

Matemática

Previsão de vazões totais para horizonte de médio prazo via modelagem guiada por dados

Forecast of total flows for medium-term horizon via data-driven modeling

Rute Santos Porto Lima¹ , Alcigeimes Batista Celeste¹ 

¹Universidade Federal de Sergipe, SE, Brasil

RESUMO

Este trabalho apresenta a aplicação de redes neurais artificiais (RNA), algoritmo dos k-vizinhos mais próximos (KNN) e máquinas de regressão por vetores de suporte (SVR) para modelar a previsão de vazões afluentes ao reservatório de Sobradinho, na Bahia. Utilizando o software Weka, seis formulações foram testadas para previsões de 3 meses, com dados divididos em esquemas de validação cruzada e calibração-validação. A eficiência foi avaliada pelo coeficiente de Nash-Sutcliffe, destacando a superioridade da SVR. A inclusão de atributos como vazão média e precipitação melhorou as eficiências. O modelo usando KNN com 13 vizinhos foi incorporado à uma estratégia de otimização estocástica implícita aprimorada para a operação do reservatório, a qual foi comparada a outros métodos de operação, mostrando superioridade no índice de vulnerabilidade.

Palavras-chave: Previsão de afluências; Aprendizado de máquina; Weka; Operação de reservatório

ABSTRACT

This study presents the application of artificial neural networks (ANN), k-nearest neighbors algorithm (KNN), and support vector regression (SVR) for modeling the prediction of inflows to the Sobradinho reservoir in Bahia, Brazil. Using the Weka software, six formulations were tested for 3-month forecasts, with data divided into cross-validation and calibration-validation schemes. Efficiency was evaluated using the Nash-Sutcliffe coefficient, highlighting the superiority of SVR. The inclusion of attributes such as average flow and precipitation improved efficiencies. The model using KNN with 13 neighbors was incorporated into an enhanced implicit stochastic optimization strategy for the operation of the reservoir, which was compared to other operational methods, showing superiority in vulnerability index.

Keywords: Inflow forecast; Machine learning; Weka; Reservoir operation

1 INTRODUÇÃO

Reservatórios de águas superficiais são meios eficazes para atender aos requisitos crescentes e muitas vezes conflitantes relacionados à utilização e controle dos recursos hídricos. Esses sistemas são projetados com o objetivo de gerenciar os recursos disponíveis de maneira sustentável, com o objetivo de garantir sua viabilidade a longo prazo. A implementação de sistemas de reservatórios com múltiplos usos não só ajuda a lidar com as discrepâncias decorrentes de variações na distribuição temporal e espacial dos recursos hídricos, mas também a atender às diversas demandas impostas (Nagy et al., 2002).

Técnicas de otimização matemática desempenham um papel importante na operação de reservatórios, fornecendo soluções eficientes para problemas complexos de gerenciamento de água. Modelos de otimização ajudam a determinar políticas de liberação ideais, maximizando os benefícios e minimizando os custos associados (Wu et al., 2023; Giuliani et al., 2021). As incertezas hidrológicas envolvidas são levadas em consideração por procedimentos de otimização estocástica (Celeste & Billib, 2009), sendo frequente o uso da programação dinâmica estocástica (PDE), embora, conforme Giuliani et al. (2021), esta técnica possua três assim chamadas “maldições”: da dimensionalidade, da modelagem, e dos múltiplos objetivos. Uma alternativa à PDE é a aplicação da otimização estocástica implícita (OEI), que tem sido amplamente adotada devido à sua capacidade de lidar com as limitações da PDE (Giuliani et al., 2021; Celeste & Billib, 2009; Khadr & Schlenkhoff, 2021; Ávila et al., 2020).

O estudo de Santana & Celeste (2021) comparou vários métodos de operação aplicados ao reservatório de Sobradinho (Bahia), incluindo OEI incorporada com previsão da afluência média prevista para um horizonte de alguns meses futuro. O modelo de OEI com previsões para mais de 3 meses mostrou menos vulnerabilidade do que outros modelos quando testado em cenários de vazões sintéticas. Esse estudo evidenciou que a previsão de vazões afluentes é importante em modelos de OEI.

Os modelos de previsão de vazão podem ser categorizados em conceituais, em que as funções utilizadas incorporam uma descrição detalhada dos processos físicos (Refsgaard, 1990), e empíricos, ou guiados por dados, que relacionam os dados de entrada e saída por meio de funções que podem não ter relevância direta com os processos físicos envolvidos (Solomatine, 2006). Modelos conceituais normalmente

requerem conhecimento de parâmetros hidrológicos, alguns dos quais não são diretamente mensuráveis. Consequentemente, devido a essa restrição específica, os modelos guiados por dados estão ganhando popularidade (Solomatine et al., 2007).

Este trabalho tem como objetivo avaliar a aplicação de modelos guiados por dados, baseados em redes neurais artificiais, no algoritmo dos k-vizinhos mais próximos, e em máquinas de vetores de suporte, identificando seus desempenhos na previsão de vazões totais para um horizonte de médio prazo. O propósito é auxiliar em estratégias de otimização estocástica implícita para a operação de reservatórios.

No modelo de operação ótima proposto por Santana & Celeste (2021), mencionado anteriormente, a previsão de vazão foi realizada de forma “perfeita”, com base em valores observados de cenários sintéticos gerados para afluências futuras, ou seja, não foi efetivamente utilizado um modelo de previsão. Portanto, uma tarefa adicional do presente trabalho foi incorporar o melhor modelo de previsão desenvolvido à modelagem OEI de Santana & Celeste (2021) para verificar seu desempenho na operação de Sobradinho.

2 MATERIAL E MÉTODOS

2.1 Formulação

A proposta metodológica desta pesquisa foi aplicar algoritmos baseados em modelagem guiada por dados (DDM: *data-driven modeling*) a fim de prever vazões mensais totais futuras para um horizonte de médio prazo. Seja Q_t^H a vazão total para o horizonte de H meses no futuro a partir do mês t , definida como

$$Q_t^H = \sum_{\tau=t}^{t+H-1} Q_{\tau} \quad (1)$$

em que Q_{τ} é a vazão no mês τ . Dessa forma, $Q_t^H = Q_t + Q_{t+1} + \dots + Q_{t+H-1}$, ou seja, a soma dos deflúvios nos próximo H meses, incluindo o mês atual. O objetivo foi, portanto, prever o valor de Q_t^H por meio do modelo

$$Q_t^H = \mathcal{F}(x_1, x_2, \dots, x_n) \quad (2)$$

onde \mathcal{F} é a função que correlaciona a saída (vazão a ser prevista) com as entradas (n atributos x_1 até x_n). Os atributos podem incluir, por exemplo, chuvas e/ou vazões de períodos passados. Cada modelo guiado por dados define uma função \mathcal{F} correspondente.

Tabela 1 – Formulações investigadas

Formulação	Saída	Entrada (atributos)	N.º de Atributos
1	Q_t^H	$\sum_{i=1}^L Q_{t-i}$ para $L = 1, 2, 3, 4, 12, 13, 14, 15, 16$	9
2		$Q_{t-1}, Q_{t-2}, Q_{t-3}$	3
3		$P_t, P_{t-1}, P_{t-2}, P_{t-3}, Q_{t-1}, Q_{t-2}, Q_{t-3}$	7
4		$P_{t-1}, P_{t-2}, P_{t-3}, Q_{t-1}, Q_{t-2}, Q_{t-3}$	6
5		$Q_{t-1}, Q_{t-2}, Q_{t-3}, \bar{Q}_t$	4
6		$Q_{t-1}, Q_{t-2}, Q_{t-3}, \bar{Q}_t, \bar{P}_t$	5

Fonte: autores (2025)

Para este trabalho, o horizonte de previsão utilizado foi de 3 meses, de forma que $Q_t^3 = Q_t + Q_{t+1} + Q_{t+2}$. A formulação dos atributos foi composta de maneira variada, tendo sido utilizados dados de vazão e/ou precipitações passadas (acumuladas ou não) e, em alguns casos, acrescidos de vazão e/ou precipitação média, estas denotadas por \bar{Q}_t e \bar{P}_t , respectivamente. As formulações foram selecionadas heurísticamente e estão detalhadas na Tabela 1. Para a formulação 1, a composição de cada atributo de entrada consistiu na vazão acumulada para *lags* (defasagens) de L meses anteriores ao mês t , usando os 9 valores de L mostrados na tabela, cada um para o respectivo atributo. Assim, os atributos foram Q_{t-1} (para $L = 1$), $Q_{t-1} + Q_{t-2}$ (para $L = 2$), $Q_{t-1} + Q_{t-2} + Q_{t-3}$ (para $L = 3$), e assim por diante até $L = 16$. Essa formulação conta com 9 atributos de entrada, sendo que outras formulações com número de atributos diferentes também foram testadas, mas seria inviável expor todas.

2.2 Algoritmos e Plataforma Utilizados

A função \mathcal{F} da equação (2) pode ser modelada por qualquer estratégia de aprendizado de máquina. Neste trabalho, foram usadas as técnicas de redes neurais artificiais, algoritmo dos k -vizinhos mais próximos e máquinas de regressão por vetores de suporte.

2.2.1 Redes Neurais Artificiais (RNA)

As redes neurais artificiais (RNA) são modelos computacionais baseados na estrutura do cérebro humano, formados por neurônios artificiais interligados que processam e transmitem informações (Morettin & Singer, 2023; Witten et al., 2016; Shu et al., 2021; Chiamsathit et al., 2016). Uma arquitetura amplamente utilizada em tarefas de previsão é o perceptron multicamadas, uma rede *feedforward*, na qual o sinal flui em uma única direção, passando por uma camada de entrada, uma ou mais camadas ocultas, até alcançar a camada de saída (ver Figura 1, apresentada posteriormente na Seção 2.4.2). Esta pesquisa utilizou redes do tipo perceptron multicamadas.

As RNA são capazes de aprender a partir dos dados, ajustando os pesos das conexões entre os neurônios durante o treinamento. Matematicamente, a saída de um neurônio em uma camada oculta é descrita por

$$a_i = \phi \left(\sum_j W_{ij} \cdot x_j + b_j \right) \quad (3)$$

onde W_{ij} são os pesos das conexões entre os neurônios, x_j são as entradas, b_j é o termo de *bias*, e ϕ é a função de ativação.

A função de ativação transforma a soma ponderada das entradas para gerar uma saída não linear. Uma função de ativação comum é a função sigmoide, descrita por:

$$\phi(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

Neste estudo, a função sigmoide foi utilizada nas camadas ocultas para introduzir não linearidade, enquanto a função linear foi empregada na camada de saída.

2.2.2 Algoritmo dos k-vizinhos mais próximos (KNN)

O KNN (*k-nearest neighbors algorithm*) é um método de aprendizado baseado em instâncias que funciona buscando os k pontos mais próximos de uma amostra no espaço de características e realizando uma previsão com base na média desses

pontos. Para medir a proximidade, utiliza-se, por exemplo, a distância euclidiana:

$$d(\mathbf{x}_q, \mathbf{x}_j) = \sqrt{\sum_{i=1}^n (x_{ij} - x_{iq})^2} \quad (5)$$

onde \mathbf{x}_q é o vetor de consulta (*query vector*) e \mathbf{x}_j é o j -ésimo vetor de atributos.

O valor previsto y_q para a amostra \mathbf{x}_q é calculado como uma média ponderada dos valores de saída y_j dos k vizinhos mais próximos:

$$y_q = \frac{\sum_{j=1}^k w_j^V y_j}{\sum_{j=1}^k w_j^V} \quad (6)$$

onde w_j^V é o peso do j -ésimo vizinho, que pode ser inversamente proporcional à sua distância. Vários tipos de funções de peso (*kernel*) podem ser aplicados, como funções lineares ou gaussianas. O valor de k é um parâmetro que precisa ser definido e influencia na suavidade da decisão final (Morettin & Singer, 2023; Witten et al., 2016; Maddu et al., 2022; Akbari et al., 2010).

2.2.3 Máquinas de regressão por vetores de suporte (SVR)

As máquinas de regressão por vetores de suporte (SVR: *support vector regression*) são uma extensão das máquinas de vetores de suporte (SVM: *support vector machines*) para problemas de regressão. Enquanto as SVM são amplamente utilizadas para classificação, as SVR são específicas para prever valores numéricos em vez de classes. Elas são baseadas no conceito de encontrar um hiperplano ótimo que, em vez de classificar dados, aproxima uma função dentro de uma margem de tolerância ϵ , minimizando erros de predição (Morettin & Singer, 2023; Witten et al., 2016; Wu et al., 2012).

O objetivo da SVR é encontrar uma função linear $f(\mathbf{x})$ que tenha o menor desvio possível em relação aos dados de treinamento, mas que, ao mesmo tempo, mantenha uma certa margem de tolerância para erros. A equação geral de uma SVR é:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad (7)$$

onde \mathbf{w} é o vetor de pesos, \mathbf{x} é vetor de entrada e b é o termo de *bias*.

Para que a SVR seja eficaz em regressão, utiliza-se a função de perda ϵ -insensível, que introduz uma faixa de tolerância ao redor da linha de regressão onde os desvios não são penalizados:

$$L_{\epsilon}(y, f(\mathbf{x})) = \begin{cases} 0, & \text{se } |y - f(\mathbf{x})| \leq \epsilon \\ |y - f(\mathbf{x})| - \epsilon & \text{se } |y - f(\mathbf{x})| > \epsilon \end{cases} \quad (8)$$

Aqui, ϵ define a largura do tubo ao redor da função de regressão onde nenhum erro é penalizado. Somente valores fora dessa faixa de tolerância recebem uma penalidade.

Uma das principais vantagens da SVR é a capacidade de usar funções *kernel* para modelar dados não lineares. As funções *kernel* permitem mapear os dados para um espaço de maior dimensionalidade, onde as relações entre as variáveis podem ser linearmente separáveis. Um *kernel* comum é o de base radial (RBF: *radial basis function*), dado por:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \quad (9)$$

em que σ controla a amplitude do *kernel* e afeta a complexidade do modelo.

2.2.4 Plataforma

Todas as simulações foram realizadas por meio da plataforma Weka (*Waikato Environment for Knowledge Analysis*: <https://ml.cms.waikato.ac.nz/weka>), um pacote de software de código aberto para mineração de dados e aprendizado de máquina desenvolvido na Universidade de Waikato, Nova Zelândia. O Weka oferece uma interface gráfica que permite aos usuários explorar uma variedade de algoritmos de aprendizado de máquina, incluindo classificação, regressão, clusterização e seleção de atributos, sem a necessidade de programação.

Para a aplicação de RNA, KNN e SVR no Weka, foram utilizadas, respectivamente, as funções MultilayerPerceptron, IBk e SM0reg. Por padrão, MultilayerPerceptron e SM0reg realizam automaticamente a normalização dos dados de entrada. Os algoritmos e o uso do software são detalhados no livro de Witten et al. (2016), desenvolvido pelos criadores do programa.

2.3 Estudo de Caso e Obtenção dos Dados

O estudo de caso deste trabalho refere-se à bacia de contribuição do reservatório de Sobradinho, na Bahia, localizada dentro da bacia hidrográfica do rio São Francisco. Para esta pesquisa, foram utilizadas oito sub-bacias, numeradas de acordo com a classificação da Agência Nacional de Águas e Saneamento Básico (ANA), todas localizadas a montante de Sobradinho, e que, portanto, fazem parte de sua área de drenagem: 40, 41, 42, 43, 44, 45, 46 e 47. Essas sub-bacias contribuem para o reservatório de Sobradinho. Os mapas de todas as sub-bacias, com estações da rede hidrometeorológica nacional (fluviométricas e pluviométricas), podem ser encontrados no portal de metadados da ANA (<http://metadados.ana.gov.br>).

Os dados pluviométricos usados na modelagem foram obtidos do portal HidroWeb (<https://www.snirh.gov.br/hidroweb>), plataforma da ANA que permite o acesso a informações de estações hidrometeorológicas. Por outro lado, os dados de vazões naturais afluentes ao reservatório de Sobradinho foram baixados do portal da Câmara de Comercialização de Energia Elétrica (CCEE) (<https://www.ccee.org.br>).

Tabela 2 – Informações sobre os postos pluviométricos nas sub-bacias do barramento de Sobradinho

Sub-bacia	Total	Operando	Responsável - ANA	Dados Utilizados
40	299	183	60	58
41	211	130	46	40
42	175	100	39	37
43	48	30	14	13
44	179	101	44	31
45	142	94	52	36
46	281	169	87	40
47	170	68	24	24

Fonte: autores (2025)

Para realizar a modelagem, foram utilizadas vazões mensais médias acumuladas e precipitação para o período de janeiro de 1931 a dezembro de 2020 (89 anos). A quantidade de postos utilizados para coleta de dados pluviométricos pode ser observada na Tabela 2, sendo que os postos com dados incompletos foram descartados. Assim, inicialmente, foi calculada a média dos dados pluviométricos de

cada sub-bacia, desconsiderando as falhas, e posteriormente foi calculada a média das chuvas médias nas sub-bacias para obter a série temporal de precipitações médias.

2.4 Processamento dos Dados e Configuração dos Algoritmos

2.4.1 Processamento dos Dados

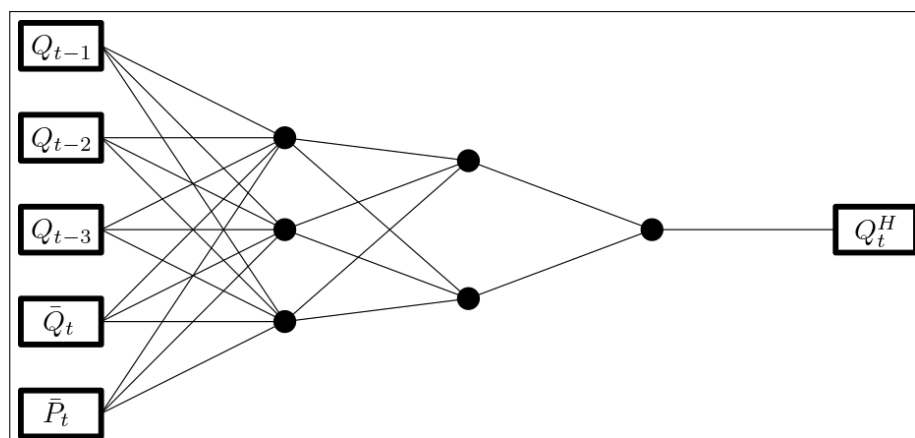
Os dados coletados para a modelagem foram submetidos a dois tipos de separação. As primeiras simulações foram realizadas utilizando validação cruzada do tipo *k-fold*, que consiste em dividir o conjunto de dados em *k* partes ou *folds* de tamanhos aproximadamente iguais. O modelo é treinado utilizando $k - 1$ partes e testado na parte restante. Esse processo é repetido *k* vezes, de modo que, a cada repetição, um *fold* diferente é utilizado para teste e os outros para treinamento. No final, os resultados de cada teste são combinados para obter uma estimativa média do desempenho do modelo. Utilizou-se o valor $k = 10$, padrão do Weka.

Os valores de entrada e saída obtidos por validação cruzada são embaralhados em relação à ordem original, o que inviabiliza sua aplicação na operação do reservatório, que deve ser gerida de forma sequencial, mês a mês. Para permitir a comparação com os resultados da validação cruzada e, principalmente, viabilizar a futura operação do reservatório, foram realizadas simulações adicionais, dividindo os dados em conjuntos de treinamento e validação. Nesse caso, o conjunto de treinamento foi composto por 70% dos dados, enquanto o conjunto de validação incluiu os 30% restantes, selecionados aleatoriamente. No presente artigo, essa abordagem é denominada calibração-validação.

2.4.2 Configuração RNA

Ao selecionar a RNA, o Weka gera uma configuração padrão com base nos dados de entrada, mas também é possível modificar essa configuração e criar uma própria topologia de RNA, editando o número de neurônios e a quantidade de camadas ocultas. Além disso, o software disponibiliza a visualização da composição da RNA configurada ou padrão, conforme exemplificado na Figura 1.

Figura 1 – Adaptação de visualização do Weka para a RNA-3,2 da formulação 6



Fonte: autores (2025)

Neste trabalho, diversos arranjos diferentes para RNA foram testados. Os arranjos foram escolhidos de maneira heurística e se repetiram para todas as formulações apresentadas na Tabela 1. Os arranjos das RNA podem ser visualizados na Tabela 3, em que os valores numéricos representam a quantidade de neurônios em uma camada oculta, enquanto a vírgula indica a adição de novas camadas. No Weka, a RNA é treinada, ou seja, seus pesos são ajustados, pelo algoritmo de retropropagação de erro (*backpropagation*) (Witten et al., 2016).

Tabela 3 – Topologias RNA investigadas

RNA-2	RNA-3,2	RNA-5,4,4	RNA-8,6,4
RNA-2,1	RNA-3,3	RNA-5,5,4	RNA-8,6,8
RNA-2,2	RNA-4,3,2	RNA-5,5	RNA-8,7,6
RNA-2,2,1	RNA-4,3,3	RNA-6,4,2	RNA-8,8,7
RNA-3,2,1	RNA-4,4	RNA-6,6,4	RNA-8,8,8

Fonte: autores (2025)

2.4.3 Configuração KNN

Várias simulações foram conduzidas utilizando o método KNN, conforme detalhado na Tabela 4. O número de vizinhos k foi selecionado empiricamente. Vale ressaltar que foi empregada a abordagem mais básica do KNN (média aritmética), sem a utilização de funções *kernel* que dão mais pesos aos vizinhos mais próximos ao ponto de consulta.

Tabela 4 – Configurações KNN investigadas

1-NN	7-NN	15-NN
3-NN	10-NN	17-NN
5-NN	13-NN	20-NN

Fonte: autores (2025)

2.4.4 Configuração SVR

Ao escolher a função SVR, o Weka permite a escolha do tipo de *kernel* a ser empregado na modelagem, bem como os parâmetros de configuração dentro do *kernel* escolhido. Neste trabalho, foram utilizados o Polykernel (*polynomial kernel*) e o RBFKernel (*radial basis function*).

2.5 Avaliação de Eficiência dos Modelos

A avaliação do desempenho dos modelos RNA, KNN e SVR foi realizada com base nos resultados obtidos pelo coeficiente de eficiência de Nash-Sutcliffe (Nash & Sutcliffe, 1970):

$$NSE = 1 - \frac{\sum_{i=1}^m (q_i - Q_i)^2}{\sum_{i=1}^m (q_i - \bar{q})^2} \quad (10)$$

em que q_i é o valor observado, Q_i é o valor modelado; \bar{q} é a média dos valores observados, e m é o número de amostras (instâncias). As eficiências de calibração, validação e total (utilizando todos os dados disponíveis) foram calculadas. Quando a validação dos dados foi realizada por meio da validação cruzada, o NSE foi calculado para os dados modelados.

A eficiência de Nash-Sutcliffe varia de $-\infty$ a 1, sendo que valores mais próximos de 1 indicam um melhor desempenho do modelo. Neste trabalho, o valor obtido para o NSE foi classificado conforme abaixo (Moriasi et al., 2007):

- Muito bom: $0,75 < NSE \leq 1,00$
- Bom: $0,65 < NSE \leq 0,75$
- Satisfatório: $0,50 < NSE \leq 0,65$
- Insatisfatório: $NSE \leq 0,50$

2.6 Operação de Sobradinho

A operação de um reservatório por otimização estocástica implícita envolve a análise de diversos cenários distintos de vazões afluentes ao reservatório, os quais podem ser derivados por modelos de geração de vazões sintéticas, que preservam as propriedades estatísticas históricas nos dados sintéticos. Posteriormente, esses cenários são submetidos a um modelo de otimização determinística sob previsão perfeita, resultando em um banco de dados contendo informações sobre armazenamento e alocação ao longo do tempo. Os dados são então analisados para identificar uma relação da alocação com o armazenamento e a afluência prevista, gerando regras operacionais ótimas (Celeste & Billib, 2009).

Neste trabalho, foi analisado um cenário futuro, com 110 anos de afluências mensais. Fez-se o uso da OEI auxiliada pelo melhor modelo de previsão de afluências encontrado nesta pesquisa. Para gerar regras operacionais, a OEI utilizou a metodologia OEI-LTF-KNN proposta por Santana & Celeste (2021), segundo a qual a alocação R_t proposta para o mês t é dada por

$$R_t = \mathcal{G}(S_{t-1}, Q_t^H) \quad (11)$$

em que S_{t-1} é o armazenamento no início do mês t , Q_t^H é a vazão afluente total esperada para o horizonte H , e \mathcal{G} é a regra operacional. No modelo OEI-LTF-KNN, a regra \mathcal{G} é implementada pelo algoritmo KNN.

Os resultados obtidos foram comparados com regras geradas por diferentes métodos: política de operação padrão (SOP: *standard operating policy*), otimização determinística sobre previsão perfeita (ODPP) e programação dinâmica estocástica (PDE), aplicados ao mesmo cenário. Todos esses métodos são detalhados em Santana & Celeste (2021). O parâmetro adotado para comparar as diferentes abordagens foi a taxa de vulnerabilidade α , expressa pela equação

$$\alpha = \frac{1}{N} \sum_{t=1}^N \left(\frac{R_t - D_t}{D_t} \right)^2 \quad (12)$$

onde D_t é a demanda no mês t e N é o período total de operação. Quanto menor o valor de α , melhor é a alocação de água.

3 RESULTADOS E DISCUSSÃO

3.1 Apresentação dos Resultados

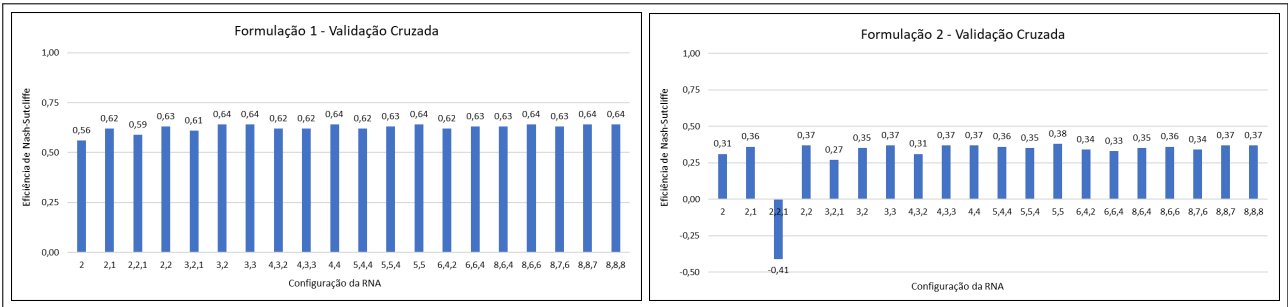
Os resultados das modelagens DDM empregadas com cada formulação proposta foram explicitados por meio de gráficos, que relacionam o coeficiente de Nash-Sutcliffe para os dados de validação com a configuração utilizada na modelagem. Os resultados serão apresentados de acordo com o tipo de validação empregada, seja validação cruzada ou calibração-validação. Posteriormente, serão apresentados os resultados obtidos em relação à aplicação do melhor modelo para a operação do reservatório.

3.2 Resultados Obtidos por Validação Cruzada

3.2.1 Resultados RNA

Para obter os resultados da modelagem RNA, foram utilizadas as configurações da Tabela 3. Os resultados de eficiência de Nash-Sutcliffe para as formulações 1 até 6 podem ser observados nas Figuras 2 até 4, respectivamente.

Figura 2 – Eficiências de Nash-Sutcliffe para formulações 1 e 2 com RNA (validação cruzada)



Fonte: autores (2025)

Figura 3 – Eficiências de Nash-Sutcliffe para formulações 3 e 4 com RNA (validação cruzada)

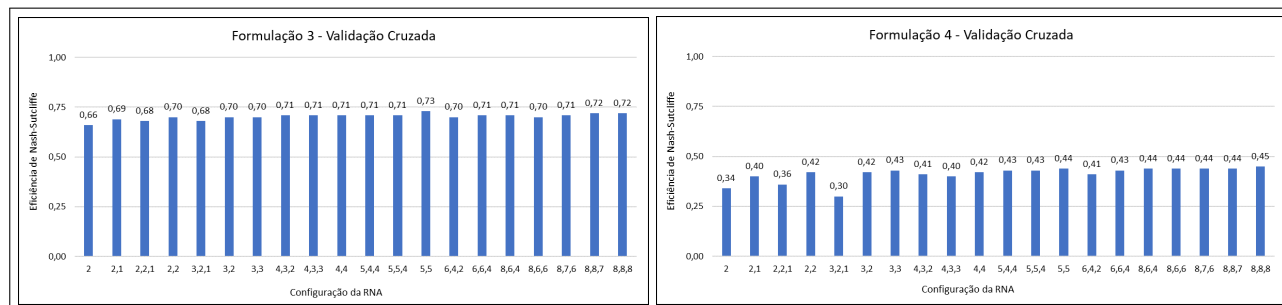
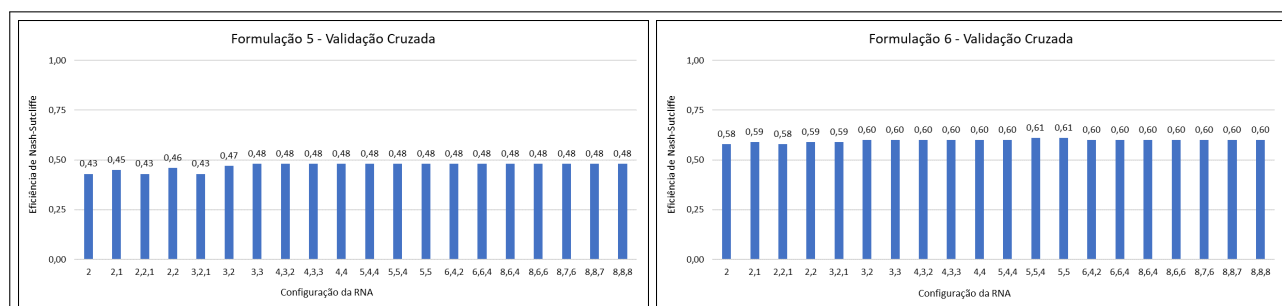


Figura 4 – Eficiências de Nash-Sutcliffe para formulações 5 e 6 com RNA (validação cruzada)

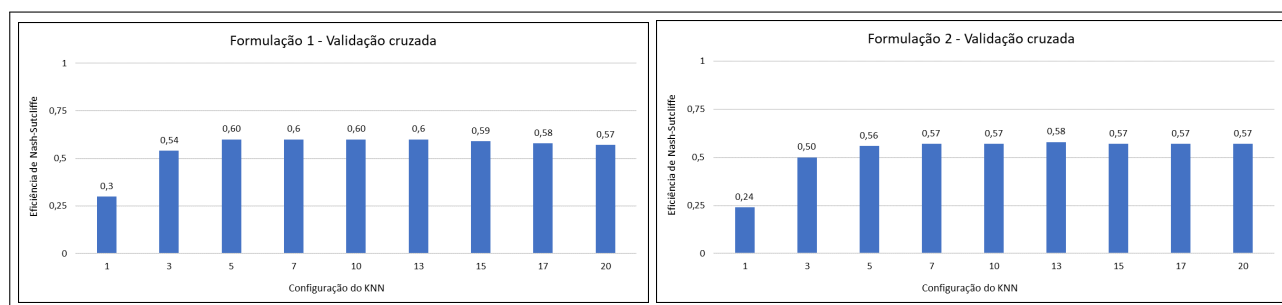


Fonte: autores (2025)

3.2.2 Resultados KNN

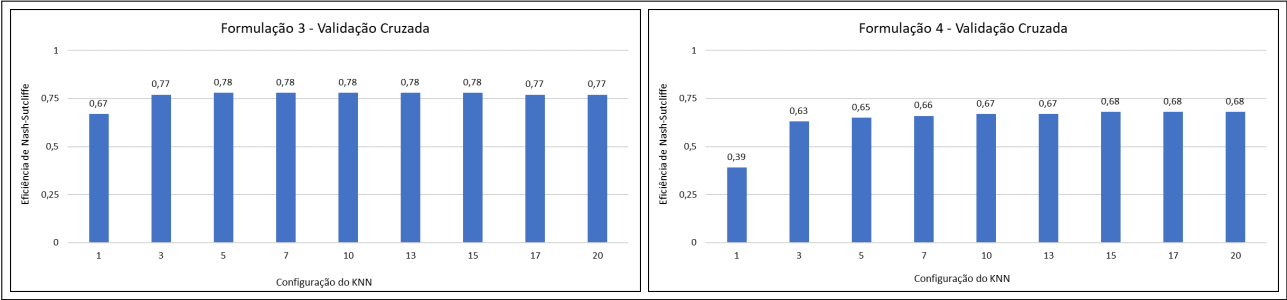
Para obter os resultados da modelagem KNN, foram utilizadas as configurações da Tabela 4. Os resultados de eficiência de Nash-Sutcliffe para as formulações 1 até 6 podem ser observados nas Figuras 5 até 7, respectivamente.

Figura 5 – Eficiências de Nash-Sutcliffe para formulações 1 e 2 com KNN (validação cruzada)



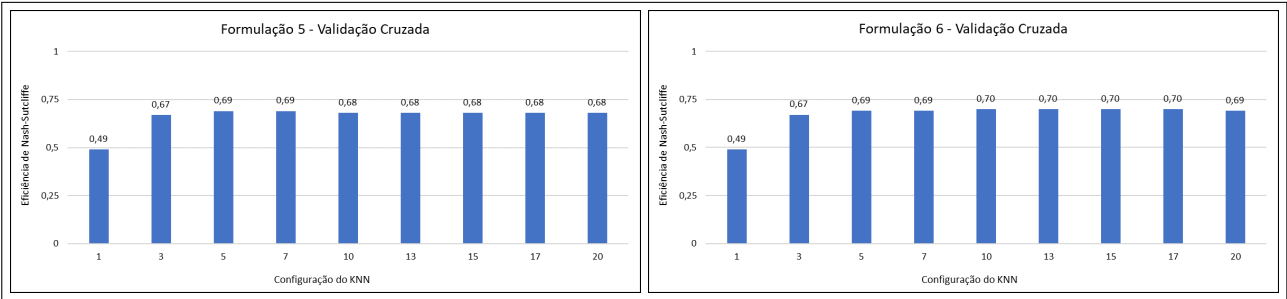
Fonte: autores (2025)

Figura 6 – Eficiências de Nash-Sutcliffe para formulações 3 e 4 com KNN (validação cruzada)



Fonte: autores (2025)

Figura 7 – Eficiências de Nash-Sutcliffe para formulações 5 e 6 com KNN (validação cruzada)

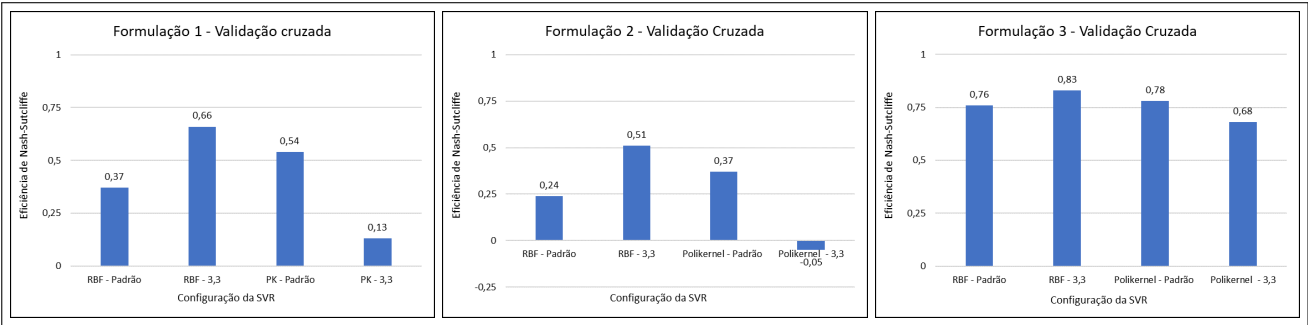


Fonte: autores (2025)

3.2.3 Resultados SVR

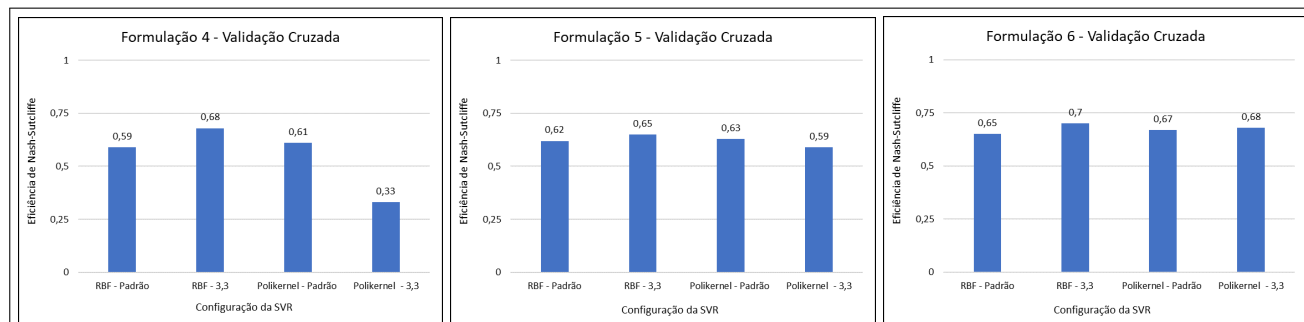
Os resultados de eficiência de Nash-Sutcliffe para as formulações 1 até 6 via SVR podem ser observados nas Figuras 8 e 9, respectivamente.

Figura 8 – Eficiências de Nash-Sutcliffe para formulações 1, 2 e 3 com SVR (validação cruzada)



Fonte: autores (2025)

Figura 9 – Eficiências de Nash-Sutcliffe para formulações 4, 5 e 6 com SVR (validação cruzada)



Fonte: autores (2025)

3.2.4 Melhores Resultados para Validação Cruzada

Os resultados com os melhores valores do coeficiente de Nash-Sutcliffe e correspondentes configurações para os dados de validação cruzada das modelagens RNA, KNN e SVR são listados a seguir. Os valores de *NSE* estão entre parênteses.

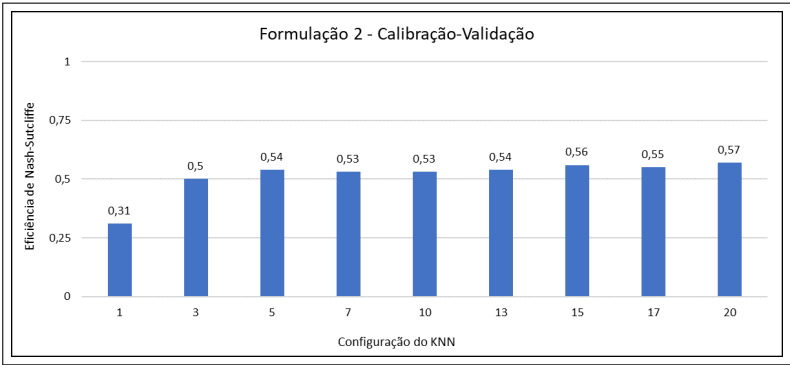
- Formulação 1: RNA-3,2 (0,64); 5-NN (0,60); SVR-RBF (0,66)
- Formulação 2: RNA-5,5 (0,38); 13-NN (0,58); SVR-RBF (0,51)
- Formulação 3: RNA-5,5 (0,73); 5-NN (0,78); SVR-RBF (0,83)
- Formulação 4: RNA-8,8,8 (0,45); 15-NN (0,68); SVR-RBF (0,68)
- Formulação 5: RNA-3,3 (0,48); 5-NN (0,69); SVR-RBF (0,65)
- Formulação 6: RNA-5,5,4 (0,64); 10-NN (0,70); SVR-RBF (0,70)

3.3 Resultados Obtidos por Calibração-Validação

3.3.1 Resultados RNA

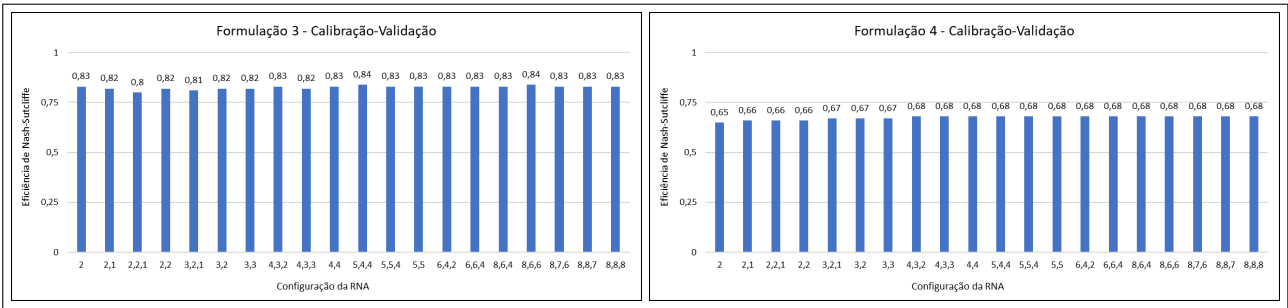
Para obter os resultados da modelagem RNA, foram utilizadas as configurações da Tabela 3. Os resultados de eficiência de Nash-Sutcliffe para as formulações 2 até 6 podem ser observados nas Figuras 10 até 12, respectivamente.

Figura 10 – Eficiências de Nash-Sutcliffe para formulação 2 com RNA (calibração-validação)



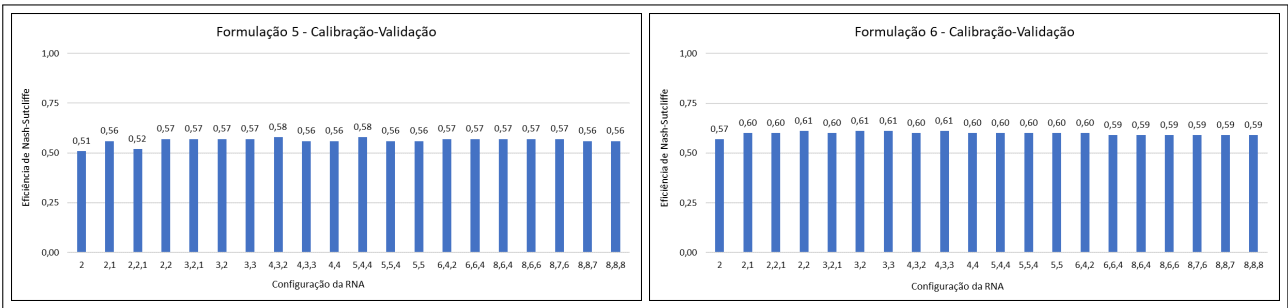
Fonte: autores (2025)

Figura 11 – Eficiências de Nash-Sutcliffe para formulações 3 e 4 com RNA (calibração-validação)



Fonte: autores (2025)

Figura 12 – Eficiências de Nash-Sutcliffe para formulações 5 e 6 com RNA (calibração-validação)

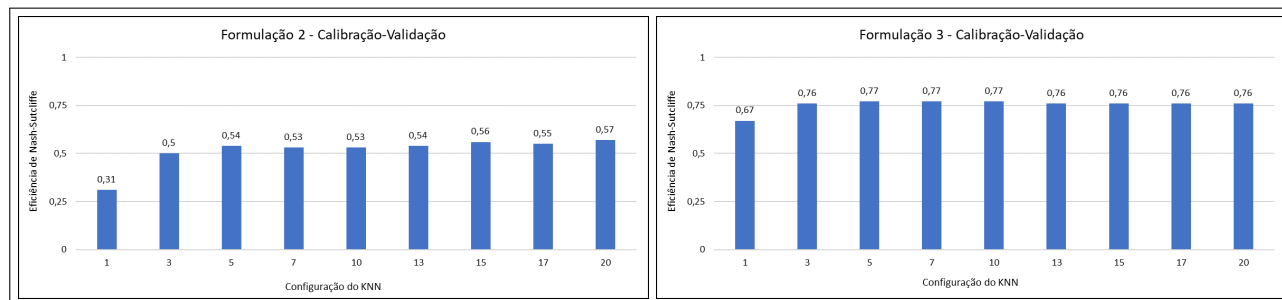


Fonte: autores (2025)

3.3.2 Resultados KNN

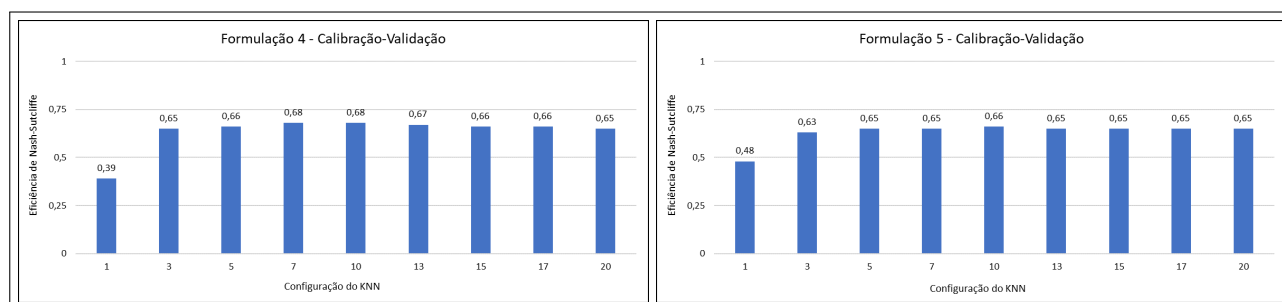
Para obter os resultados da modelagem KNN, foram utilizadas as configurações da Tabela 4. Os resultados de eficiência de Nash-Sutcliffe para as formulações 2 até 6 podem ser observados nas Figuras 13 até 15, respectivamente.

Figura 13 – Eficiências de Nash-Sutcliffe para formulações 2 e 3 com KNN (calibração-validação)



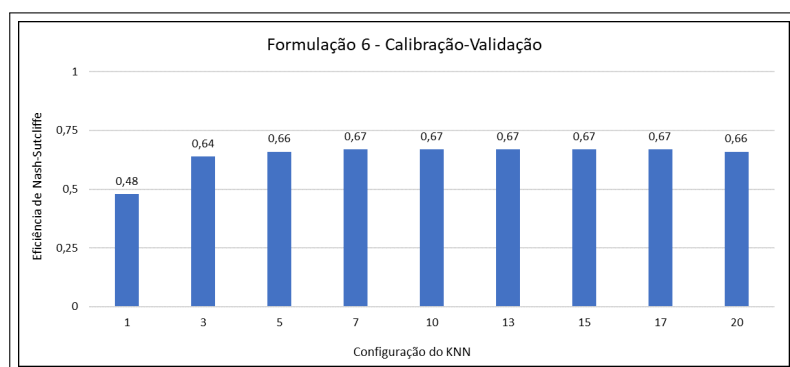
Fonte: autores (2025)

Figura 14 – Eficiências de Nash-Sutcliffe para formulações 4 e 5 com KNN (calibração-validação)



Fonte: autores (2025)

Figura 15 – Eficiências de Nash-Sutcliffe para formulação 6 com KNN (calibração-validação)

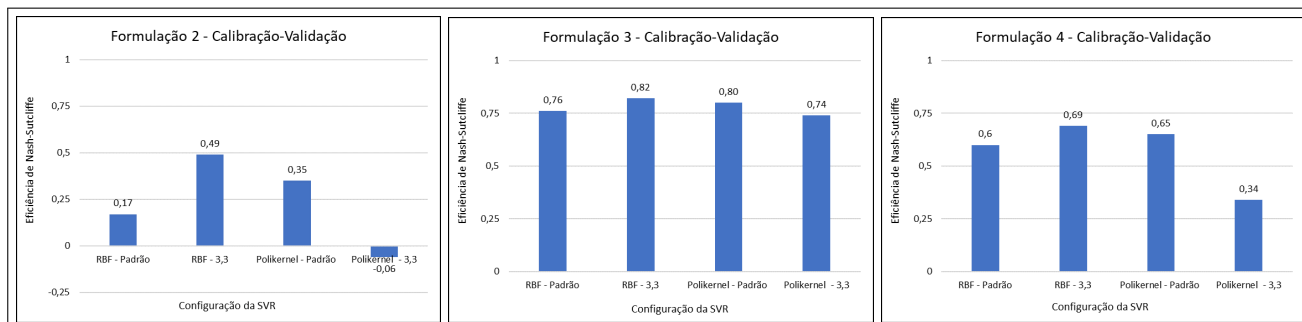


Fonte: autores (2025)

3.3.3 Resultados SVR

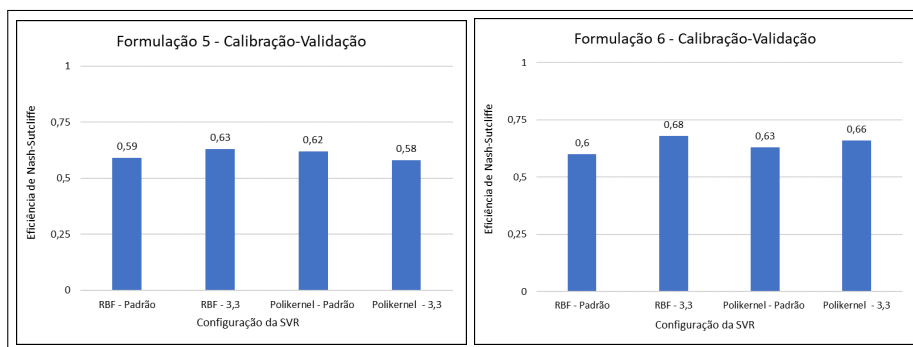
Os resultados de eficiência de Nash-Sutcliffe para as formulações 2 até 6 via SVR podem ser observados nas Figuras 16 e 17, respectivamente.

Figura 16 – Eficiências de Nash-Sutcliffe para formulações 2, 3 e 4 com SVR (calibração-validação)



Fonte: autores (2025)

Figura 17 – Eficiências de Nash-Sutcliffe para formulações 5 e 6 com SVR (calibração-validação)



Fonte: autores (2025)

3.3.4 Melhores Resultados para Calibração-Validação

Os resultados com os melhores valores do coeficiente de Nash-Sutcliffe e correspondentes configurações para os dados de calibração-validação das modelagens RNA, KNN e SVR são listados a seguir. Os valores de *NSE* estão entre parênteses.

- Formulação 2: RNA-3,3 (0,53); 20-NN (0,57); SVR-RBF (0,49)
- Formulação 3: RNA-2 (0,83); 5-NN (0,77); SVR-RBF (0,82)
- Formulação 4: RNA-4,3,2 (0,68); 7-NN (0,68); SVR-RBF (0,69)
- Formulação 5: RNA-4,3,2 (0,58); 5-NN (0,65); SVR-RBF (0,63)
- Formulação 6: RNA-2,2 (0,61); 5-NN (0,67); SVR-RBF (0,68)

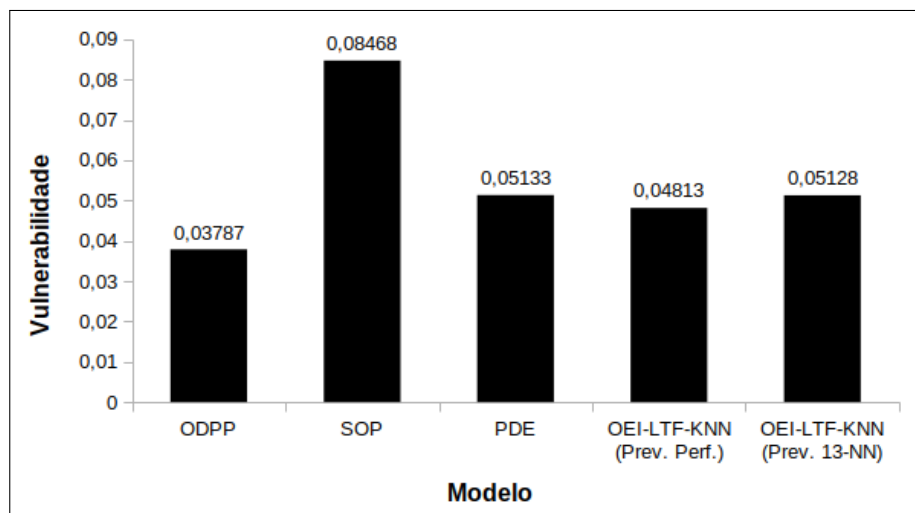
3.4 Resultados da Operação de Sobradinho

Os resultados com os melhores valores do coeficiente de Nash-Sutcliffe para os dados de calibração-validação das modelagens RNA, KNN e SVR para o cenário de operação do reservatório são listados a seguir. Os valores de NSE estão entre parênteses.

- Formulação 6: RNA-2,2 (0,63); 13-NN (0,68); SVR-RBF (0,66)

O desempenho da operação do reservatório para os diferentes modelos de operação foi explicitado por meio de gráfico que indica a vulnerabilidade para cada modelo operacional. A Figura 18 torna possível avaliar o desempenho da aplicação do modelo OEI-LTF-KNN contra os modelos SOP, ODPP e PDE. O OEI-LTF-KNN foi aplicado com previsão perfeita de Q_t^H (conforme Santana & Celeste (2021)) e com previsão de Q_t^H pelo modelo de previsão que forneceu melhor NSE (Formulação 6 com 13-NN, conforme mencionado acima).

Figura 18 – Vulnerabilidades obtidas para cada modelo de operação do reservatório



Fonte: autores (2025)

Os resultados elencados na Tabela 5 mostram a vulnerabilidade de cada modelo comparadas em termos de porcentagem ao resultado obtido pela ODPP (usada como referência).

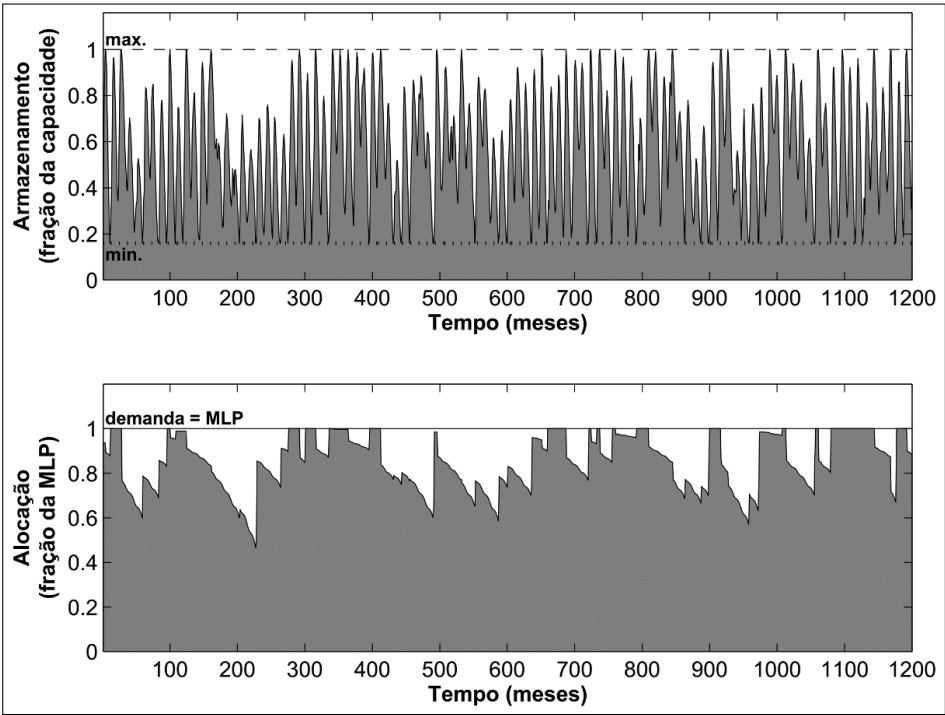
Tabela 5 – Percentagem das vulnerabilidades em relação à ODPP

SOP	PDE	OEI-LTF-KNN	
		Prev. Perf.	Prev. 13-NN
123,58	35,53	27,08	35,39

Fonte: autores (2025)

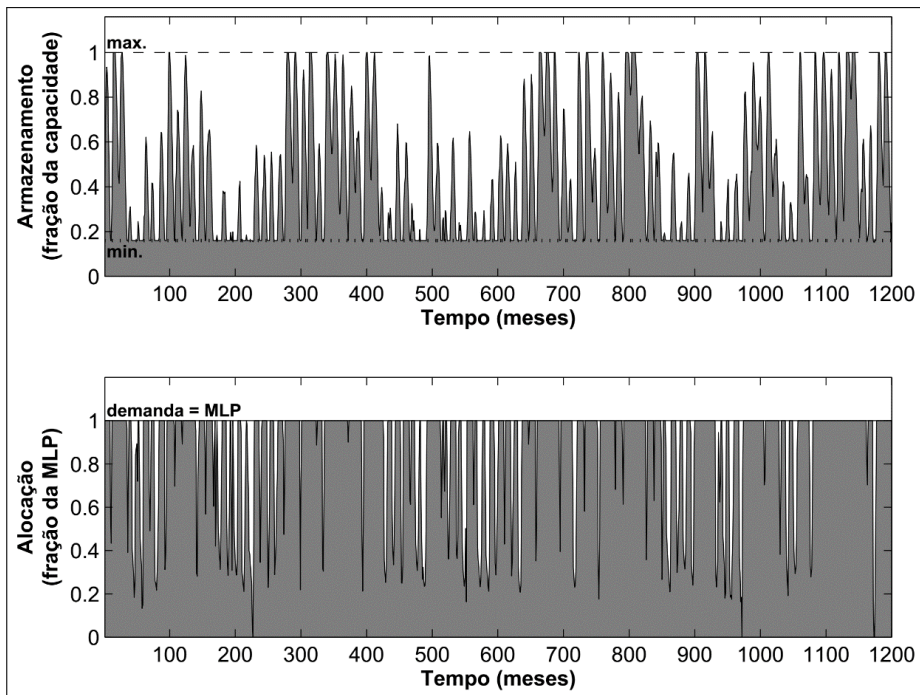
Para o cenário de afluência em cada modelo, foram gerados gráficos de armazenamento e alocação versus tempo. As Figuras 19 até 23 mostram os resultados obtidos na operação. Nesses gráficos, a alocação é representada em fração da média de longo período (MLP).

Figura 19 – Operação do reservatório via ODPP



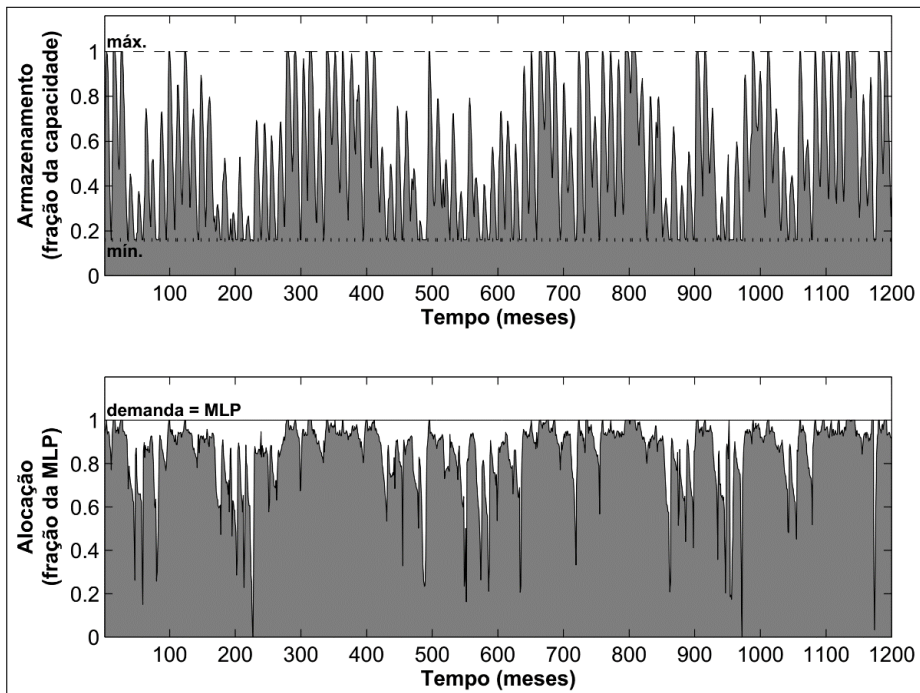
Fonte: autores (2025)

Figura 20 – Operação do reservatório via SOP



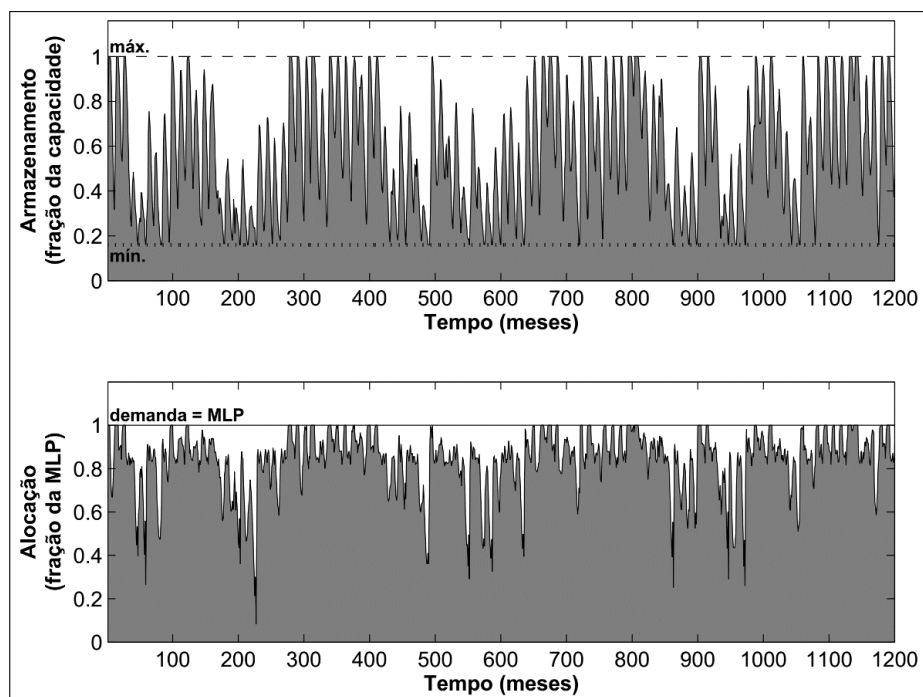
Fonte: autores (2025)

Figura 21 – Operação do reservatório via PDE



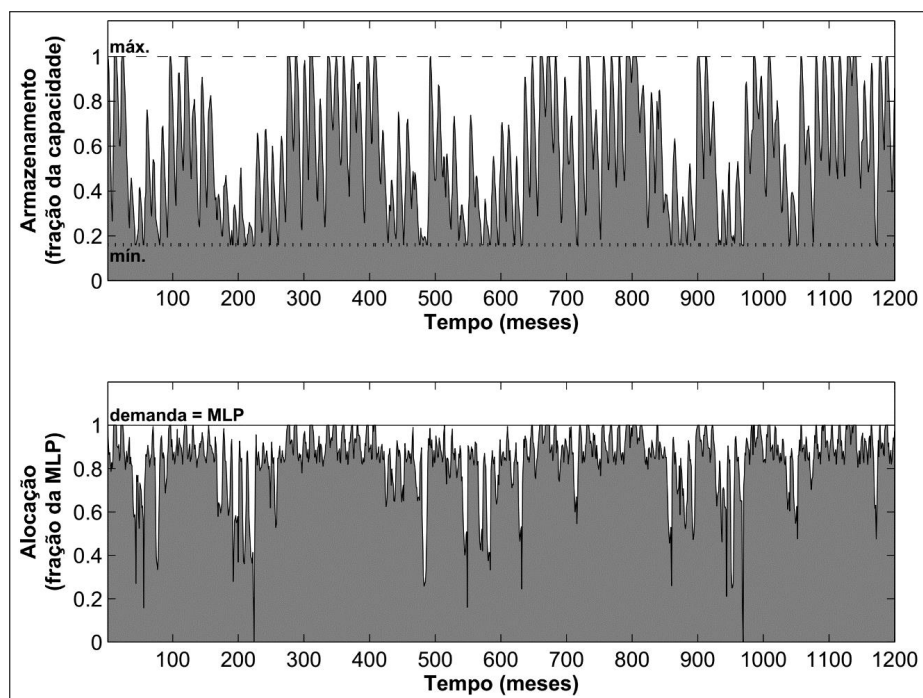
Fonte: autores (2025)

Figura 22 – Operação do reservatório via OEI-LTF-KNN Prev. Perf



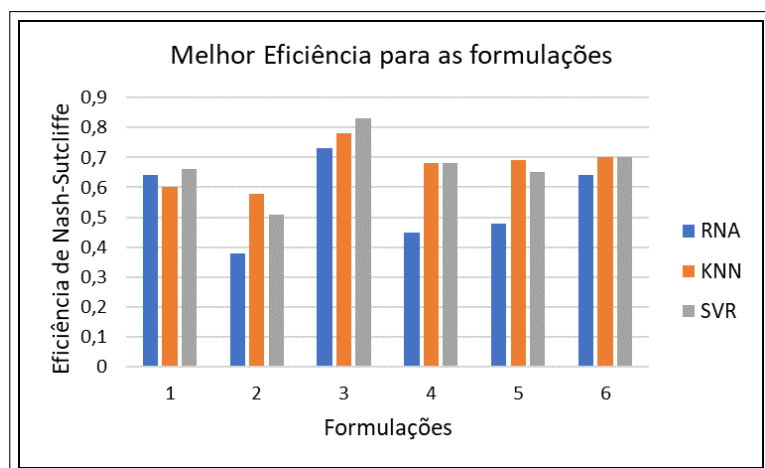
Fonte: autores (2025)

Figura 23 – Operação do reservatório via OEI-LTF-KNN Prev. 13-NN



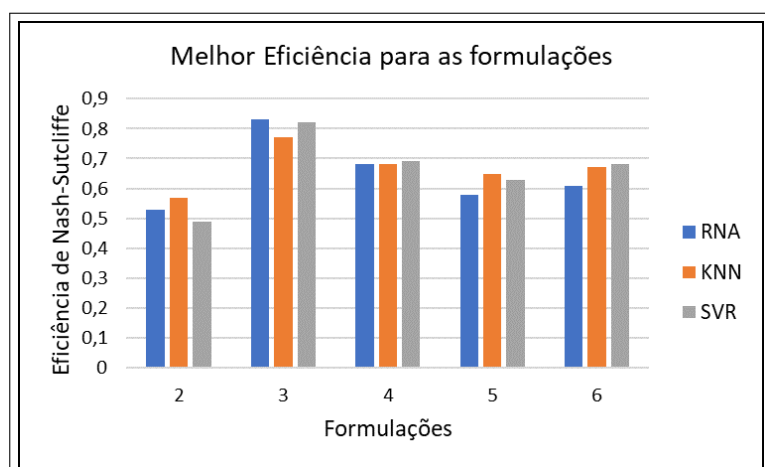
Fonte: autores (2025)

Figura 24 – Melhores eficiências para validação cruzada



Fonte: autores (2025)

Figura 25 – Melhores eficiências para calibração-validação



Fonte: autores (2025)

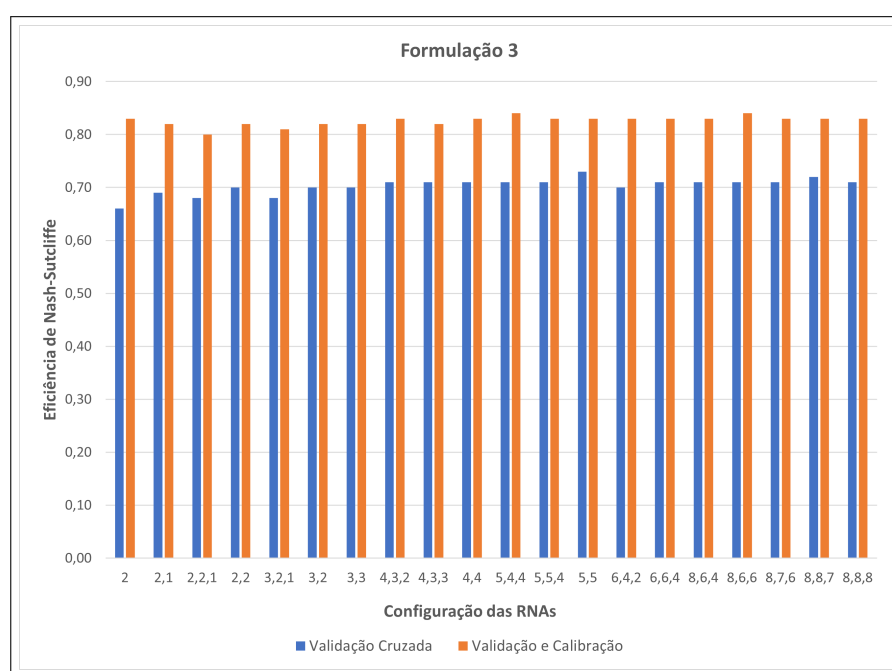
3.5 Discussão dos Resultados

Ao analisar os resultados das modelagens, verificou-se que, entre as formulações avaliadas, as que apresentaram melhor desempenho foram a 3 e a 6. Estas formulações, ao contrário das demais, incluíram variáveis adicionais como precipitação, precipitação média e vazão média. Conclui-se, portanto, que esses atributos contribuíram de forma significativa, onde a vazão média restringiu a previsão para valores mais próximos da média, minimizando grandes discrepâncias entre o valor previsto e o valor observado, enquanto a precipitação foi um parâmetro favorável ao processo de transformação de chuva em vazão. As melhores formulações

para validação cruzada e calibração-validação podem ser observadas nas Figuras 24 e 25, respectivamente.

Em relação à comparação da validação do modelo por meio de validação cruzada e calibração-validação, observou-se que a modelagem baseada em RNA apresentou melhores resultados quando validada por calibração-validação, conforme ilustrado na Figura 26, tendo a formulação 3 como exemplo. Por outro lado, as modelagens baseadas em KNN e SVR, em sua maioria dos resultados, mostraram desempenhos semelhantes.

Figura 26 – Variação da eficiência de Nash-Sutcliffe para formulação 3 via RNA



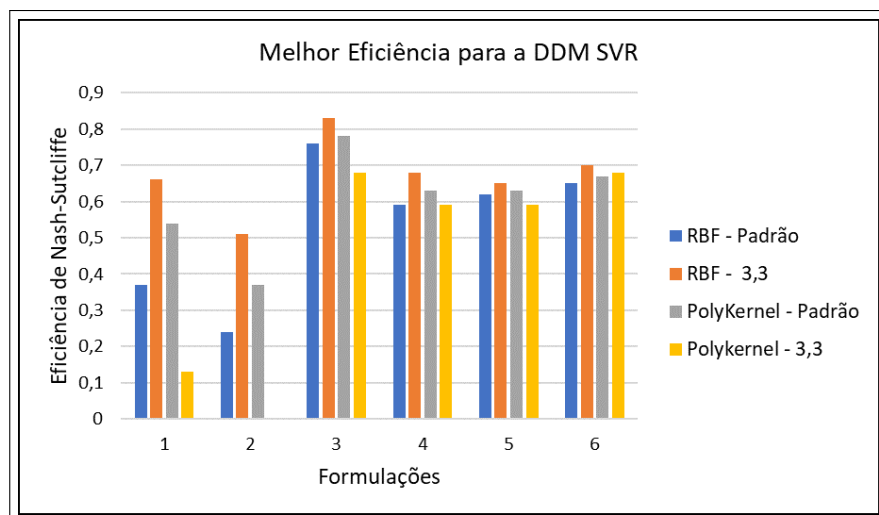
Fonte: autores (2025)

Em relação às configurações adotadas, destaca-se que, para os modelos testados por SVR, verificou-se que a configuração do RBFKernel, com os parâmetros *cacheSize* e *exponent* alterados para o valor 3 (diferentes dos valores padrões do Weka), apresentou melhores resultados. Por outro lado, o Polykernel, na maioria dos casos, teve melhores desempenhos com a configuração padrão do próprio Weka, como observado nas Figuras 27 e 28.

Os modelos baseados em KNN, na maioria das simulações, mostraram resultados melhores quando a configuração adotada foi para mais de $k = 5$ vizinhos mais próximos. Em relação às redes neurais, as variações não foram tão significativas, mas obtiveram-se melhores resultados para o esquema de calibração-validação

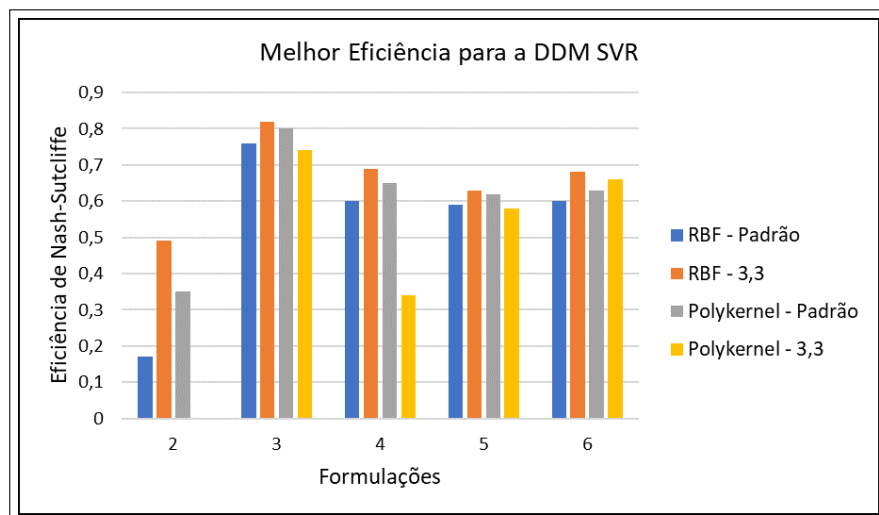
quando a configuração adotada incluiu mais de 2 camadas ocultas e 3 e 2 neurônios em cada camada, respectivamente.

Figura 27 – Variação da eficiência de Nash-Sutcliffe para as formulações via SVR (validação cruzada)



Fonte: autores (2025)

Figura 28 – Variação da eficiência de Nash-Sutcliffe para as formulações via SVR (calibração-validação)



Fonte: autores (2025)

Além disso, é importante destacar que a escolha de treinar modelos e validar de duas maneiras distintas, além de avaliar a variação nos resultados, foi motivada pela questão de os resultados do modelo, quando validados por meio de validação cruzada, saírem em posições diferentes das posições originais. Isso é crucial para a

aplicabilidade do modelo implementado pela OEI na operação do reservatório, pois essa utiliza previsões mensais para operar o reservatório. Portanto, era fundamental que os meses estivessem na ordem cronológica original. Além disso, a formulação para que o modelo seja aplicável à OEI deve conter apenas dados de vazão, ou dados adicionais de precipitação média. Nesse sentido, devido a essas restrições, a formulação 6, validada por meio de calibração-validação, apresentou melhores resultados.

Portanto, a partir de um novo cenário com 1320 instâncias de aflúências, a formulação 6 foi testada para as modelagens de RNA, KNN e SVR. Foi constatado que a melhor modelagem foi feita através do KNN para $k = 13$ vizinhos, resultando em um coeficiente de Nash-Sutcliffe de 0,68, como apresentado na Seção 3.4. Este resultado foi considerado bom de acordo com os critérios abordados na Seção 2.5. Observou-se (ver Tabela 5) que a vulnerabilidade do modelo de operação OEI-LTF-KNN (Prev. 13-NN), em comparação com a PDE, foi relativamente menor, sendo considerada satisfatória, uma vez que o modelo PDE é mais complexo. Além disso, em relação à SOP, a vulnerabilidade foi significativamente inferior. Como era esperado, o modelo ODPP apresentou as menores vulnerabilidades, já que opera sob previsão perfeita, o que na prática é impossível. Essas observações foram feitas através do gráfico da Figura 18.

4 CONCLUSÃO

A aplicação de três abordagens guiadas por dados – RNA, KNN e SVR – para prever aflúências totais futuras para um horizonte de médio prazo no reservatório de Sobradinho revelou a superioridade da SVR em comparação com as outras técnicas investigadas. Esse resultado ressalta o potencial significativo da SVR na modelagem hidrológica. Além disso, ao integrar DDM na operação estocástica implícita do reservatório, observaram-se resultados positivos, mesmo em comparação com abordagens mais complexas. Isso destaca a eficácia da metodologia OEI-LTF-KNN incorporada com previsões geradas a partir de modelos guiados por dados, que, embora mais simples, demonstrou resultados de vulnerabilidade próximos ao modelo PDE. Em resumo, este estudo contribuiu para os estudos em modelagem hidrológica ao apresentar uma abordagem diferenciada e pela aplicabilidade bem-sucedida da

previsão na operação do reservatório, demonstrando uma alternativa eficaz em relação às abordagens tradicionais baseadas em previsão perfeita.

REFERÊNCIAS

- Akbari, M., Overloop, P. J., & Afshar, A. (2010). Clustered k nearest neighbor algorithm for daily inflow forecasting. *Water Resources Management*, 25(5), 1341–1357. <https://doi.org/10.1007/s11269-010-9748-z>.
- Ávila, L., Mine, M. R. M., & Kaviski, E. (2020). Probabilistic long-term reservoir operation employing copulas and implicit stochastic optimization. *Stochastic Environmental Research and Risk Assessment*, 34(7), 931–947. <https://doi.org/10.1007/s00477-020-01826-9>.
- Celeste, A. B. & Billib, M. (2009). Evaluation of stochastic reservoir operation optimization models. *Advances in Water Resources*, 32(9), 1429–1443. <http://dx.doi.org/10.1016/j.advwatres.2009.06.008>.
- Chiamsathit, C., Adeloye, A. J., & Bankaru-Swamy, S. (2016). Inflow forecasting using artificial neural networks for reservoir operation. *Proceedings of the International Association of Hydrological Sciences*, 373, 209–214. <https://doi.org/10.5194/piahs-373-209-2016>.
- Giuliani, M., Lamontagne, J. R., Reed, P. M., & Castelletti, A. (2021). A state-of-the-art review of optimal reservoir control for managing conflicting demands in a changing world. *Water Resources Research*, 57(12), e2021WR029927. <https://doi.org/10.1029/2021WR029927>.
- Khadr, M. & Schlenkhoff, A. (2021). GA-based implicit stochastic optimization and RNN-based simulation for deriving multi-objective reservoir hedging rules. *Environmental Science and Pollution Research*, 28(15), 19107–19120. <https://doi.org/10.1007/s11356-020-12291-w>.

- Maddu, R., Pradhan, I., Ahmadisharaf, E., Singh, S. K., & Shaik, R. (2022). Short-range reservoir inflow forecasting using hydrological and large-scale atmospheric circulation information. *Journal of Hydrology*, 612, e128153. <https://doi.org/10.1016/j.jhydrol.2022.128153>.
- Morettin, P. A. & Singer, J. M. S. (2023). *Estatística e Ciência de Dados*. LTC.
- Moriasi, D. N., Arnold, J. G., Liew, M. W. V., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3), 885–900. <http://dx.doi.org/10.13031/2013.23153>.
- Nagy, I. V., Asante-Duah, K., & Zsuffa, I. (2002). *Hydrological Dimensioning and Operation of Reservoirs: Practical Design Concepts and Principles*. Springer Netherlands.
- Nash, J. & Sutcliffe, J. (1970). River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6).
- Refsgaard, J. C. (1990). Terminology, modelling protocol and classification of hydrological model codes. In Abbott, M. B. & Refsgaard, J. C., editors, *Distributed Hydrological Modelling*. (pp. 17-39). Springer Netherlands. https://doi.org/10.1007/978-94-009-0257-2_2.
- Santana, R. F. & Celeste, A. B. (2021). Stochastic reservoir operation with data-driven modeling and inflow forecasting. *Journal of Applied Water Engineering and Research*, 10(3), 212–223. <https://doi.org/10.1080/23249676.2021.1964389>.
- Shu, X., Ding, W., Peng, Y., Wang, Z., Wu, J., & Li, M. (2021). Monthly streamflow forecasting using convolutional neural network. *Water Resources Management*, 35(15), 5089–5104. <http://dx.doi.org/10.1007/s11269-021-02961-w>.
- Solomatine, D. P. (2006). Data-driven modeling and computational intelligence methods in hydrology. In Anderson, M. G. & McDonnell, J. J., editors, *Encyclopedia of Hydrological Sciences*. Wiley. <https://doi.org/10.1002/0470848944.hsa021>.

- Solomatine, D. P., Maskey, M., & Shrestha, D. L. (2007). Instance-based learning compared to other data-driven methods in hydrological forecasting. *Hydrological Processes*, 22(2), 275–287. <https://doi.org/10.1002/hyp.6592>.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. (4th ed). Morgan Kaufmann.
- Wu, M.-C., Lin, G.-F., & Lin, H.-Y. (2012). Improving the forecasts of extreme streamflow by support vector regression with the data extracted by self-organizing map. *Hydrological Processes*, 28(2), 386–397. <https://doi.org/10.1002/hyp.9584>.
- Wu, W., Eamen, L., Dandy, G., Razavi, S., Kuczera, G., & Maier, H. R. (2023). Beyond engineering: A review of reservoir management through the lens of wickedness, competing objectives and uncertainty. *Environmental Modelling & Software*, 167, e105777. <https://doi.org/10.1016/j.envsoft.2023.105777>.

Contribuições dos autores

1 – Rute Santos Porto Lima

Estudante de Engenharia Civil

<https://orcid.org/0009-0002-5274-7141> • ruteporto.engcivil@gmail.com

Contribuição: Validação; Análise Formal; Investigação; Curadoria de Dados; Escrita – Primeira Redação; Visualização de Dados

2 – Alcigeimes Batista Celeste (Autor Correspondente)

Professor Associado de Engenharia Civil

<https://orcid.org/0000-0002-3794-6583> • geimes@yahoo.com

Contribuição: Supervisão; Conceituação; Obtenção de Financiamento; Metodologia; Escrita – Revisão e Edição

Como citar este artigo

Lima, R. S. P., & Celeste, A. B. (2025). Previsão de vazões totais para horizonte de médio prazo via modelagem guiada por dados. *Ciência e Natura*, Santa Maria, v. 47, e87856. DOI: <https://doi.org/10.5902/2179460X87856>.