

# Predição das vazões observadas na bacia “Jaguari – Buenópolis” com modelos conceituais e redes neurais LSTM

Arlan dos Reis Scortegagna  
Curitiba, Brasil  
arlanscort@gmail.com

**Abstract** – Following recent applications of machine learning methods in hydrological modelling, artificial neural networks with Long Short-Term Memory (LSTMs) cells have shown competitive results when compared to consolidated rainfall-runoff models. In this work, LSTM networks were applied to predict streamflow discharges of a regulated catchment, located downstream the Reservoir Jaguari/Jacareí of Cantareira System, in São Paulo State, Brazil. The LSTMs were fine tuned and with the optimum hyperparameters they were trained both to predict streamflow series directly, given meteorological forcings and reservoir outflows as inputs, and to predict the residues of simulated streamflow series that were obtained with the consolidated Sacramento Soil Moisture Accounting (SAC-SMA) model. Therefore, the second approach, is an attempt to improve the rainfall-runoff predictions with a machine learning method, namely, the LSTM network. The results obtained were, in general, good for all models tested, while the combined model, SAC-SMA-LSTM, outperformed both SAC-SMA-only and LSTM-only.

**Keywords** – rainfall-runoff modelling, machine learning, LSTM neural networks.

## I. INTRODUÇÃO

A modelagem “chuva-vazão” consiste na técnica utilizada para estimar as vazões não observadas em uma seção de rio, denominada seção exutória, a partir das variáveis meteorológicas (precipitação, temperatura do ar, radiação solar etc.) e hidráulicas (descargas de reservatórios) observadas na bacia hidrográfica definida por essa seção. Trata-se de um dos principais problemas de engenharia na área de Hidrologia, e segue como um importante desafio e tópico de pesquisas científicas e inovações tecnológicas.

Dentre as soluções frequentemente empregadas na obtenção de séries temporais de estimativas de vazão, destacam-se os modelos hidrológicos conceituais, que são simplificações matemáticas dos processos físicos que ocorrem no ciclo hidrológico e nas bacias hidrográficas. Esses modelos vêm sendo utilizados desde os anos 1960 e, embora apresentem importantes limitações, sobretudo devido às incertezas estruturais, possuem boa aceitação entre a comunidade profissional de hidrólogos.

Há menos de uma década, técnicas de aprendizagem de máquina têm sido empregadas com crescente sucesso na modelagem “chuva-vazão”. Segundo Shen *et al.* (2021), a introdução de *Deep Learning* na Hidrologia ocorreu somente por volta de 2016 – 2018. Dentre os trabalhos mais citados desse período, destacam-se as aplicações de redes neurais recorrentes do tipo Long Short-Term Memory (LSTM) de Kratzert *et al.* (2018). Recentemente, Cho e Kim (2022) publicaram um artigo no qual utilizaram redes LSTM para aprimorar as previsões de um modelo hidrológico e apresentaram resultados bastante promissores. Esses dois artigos foram escolhidos neste trabalho para revisão e aprofundamento, embora outras referências tenham sido consultadas.

O objetivo deste trabalho é aplicar técnicas de aprendizagem de máquina, voltadas para séries temporais, a fim de aprimorar as previsões de um modelo hidrológico conceitual, que consistem em séries temporais de vazões

simuladas na seção exutória de uma bacia hidrográfica. Trata-se de um problema típico de modelagem chuva-vazão, em escala diária, que antecede a implementação do melhor modelo atingido, ou seja, daquele que apresentou o melhor desempenho na validação, em um sistema computacional para gerar previsões hidrológicas operacionais. Para tanto, foram realizados dois experimentos com redes LSTM: (a) as redes foram treinadas com as forçantes meteorológicas e hidráulicas e as variáveis de saída (alvo) foram consideradas diretamente como as vazões observadas somente; e (b) as redes foram treinadas com as mesmas variáveis de entrada, mas considerando como variáveis de saída a série temporal dos resíduos resultantes das vazões simuladas com o modelo hidrológico. Dessa forma, este trabalho buscou responder a seguinte pergunta: *para os dados em análise, é possível melhorar o desempenho das previsões hidrológicas combinando modelos conceituais com modelos de aprendizagem de máquina?*

## II. ESTADO DA ARTE

O trabalho de Kratzert *et al.* (2018) atraiu a atenção da comunidade internacional por diversos motivos. O primeiro, e provavelmente o mais importante, foi o desempenho obtido na aplicação de uma rede LSTM para a simulação de vazões diárias quando comparado a um modelo hidrológico chuva-vazão, o Sacramento Soil Moisture Accounting (SAC-SMA), que é consagrado e usado operacionalmente nos Estados Unidos da América (EUA). As LSTMs produziram resultados comparáveis ou superiores aos do SAC-SMA e, adicionalmente, apresentaram uma enorme vantagem por permitirem a utilização simultânea dos dados de diferentes bacias na predição das descargas locais, o que não é possível no método convencional de calibração do modelo hidrológico conceitual. O segundo motivo de destaque foi a utilização de um amplo conjunto de dados que contempla diversas bacias hidrográficas dos EUA, distribuídas por regiões com características geológicas e climáticas bastante distintas. Esse conjunto é denominado *Catchment Attributes for Large-Sample Studies* (CAMELS) e foi utilizado anteriormente no trabalho Newman *et al.* (2015) que, por sua vez, calibrou e validou o SAC-SMA, permitindo a Kratzert *et al.* (2018) a realização do estudo comparativo. Atenta-se para a possibilidade de ampliar os resultados desse estudo e verificar a eficiência dessa metodologia no caso de dados do Brasil, uma vez que um conjunto de dados semelhante foi produzido por Chagas *et al.* (2020), o qual foi denominado CAMELS-BR.

Além dos motivos supracitados, Kratzert e colaboradores apresentaram uma arquitetura básica que pode servir como ponto de partida para a aplicação de redes LSTM na modelagem chuva-vazão em escala diária. Eles utilizaram apenas duas camadas escondidas, com 20 neurônios LSTM cada e regularização por *dropout* de 10% entre as camadas. O treinamento foi feito com 50 épocas, *batch size* de 512 e a representação considerou 365 passos anteriores para definição do tamanho das séries de entrada. Além disso, os autores repetiram o experimento com redes neurais recorrentes, sem células LSTM, e demonstraram que as mesmas não conseguem aprender a “memória” de longo-prazo, o que já era uma limitação conhecida das RNNs, mas que foi ressaltada para o problema em estudo, uma vez que as bacias hidrológicas têm suas vazões condicionadas ao

escoamento de base (subterrâneo) que pode ser alimentado de forma extremamente lenta, a depender das condições geológicas e fisiográficas. Por fim, Kratzert *et al.* (2018) apresentaram uma interpretação hidrológica das redes LSTM, na qual correlacionaram os estados escondidos com as variáveis de entrada e argumentaram que o comportamento é semelhante aos reservatórios conceituais dos modelos chuva-vazão.

O segundo artigo revisado foi o de Cho e Kim (2022), que se destacou por ser uma publicação muito recente em uma das principais revistas científicas da área de Hidrologia. Nesse trabalho, os autores realizaram a modelagem chuva-vazão das vazões afluentes a um reservatório na Coréia do Sul utilizando o modelo hidrológico WRF-Hydro. O WRF-Hydro é considerado um dos modelos hidrológicos mais avançados na atualidade e possui um forte embasamento físico, tanto na fase de geração de escoamento superficial quanto na fase de propagação. Não obstante, os parâmetros relativos ao uso do solo e o coeficiente de Manning, por exemplo, devem ser atribuídos na escala de bacia e, portanto, calibrados de modo que podem propagar incertezas de ordem estrutural para os resultados. Para as previsões de vazão na seção exutória, o modelo deve ser forçado com oito variáveis meteorológicas e isso deve ser feito de forma distribuída, em alta resolução espacial (100m x 100m) e em escala sub-diária, ou horária. Desse modo, a demanda de dados do modelo WRF-Hydro é relativamente intensiva. Para a previsão realizada diretamente com o WRF-Hydro, os autores denominaram o experimento de *WRF-Hydro-only*.

Com as vazões resultantes desse modelo, os autores calcularam a série temporal dos resíduos, ou seja, das diferenças entre as vazões observadas e simuladas. Esses resíduos foram utilizados como variável de saída (alvo) para o treinamento de uma rede LSTM que considerou as forçantes meteorológicas como variáveis de entrada. A esse experimento, os autores denominaram *WRF-Hydro-LSTM*, uma vez que as saídas do modelo hidrológico são utilizadas indiretamente. A fim de comparar os resultados, a rede LSTM foi treinada diretamente para as vazões a partir das forçantes meteorológicas, experimento ao qual denominaram *LSTM-only*. Na validação dos três modelos, o experimento *WRF-Hydro-LSTM* mostrou um desempenho superior aos demais, embora tenha sido apenas sutilmente superior ao experimento *LSTM-only*.

Embora os autores tenham refinado as redes LSTM durante o treinamento, concluíram que a previsão de resíduos requer um esforço maior do que a previsão direta das vazões (*LSTM-only*). Uma importante limitação notada no trabalho de Cho e Kim (2022), é que os autores não especificaram o tamanho das séries de entrada, ou seja, o número de passos anteriores utilizados no treinamento das redes LSTM.

Além desses artigos, foram revisados outros trabalhos notáveis, a saber: Yuan *et al.* (2018) que lidaram com vazões mensais; Kratzert *et al.* (2019) que refinaram a metodologia aplicada em 2018; e de Hunt (2022) que se aprofundaram no melhoramento de previsões hidrológicas operacionais.

### III. METODOLOGIA

O estudo foi conduzido seguindo a proposta de Cho e Kim (2022) e atentando às recomendações de Kratzert *et al.* (2018). Todos os experimentos foram realizados por meio de programas codificados em Python 3.10 que foram

disponibilizados no repositório pessoal do GitHub, link de acesso: <https://github.com/arlanscort/machine-learning>.

Dentre as bibliotecas utilizadas, destacam-se: Pandas (versão 1.4.1), Numpy (versão 1.22.3), TensorFlow (versão 2.8.1) e Keras (2.8.1). Os programas foram executados em Debian GNU/Linux, versão 10, com processador Intel Core i7-6700 3.4 GHz, 4 núcleos, 8 threads, 8 MB de memória cache e 16 GB de memória RAM.

Os dados obtidos foram aplicados em dois experimentos: (a) *LSTM-only*, no qual as variáveis de entrada (forçantes meteorológicas e hidráulicas) foram aplicadas diretamente na previsão de vazões na seção exutória, e (b) *SAC-SMA-LSTM*, no qual essas mesmas variáveis foram aplicadas na previsão com uma rede LSTM dos resíduos da série de vazões simuladas pelo modelo SAC-SMA.

#### A. Dados

Neste trabalho, foram utilizados os dados da bacia hidrográfica localizada entre o Reservatório Jaguari - Jacaré, que integra o Sistema Cantareira no Estado de São Paulo, e o posto de controle fluviométrico código 3D-009T, denominado “Rio Jaguari em Buenópolis”, que faz parte da rede de monitoramento hidrológico do Departamento de Águas e Energia Elétrica (DAEE). Trata-se de uma bacia incremental regularizada, uma vez que recebe a montante as descargas do reservatório Jaguari – Jacaré. Possui uma área de 653,2 km<sup>2</sup> e sua localização está apresentada na Figura 1. As vazões na seção exutória são monitoradas continuamente pelo posto de controle desde meados de 2012. Por esses motivos, os dados contemplam o período de janeiro de 2012 a 29 de maio de 2022.

A bacia “Jaguari – Buenópolis” foi escolhida por estar inserida no sistema de previsão hidrometeorológica desenvolvido pelo autor, e por ter apresentado um desempenho limitado em relação às demais bacias contempladas nesse sistema. Desse modo, encontra-se disponível uma série de vazões simuladas obtidas com o modelo chuva-vazão SAC-SMA, que foi calibrado e validado pelo autor em desenvolvimentos anteriores. A essa série foi atribuída a denominação *SAC-SMA only*, e ela foi utilizada nos experimentos a fim de aprimorar as previsões obtidas com o emprego das redes LSTM.

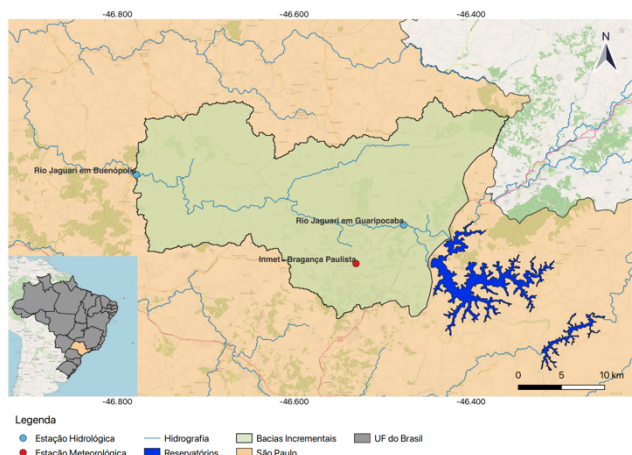


Fig. 1. Localização da bacia “Jaguari – Buenópolis”.

Todos os dados utilizados no desenvolvimento são abertos e se encontram disponíveis nos sistemas do DAEE, do Instituto Nacional de Meteorologia (Inmet) e da Companhia de Saneamento Básico do Estado de São Paulo

(Sabesp). As séries temporais usadas como variáveis de entrada foram: (a) precipitações médias espaciais (PME) na bacia, calculadas a partir dos dados dos postos “Rio Jaguari em Buenópolis” e “Rio Jaguari em Guaripocaba” (3D-015T), ambos do DAEE; (b) temperaturas do ar (TA) medidas na estação meteorológica do Inmet em Bragança Paulista, localizada próxima ao centro da bacia; e (c) vazões de montante (Qmon) ou descargas do Reservatório Jaguari-Jacaré ao Rio Jaguari, disponíveis no “Portal dos Mananciais” da Sabesp (<https://mananciais.sabesp.com.br/>). Como variável de saída (alvo), utilizou-se a série de vazões observadas a jusante da bacia (Qjus), na seção exutória, ou seja, no posto de controle Jaguari em Buenópolis. Todas as séries dessas variáveis foram agregadas para a escala diária e as eventuais falhas foram imputadas por meio de interpolação com *splines* e verificadas visualmente para fins de consistência.

### B. Experimento LSTM-only

Neste experimento, uma rede neural LSTM com duas camadas escondidas e um número variável de neurônios foi treinada, tomando como variáveis de entrada PME, TA e Qmon e como variável de saída Qjus. A rede foi configurada com uma terceira camada *fully connected* com apenas uma célula, de modo a realizar previsões para um passo de tempo de Qjus. A representação consiste em 365 passos anteriores de cada variável de entrada. O período de dados foi dividido em um conjunto de treino de jan/2013 a dez/2018, e um conjunto de validação de jan/2019 a maio/2022.

O número de neurônios foi mantido igual para ambas as camadas e variado no intervalo [16, 32, 64, 128]. Além deste, os hiperparâmetros *batch size* e a taxa de aprendizado (*learning rate*) foram variados nos intervalos [32, 64, 128, 256 e 512] e [0.001 e 0.0001], respectivamente. Os limites desses intervalos foram atribuídos com base nos trabalhos de Cho e Kim (2022) e Kratzert *et al.* (2018).

Entre as camadas escondidas, foi aplicada regularização por *dropout* a 10%. A função de perda escolhida foi o *mean squared error* (MSE), seguindo Kratzert *et al.* (2018) e o algoritmo de otimização adotado foi o Adam. Reitera-se que os códigos desenvolvidos se encontram disponíveis e minimamente detalhados no link do repositório GitHub.

### C. Experimento SAC-SMA LSTM

Neste experimento, outra rede LSTM foi treinada a partir das mesmas variáveis de entrada (PME, TA e Qmon), mas para realizar a previsão dos resíduos do modelo hidrológico, ou seja, da variável  $res(t) = Qjus(t) - Qsim(t)$ . Nesse caso, o número de passos anteriores, definido em 365 no experimento *LSTM-only*, foi variado. Isso foi feito porque o modelo hidrológico já deve considerar os efeitos de longo prazo da bacia e os resíduos mais significativos foram observados nos picos de vazão, para os quais antecedências de até 30 dias das variáveis de entrada pareciam de maior relevância.

Da mesma forma que no experimento anterior, os dados foram divididos nos conjuntos de treinamento (jan/2013 – dez/2018) e validação (jan/2019 – maio/2022).

## IV. RESULTADOS E DISCUSSÃO

Os resultados da avaliação de sensibilidade dos hiperparâmetros, referente aos treinamentos executados no experimento *LSTM-only*, estão apresentados na Figura 2. Dentre as 40 configurações testadas, o melhor resultado, ou

seja, que produziu o menor MSE para o conjunto de validação, foi atingido com os seguintes hiperparâmetros:

- número de neurônios = 64;
- *batch size* = 32;
- taxa de aprendizado = 0.001;
- e número de épocas = 5.

Analisando a Figura 2, verifica-se que essa configuração apresentou também uma variabilidade reduzida em relação aos demais resultados. Trata-se, portanto, da configuração ótima obtida nesse experimento. Nota-se que o *batch size* = 32 é muito menor do que a recomendação de Kratzert *et al.* (2018) (*batch size* = 512). Não obstante, observa-se que, a depender dos demais hiperparâmetros, mesmo para *batch size* = 512 é possível atingir resultados semelhantes em termos de MSE.

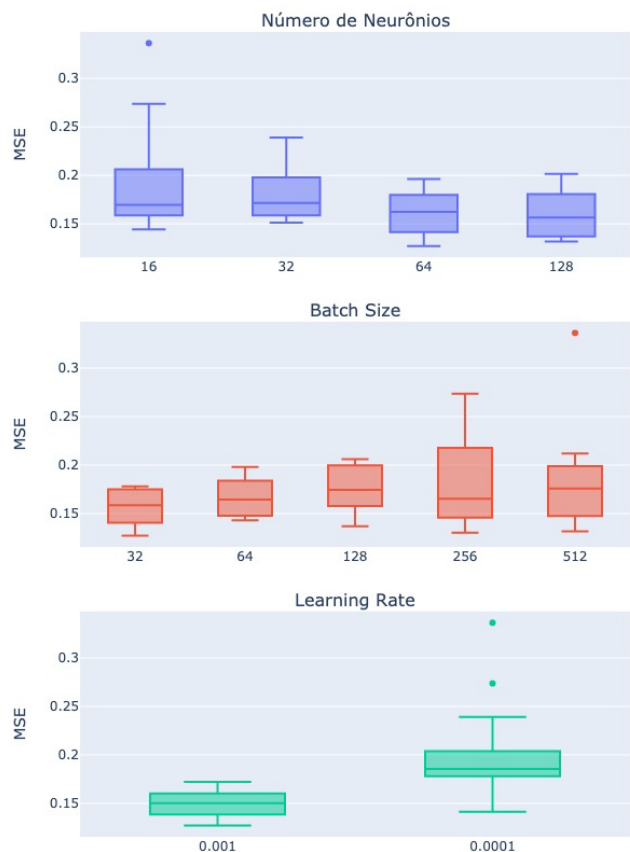


Fig. 2. Análise de sensibilidade aos hiperparâmetros para o treinamento do experimento *LSTM-only*.

Essa configuração ótima também foi empregada no experimento *SAC-SMA-LSTM*. Entende-se que essa não seja a situação ideal, e que uma avaliação de sensibilidade também deveria ter sido feita para os hiperparâmetros supracitados. Contudo, no caso desse experimento, avaliou-se a sensibilidade do treinamento somente ao número de passos anteriores, que foi variado de 1 a 365. Os resultados dessa avaliação estão apresentados na Figura 3.

De acordo com a Figura 3, para um número de passos anteriores igual a 100, e sete épocas (não apresentado), obtém-se o MSE mínimo para o conjunto de validação. Contudo, nota-se que esses resultados não são conclusivos: para cinco passos anteriores e 21 épocas também foi possível obter um MSE baixo. Não obstante, a configuração ótima para o treinamento do experimento *SAC-SMA-LSTM* foi definida em 100 passos anteriores e sete épocas.

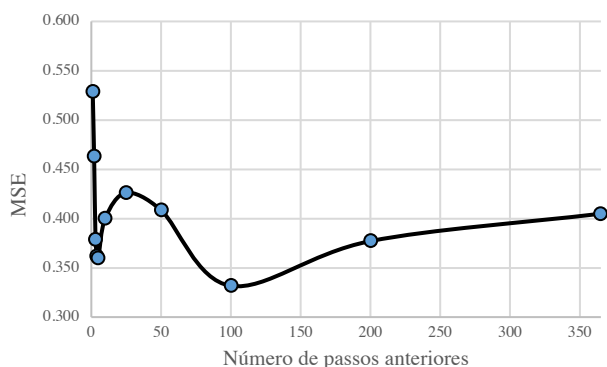


Fig. 3. Análise de sensibilidade ao número de passos anteriores usado na representação para o treinamento do experimento *SAC-SMA-LSTM*.

Uma vez definidas as configurações ótimas de ambos os experimentos, com base nas análises de sensibilidade, foram realizadas previsões para o período de validação e calculados os mesmos índices de eficiência avaliados no trabalho de Cho e Kim (2022): (a) coeficiente de eficiência de Nash-Sutcliffe (NSE); (b) NSE dos logaritmos das vazões (logNSE); e (c) viés percentual (PBIAS). Os resultados desses índices estão apresentados na Tabela 1 e as previsões de vazão de todos os modelos para último ano hidrológico completo (out/2020 – set/2021) estão ilustradas na Figura 4.

TABELA 1 – ÍNDICES DE EFICIÊNCIA NA VALIDAÇÃO

Experimento	NSE	logNSE	PBIAS (%)
<i>SAC-SMA-only</i>	0.66	0.77	8.5
<i>LSTM-only</i>	0.72	0.52	7.6
<i>SAC-SMA-LSTM</i>	0.76	0.83	8.8

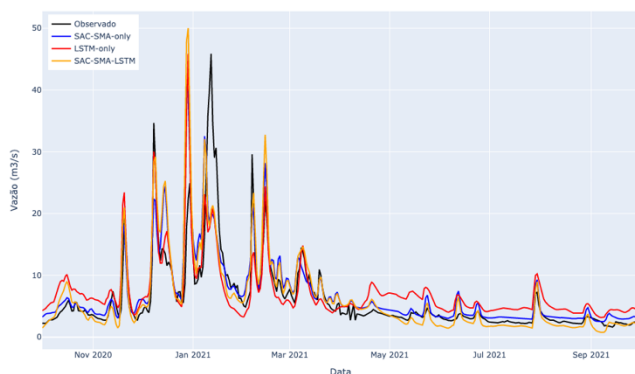


Fig. 4. Hidrograma de vazões observadas e simuladas para o ano hidrológico 2020/2021.

De acordo com a Tabela 1, para o índice PBIAS, todos os modelos atingiram desempenhos considerados “muito bons” ( $PBIAS < \pm 10\%$ ) pela literatura (Moriassi *et al.*, 2007). Já para o índice NSE, os modelos atingiram desempenhos “bons” ( $NSE > 0.65$ ) ou “muito bons” ( $NSE > 0.75$ ). O resultado modelo combinado *SAC-SMA-LSTM* possibilitou um notável aprimoramento tanto do modelo *SAC-SMA-only*, limitado pelo NSE, quanto do *LSTM-only*, limitado pelo logNSE.

A Figura 4 esclarece esses resultados. O índice NSE prioriza as vazões mais altas e, como se observa na figura, o modelo chuva-vazão tinha dificuldades em reproduzir os picos do hidrograma que foram parcialmente superadas com

o modelo combinado. A rede neural, por sua vez, era limitada nos eventos de recessão, ou seja, nas vazões mais baixas, e essa limitação também foi superada pelo modelo combinado, o que refletiu na melhora de logNSE, que prioriza as vazões menores do hidrograma.

De um modo geral, pode-se afirmar que houve um aprimoramento significativo com o modelo combinado e que as limitações e potencialidades do modelo chuva-vazão e da rede neural foram aproveitadas de forma bastante eficaz.

## V. CONCLUSÃO

Os resultados obtidos são promissores e demonstraram o potencial de aplicabilidade das redes neurais LSTM no aprimoramento de um modelo hidrológico usual. Isso ocorreu porque algumas incertezas estruturais não são capturadas pelo modelo chuva-vazão, ao passo que algumas componentes essenciais, como a representação dos escoamentos de recessão, são preteridas no modelo *data-driven*. A combinação de ambos, o SAC-SMA e a rede LSTM, se mostrou eficaz e capaz de superar as limitações individuais de cada modelo.

Ressalta-se que os resultados aqui obtidos são limitados a um caso específico e, como sugestão para desenvolvimentos futuros, recomenda-se a aplicação em outros conjuntos de dados, como o próprio CAMELS-BR. Além disso, recomenda-se que o modelo *SAC-SMA-LSTM* também seja refinado por meio de uma análise de sensibilidade de todos os hiperparâmetros, inclusive aqueles refinados apenas no experimento *LSTM-only*.

## REFERÊNCIAS

- [1] Chagas, Vinícius BP et al. CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data*, v. 12, n. 3, p. 2075-2096, 2020.
- [2] Cho, Kyeungwoo; kim, Yeonjoo. Improving streamflow prediction in the WRF-Hydro model with LSTM networks. *Journal of Hydrology*, v. 605, p. 127297, 2022.
- [3] Hunt, Kieran MR et al. Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrology and Earth System Sciences Discussions*, p. 1-30, 2022.
- [4] Kratzert, Frederik et al. Rainfall-runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, v. 22, n. 11, p. 6005-6022, 2018.
- [5] Kratzert, Frederik et al. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, v. 55, n. 12, p. 11344-11354, 2019.
- [6] Moriassi, Daniel N. Et al. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, v. 50, n. 3, p. 885-900, 2007.
- [7] Newman, A. J. Et al. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, v. 19, n. 1, p. 209-223, 2015.
- [8] Shen, Chaopeng, Xingyuan Chen, and Eric Laloy. "Broadening the Use of Machine Learning in Hydrology." *Frontiers in Water* 3 (2021): 38.
- [9] Yuan, Xiaohui et al. Monthly runoff forecasting based on LSTM-ALO model. *Stochastic environmental research and risk assessment*, v. 32, n. 8, p. 2199-2212, 2018.