, which were less effective in capturing nonlinear relationships associated with solute transport in the alluvial aquifer. Therefore, predictive algorithms based on readily available variables such as rainfall represent promising tools for reconstructing monitoring gaps and providing essential support for the sustainable management of groundwater resources in semi-arid regions.

Palavras-Chave – Preenchimento de dados; aprendizado de máquina; salinidade

[1]) Afiliação: UFRPE, eduardo.fonte@ufrpe.br
[2]) Afiliação: UFRPE, otobandrade@gmail.com
3) Afiliação: UFRPE, thaynaalicee@gmail.com
4) Afiliação: UFRPE, montenegro.ufrpe@gmail.com

INTRODUÇÃO

A escassez hídrica constitui um dos principais desafios das regiões semiáridas, onde a elevada variabilidade das precipitações e as altas taxas de evapotranspiração limitam a disponibilidade de água. Nesse contexto, os aquíferos aluviais desempenham papel fundamental na segurança hídrica do Nordeste brasileiro, contribuindo para o abastecimento de comunidades rurais e para a agricultura irrigada em pequena escala (Coelho et al., 2017; Almeida et al., 2024).

Apesar de sua importância, esses aquíferos apresentam elevada vulnerabilidade à salinização, processo frequentemente avaliado por meio da condutividade elétrica (CE) da água subterrânea. A dinâmica da CE é influenciada por fatores naturais e antrópicos, incluindo recarga pluviométrica, evapotranspiração, características geológicas e uso intensivo dos recursos hídricos, tornando seu monitoramento essencial para a gestão sustentável desses sistemas (Andrade et al., 2012; Almeida et al., 2024; Almeida et al., 2025a).

Falhas em séries históricas de monitoramento são frequentes em estudos hidroambientais e podem ocorrer em função de problemas operacionais, falhas em equipamentos, interrupções no fornecimento de energia elétrica ou limitações logísticas durante as campanhas de campo. E essa disponibilidade limitada de séries históricas consistentes representa um desafio recorrente no monitoramento hidrogeológico. Nesse contexto, técnicas de aprendizado de máquina têm sido amplamente utilizadas para modelagem e preenchimento de falhas em dados ambientais, permitindo identificar padrões complexos a partir de variáveis espaciais e meteorológicas (Haggerty et al., 2023; Sahour et al., 2023; Niazkar et al., 2025).

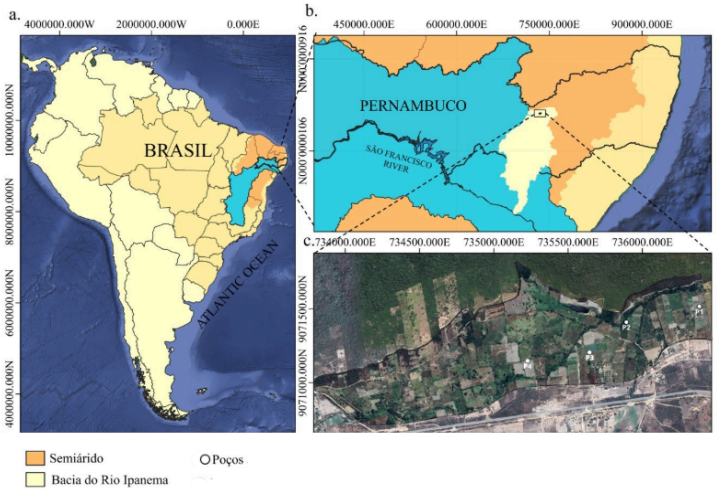
Assim, este trabalho teve como objetivo avaliar o potencial de diferentes algoritmos de aprendizado de máquina na estimativa da condutividade elétrica da água subterrânea do Vale Aluvial do Mimoso, com ênfase na aplicação dessas técnicas para o preenchimento de falhas em séries históricas de monitoramento hidroambiental.

METODOLOGIA

Área de estudo e base de dados

O estudo foi desenvolvido no Vale Aluvial do Mimoso, no município de Pesqueira-PE, utilizando dados de monitoramento da condutividade elétrica (CE) da água subterrânea obtidos em quatro poços comunitários (P1, P2, P3 e P4). Para cada registro foram consideradas informações espaciais, temporais e pluviométricas.

Figura 1: Área de estudo



As variáveis independentes foram selecionadas considerando fatores espaciais, temporais e pluviométricos potencialmente relacionados às variações da condutividade elétrica (CE) da água subterrânea. Foram consideradas a localização dos poços (latitude e longitude), a identificação do ponto de monitoramento, o mês e ano de coleta, a precipitação acumulada do mês atual (P.atual), a precipitação acumulada do mês anterior (P.anterior) e a precipitação acumulada nos 60 dias anteriores à coleta (P60).

Relação entre precipitação e condutividade elétrica

Para avaliar a influência da precipitação sobre a condutividade elétrica da água subterrânea, foram ajustados modelos de regressão linear simples para cada poço, utilizando a CE como variável dependente e P60 como variável independente, conforme a equação 1:

$$CE = \beta 0 + \beta 1 P60 + \varepsilon \quad (1)$$

onde CE é condutividade elétrica; β_0 = intercepto; β_1 = coeficiente angular; P60 = precipitação acumulada dos 60 dias anteriores; ε = erro aleatório.

Para cada poço foram calculados coeficiente de correlação de Pearson (r), coeficiente de determinação (R^2) e significância estatística do coeficiente angular (p-valor).

Correlação entre poços

Buscando avaliar a similaridade temporal do comportamento salino entre os pontos monitorados, os dados foram agrupados em dois períodos hidrológicos, conforme observações de Almeida *et al*. (2025a) em período chuvoso (abril a julho) e período seco (demais meses). Posteriormente, foram realizadas análises de correlação entre os valores de condutividade elétrica dos poços P1, P2 e P3 em relação ao P4, adotado como referência.

Para cada combinação de poço e período hidrológico foram calculados coeficiente de correlação de Pearson (r) e coeficiente de determinação (R^2) obtido por regressão linear simples. O

objetivo dessa etapa foi identificar possíveis padrões espaciais e temporais de comportamento da condutividade elétrica entre os poços monitorados.

Desenvolvimento dos modelos de aprendizado de máquina

Com o objetivo de avaliar a capacidade preditiva de diferentes algoritmos para estimativa da condutividade elétrica, foram empregados métodos de aprendizado de máquina supervisionado. Os modelos foram implementados no software R utilizando o pacote Caret, que permite padronizar os procedimentos de treinamento, validação e comparação entre algoritmos.

A base de dados foi dividida aleatoriamente em dois subconjuntos, sendo 70% dos registros destinados ao treinamento dos modelos e 30% dos registros destinados à validação independente. Foram avaliados quatro algoritmos distintos:

Foram avaliados os algoritmos KNN, Random Forest, Regressão Linear Múltipla e Cubist, implementados no pacote Caret do software R. Os modelos foram calibrados por validação cruzada, resultando nos parâmetros k = 5 (KNN), mtry = 2 (RF), neighbors = 5 e committees = 1 (Cubist), enquanto a regressão linear utilizou todas as variáveis explicativas disponíveis.

Validação dos modelos

Para reduzir riscos de sobreajuste e aumentar a robustez estatística dos resultados, todos os algoritmos foram submetidos à validação cruzada do tipo k-fold. Foi adotada validação cruzada com 10 folds. Os modelos foram calibrados por validação cruzada k-fold (k = 10), utilizando RMSE como critério para seleção dos melhores hiperparâmetros.

Avaliação do desempenho dos modelos

O desempenho dos algoritmos foi avaliado utilizando o conjunto independente de validação (30% dos dados). Foram calculados os seguintes indicadores estatísticos: Coeficiente de determinação, a Raiz do Erro Quadrático Médio (RMSE), e Erro Absoluto Médio (MAE).

Análise da importância das variáveis

A importância relativa das variáveis foi avaliada por meio da função varImp() do pacote Caret para os modelos KNN, Random Forest, Regressão Linear e Cubist. Entretanto, para o algoritmo Cubist não foram observados valores interpretáveis de importância das variáveis, sendo a discussão concentrada nos resultados obtidos pelos demais modelos.

Aplicação dos modelos para estimativa de dados ausentes

Após a etapa de avaliação e comparação dos algoritmos, os modelos treinados foram utilizados para estimar registros ausentes de condutividade elétrica identificados na série temporal monitorada.

RESULTADOS E DISCUSSÃO

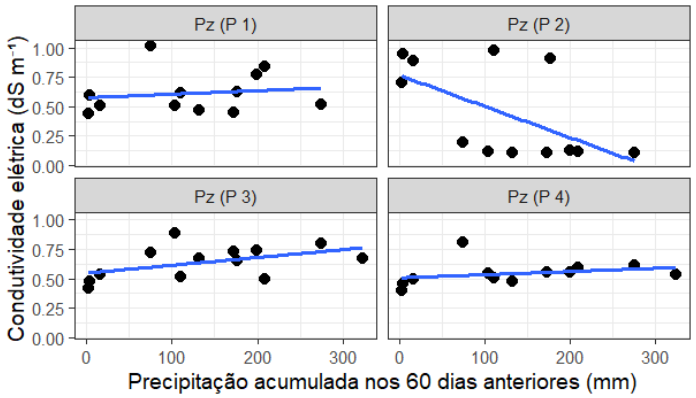
A avaliação das relações entre a precipitação acumulada dos 60 dias anteriores (P60) e a CE evidenciou que apenas o poço P2 apresentou correlação estatisticamente significativa, com relação inversa e moderada indicando que aproximadamente 34,5% da variabilidade da condutividade elétrica pode ser explicada pela precipitação antecedente (Tabela 1). O sinal negativo da correlação sugere que maiores lâminas de chuva estão associadas à redução da salinidade da água subterrânea.

Tabela 1 – Correlação dos poços com a P60

Poço	r	R^2	p-valor
P1	0,147	0,022	0,6487
P2	-0,587	0,345	0,0448
P3	0,476	0,227	0,0997
P4	0,276	0,076	0,3851

Nos poços P1, P3 e P4, as correlações mostraram-se fracas a moderadas e sem significância estatística (Figura 2).

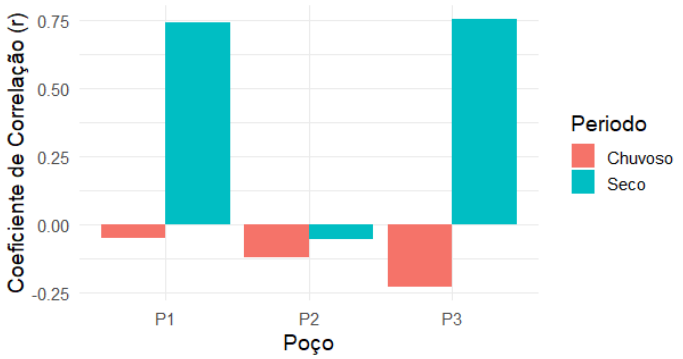
Figura 2: Correlação dos poços com a precipitação acumulada (p60)



A baixa correlação indica que outras variáveis interferem na distribuição de sais em tais poços. As diferenças locais na condutividade hidráulica do solo, as variações na textura (presença de lentes de argila ou silte) e as taxas pontuais de extração por bombeamento geram respostas de salinidade distintas mesmo em pontos geograficamente próximos (Almeida et al., 2024; Coelho et al., 2017).

A correlação espacial entre os poços (utilizando o poço P4 como variável dependente) apresentou comportamentos distintos entre as estações hidrológicas. Durante o período seco, foram observadas fortes correlações positivas entre P4 e os demais poços (Figura 3).

Figura 3: Correlação dos pontos em relação a P4 com a estação



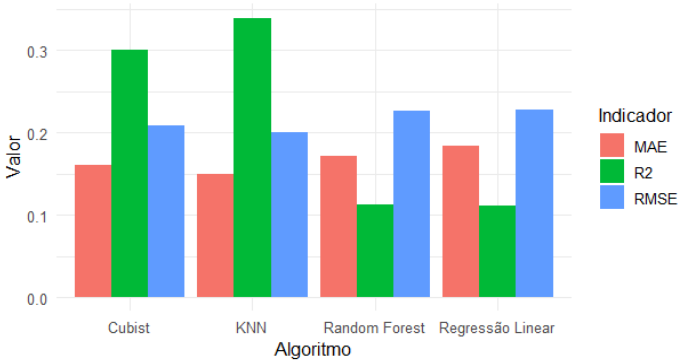
Esses resultados indicam elevado grau de similaridade temporal entre os pontos, sugerindo que mais de 55% da variabilidade observada nesses poços pode ser explicada pelo comportamento registrado em P4. No período chuvoso, entretanto, essas correlações tornaram-se praticamente nulas. O poço P2 exibiu baixa correlação em ambos os períodos.

A influência da sazonalidade sobre a correlação entre pontos de monitoramento também foi observada por Moura *et al*. (2017), que verificaram desempenho superior dos modelos em escalas temporais mais agregadas e com maior estabilidade hidrológica, devido ao aumento da correlação natural entre os registros.

As baixas correlações no período chuvoso indica que a infiltração de água não ocorre de forma uniforme, e sinalizam que outros fatores tais como as irrigações nos lotes devem ser levados em consideração. Segundo Andrade *et al*. (2012) e Almeida *et al*. (2025), a recarga em aluviões semiáridos cria "bolsões" localizados de água doce devido aos fluxos preferenciais nas frações mais arenosas do perfil, rompendo a continuidade espacial da salinidade. O comportamento isolado do poço P2 corrobora a tese de que fatores antrópicos, como o uso do solo e o manejo de irrigação no entorno imediato do poço, podem se sobrepor à dinâmica climática regional (Almeida *et al.*, 2025b). Conforme Almeida et al. (2024), a precipitação antecedente exerce dois efeitos antagônicos: promove lixiviação dos sais das camadas de solo superiores, elevando a salinidade da água na zona saturada; e possibilita a diluição dos sais da zona saturada devido à recarga, reduzindo assim a salinidade nesta camada.

No tocante ao desenvolvimento dos modelos preditivos para estimativa da CE, o KNN apresentou o melhor desempenho geral, seguido pelo método Cubist. Em divergência com certas aplicações clássicas, os algoritmos RF e MLR apresentaram baixo poder preditivo, conforme Figura 4.

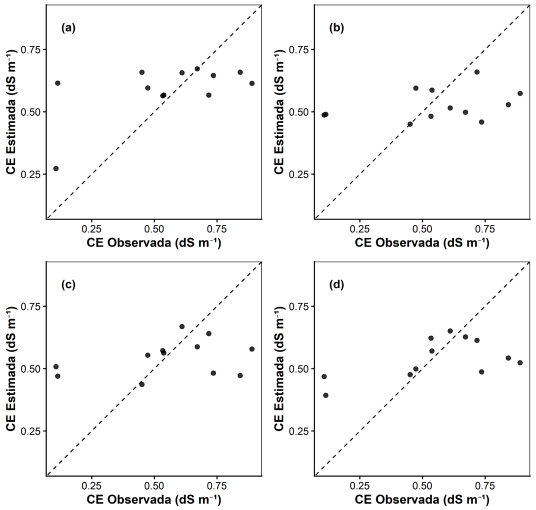
Figura 4: desempenho dos algoritmos



Ledra *et al*. (2017) encontraram R^2 médio de aproximadamente 0,26 utilizando Redes Neurais Artificiais para preenchimento de falhas em séries pluviométricas curtas. Dessa forma, o desempenho obtido pelo KNN (R^2 = 0,339) pode ser considerado satisfatório para uma base hidroambiental limitada. Por outro lado, Arriagada *et al*. (2021) obtiveram R^2 superiores a 0,60 utilizando o algoritmo MissForest em séries de vazão com mais de 15 anos de observação. Os autores demonstraram que o aumento do histórico temporal e do número de variáveis preditoras melhora significativamente a capacidade de generalização dos modelos.

Todos os modelos apresentaram dispersão moderada em relação à reta 1:1, indicando diferenças entre valores observados e estimados pelo modelo (Figura 5. Observou-se tendência à concentração das previsões em faixas intermediárias de CE, com menor capacidade de reprodução dos valores extremos.

Figura 5: comparação entre os valores estimados e observados de CE



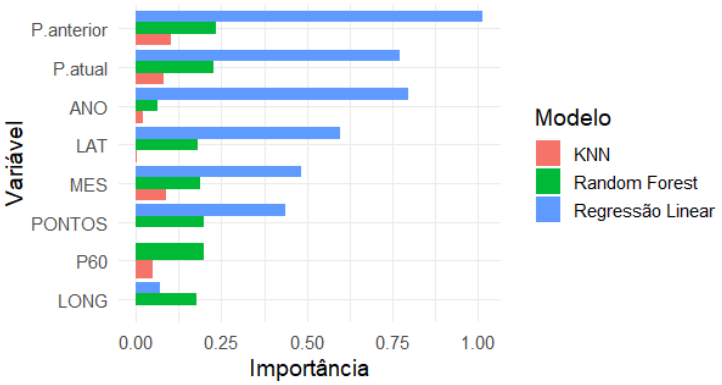
KNN (a), Regressão Linear (b), Random Forest (c) e Cubist (d)

Estudos comparativos em modelagem hidroambiental frequentemente apontam o RF como um modelo superior (Sahour *et al*., 2023; Singha *et al*., 2021). Contudo, Niazkar *et al*. (2025)

ressaltam que o desempenho dos algoritmos de aprendizado de máquina varia significativamente a depender das dimensões e características intrínsecas dos dados. O baixo desempenho da regressão linear reflete a alta não-linearidade das diluições geoquímicas e climáticas que governam a salinidade no aquífero, processos que algoritmos paramétricos globais falham em capturar (Haggerty *et al*., 2023).

A análise de importância revelou um consenso entre os modelos testados: a precipitação do mês anterior foi a variável de maior importância nos modelos avaliados (figura 6).

Figura 6: Importância das variáveis



Esses resultados indicam que a memória hidrológica recente exerce papel fundamental na determinação da condutividade elétrica da água subterrânea. A água precipitada na superfície necessita de tempo para percolar a zona não saturada e lixiviar os sais acumulados no perfil do solo até atingir o lençol freático e alterar a condutividade elétrica efetiva da água subterrânea (Almeida *et al*., 2025b).

Os algoritmos treinados foram utilizados para estimar três registros ausentes de condutividade elétrica. Os modelos geraram estimativas consistentes para os registros ausentes, de acordo com o desempenho observado na etapa de validação. Os resultados variaram, principalmente para o piezômetro P4, que oscilou entre 0,606 dS m^{-1} (KNN) e 0,731 dS m^{-1} (Regressão Linear) conforme tabela 2.

Tabela 2: Preenchimento das falhas de condutividade elétrica

Poço	KNN	RF	LM
Pz(P1)	0,580	0,516	0,489
Pz(P2)	0,272	0,405	0,414
Pz(P4)	0,606	0,657	0,731

Essa variabilidade evidencia a incerteza inerente ao processo de preenchimento de dados ambientais e reforça a importância da comparação entre múltiplos algoritmos antes da definição do

valor final a ser incorporado à série histórica. Resultados semelhantes foram relatados por Al-Jahwari *et al.* (2026), que verificaram redução significativa dos erros quando diferentes algoritmos foram combinados em modelos ensemble. Os autores demonstraram que a integração de múltiplos métodos permite reduzir vieses individuais e aumentar a robustez das estimativas.

De forma semelhante, Solanki e Mishra (2026) observaram que abordagens baseadas em florestas de regressão apresentaram elevada capacidade de reconstrução de séries hidrológicas, embora alertem para o aumento da incerteza em aplicações envolvendo eventos extremos.

CONCLUSÕES

Os resultados demonstraram que a relação entre a precipitação antecedente e a condutividade elétrica da água subterrânea apresenta comportamento espacialmente heterogêneo no vale, com maior similaridade entre os poços durante o período seco e redução das correlações no período chuvoso.

Entre os algoritmos avaliados, o KNN apresentou o melhor desempenho na estimativa da condutividade elétrica, seguido pelo Cubist. A precipitação do mês anterior foi a variável mais importante para os modelos, evidenciando a influência da memória hidrológica recente na dinâmica da salinidade. Os resultados indicam que, em bases hidroambientais de pequena dimensão, modelos mais simples podem apresentar desempenho igual ou superior a algoritmos mais complexos.

Os modelos mostraram potencial para aplicação no preenchimento de falhas em séries históricas de qualidade da água subterrânea. Entretanto, o desempenho moderado obtido evidencia que a limitada extensão da série histórica restringe a capacidade preditiva dos algoritmos, reforçando a importância da continuidade do monitoramento e da ampliação da base de dados para aumentar a precisão das estimativas. Séries históricas mais longas tendem a aumentar a robustez dos modelos e a precisão das estimativas geradas por técnicas de aprendizado de máquina.

AGRADECIMENTOS

Agradecimentos à Universidade Federal Rural de Pernambuco (UFRPE), ao Laboratório de Água e Solo (LAS) e à Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE), CNPq, CAPES, Finep pelo apoio institucional e científico.

REFERÊNCIAS

ANDRADE, T.S.; MONTENEGRO, S.M.G.L.; MONTENEGRO, A.A.A.; RODRIGUES, D.F.B. (2012). "Variabilidade espaço-temporal da condutividade elétrica da água subterrânea na região semiárida de Pernambuco". Revista Brasileira de Engenharia Agrícola e Ambiental 16(5), pp. 496–504.

AL-JAHWARI, B.; AL-RAWAS, G.; NIKOO, M.R.; ETRI, T.; GRUNDMANN, J. (2026). "Robust rainfall gap-filling in coastal arid regions using ensemble fusion models". Hydrology 13(1), pp. 1.

ALMEIDA, T.A.B.; BOAVENTURA, L.C.S.; SILVA, M.V.; FARIAS, C.W.L.A.; CHAGAS, A.M.S.; COSTA, R.S.; MOURA, C.V.S.; MONTENEGRO, A.A.A. (2025a). "Assessing Shallow Groundwater Depth and Electrical Conductivity in the Brazilian Semiarid: A Geostatistical Analysis". Geosciences 15(4), art. 136.

ALMEIDA, T.A.B.; MONTENEGRO, A.A.A.; LIMA, J.L.M.P.; FARIAS, C.W.L.A.; CARVALHO, A.A.; PAIVA, A.L.R. (2025b). "Spatiotemporal Variability of Groundwater Quality for Irrigation: A Case Study in Mimoso Alluvial Valley, Semiarid Region of Brazil". Water 17(3), art. 410.

ALMEIDA, T.A.B.; MONTENEGRO, A.A.A.; MACKAY, R.; MONTENEGRO, S.M.G.L.; COELHO, V.H.R.; CARVALHO, A.A.; SILVA, T.G.F. (2024). "Hydrogeological trends in an alluvial valley in the Brazilian semiarid: Impacts of observed climate variables change and exploitation on groundwater availability and salinity". Journal of Hydrology: Regional Studies 53, art. 101784.

ARRIAGADA, P.; KARELOVIC, B.; LINK, O. (2021). "Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm". Journal of Hydrology 598, art. 126454.

COELHO, V.H.R.; MONTENEGRO, S.; ALMEIDA, C.N.; SILVA, B.B.; OLIVEIRA, L.M.; GUSMÃO, A.C.V.; FREITAS, E.S.; MONTENEGRO, A.A.A. (2017). "Alluvial groundwater recharge estimation in semi-arid environment using remotely sensed data". Journal of Hydrology 548, pp. 1–15.

HAGGERTY, R.; SUN, J.; YU, H.; LI, Y. (2023). "Application of machine learning in groundwater quality modeling: A comprehensive review". Water Research 233, art. 119745.

LEDRA, L.; BARROS, V.G.; HENNING, E. (2017). "Preenchimento de falhas de dados diários de precipitação utilizando redes neurais artificiais e método das correlações" in Anais do XXII Simpósio Brasileiro de Recursos Hídricos.

MOURA, C.N. de; SÁ, É.A.S.; PADILHA, V.L.; RAFAELI NETO, S.L. (2017). "Desempenho do machine learning para o preenchimento de falhas em séries de vazões diárias e mensais" in Anais do XXII Simpósio Brasileiro de Recursos Hídricos.

NIAZKAR, M.; PIRAEI, R.; GOODARZI, M.R.; ABEDI, M.J. (2025). "Comparative Assessment of Machine Learning Models for Groundwater Quality Prediction Using Various Parameters". Environmental Processes 12(10).

SAHOUR, S.; KHANBEYKI, M.; GHOLAMI, V.; SAHOUR, H.; KAHVAZADE, I.; KARIMI, H. (2023). "Evaluation of machine learning algorithms for groundwater quality modeling". Environmental Science and Pollution Research 30, pp. 46004–46021.

SINGHA, S.; PASUPULETI, S.; SINGHA, S.S.; SINGH, R.; KUMAR, S. (2021). "Prediction of groundwater quality using efficient machine learning technique". Chemosphere 276, art. 130265.

SOLANKI, H.; MISHRA, V. (2026). "Filling streamflow data gaps in Indian catchments using machine learning and K-means clustering". Water Resources Research 62, pp. e2025WR042210.