XXVI SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

INTERPRETABILITY ON AGILE MACHINE LEARNING MODELS FOR HYDROLOGICAL PREDICTIONS: A CASE STUDY IN THE MEGA-DISASTER IN RIO GRANDE DO SUL, BRAZIL, IN MAY 2024

Luiz. F. Satolo [1,2]; Larissa A. da Silva [1]; Luan C. S. M. Ozelim[3]; Jaqueline A. J. P. Soares[1,2]; Matheus Correia[2]; Elton V. Escobar-Silva[1]; Glauston R. T. Lima[1], Luz A. C. Pineda[1]; Leonardo B. L. Santos[1,2]

Abstract: In May 2024, the region of the Rio Grande do Sul state experienced one of the worst floods in Brazilian history, affecting millions and causing severe damage to infrastructure. This study applies an agile hydrological forecasting approach using methods from traditional time-series models, such as ARIMA and SARIMA, and machine learning (ML) models, such as ElasticNet and LASSO. Data from streamflow monitoring stations were used to predict water levels at different lead times. The results showed that SARIMA consistently ranked among the top-performing models, while ElasticNet and LASSO demonstrated competitive performance among the ML methods. To enhance interpretability, Permutation Importance and Accumulated Local Effects were applied, highlighting the significance of autoregressive terms and upstream hydrological conditions. These findings underscore the potential of integrating traditional and ML methods in an agile approach to adaptive flood risk forecasting.

Resumo: Em maio de 2024, a região do estado do Rio Grande do Sul sofreu uma das piores enchentes da história brasileira, afetando milhões de pessoas e causando graves danos à infraestrutura. Este estudo aplica uma abordagem ágil de previsão hidrológica usando métodos de modelos tradicionais de séries temporais, como ARIMA e SARIMA, e modelos de aprendizado de máquina (*machine learning*, ML), como ElasticNet e LASSO. Dados de estações de monitoramento de vazão foram usados para prever níveis de água em diferentes horizontes de previsão. Os resultados mostraram que SARIMA consistentemente classificou-se entre os modelos de melhor desempenho, enquanto ElasticNet e LASSO demonstraram desempenho competitivo entre os métodos de ML. Para melhorar a interpretabilidade, a Importância da Permutação e os Efeitos Locais Acumulados foram aplicados, destacando a importância dos termos autorregressivos e das condições hidrológicas a montante. Essas descobertas ressaltam o potencial da integração de métodos tradicionais e de ML em uma abordagem ágil para previsão adaptativa de risco de enchentes.

Keywords: Flood, water level prediction, machine learning

1) Centro Nacional de Monitoramento e Alertas de Desastres Naturais - CEMADEN, Estrada Doutor Altino Bondesan, 500, Distrito de Eugênio de Melo, São José dos Campos/SP, CEP 12247-016, Brasil, luiz.satolo@inpe.br
2) Instituto Nacional de Pesquisas Espaciais - INPE, Avenida dos Astronautas, 1.758 - Jd. Granja - CEP 12227-010. São José dos Campos - SP - Brasil
3) Divisão de Engenharia Civil, Instituto de Tecnologia de Aeronáutica - ITA, Praça Marechal Eduardo Gomes, 50 - Vila das Acacias, São José dos Campos - SP, 12228-900, Brasil.

INTRODUCTION

In May 2024, Rio Grande do Sul suffered one of the worst floods in Brazilian history. The disaster resulted in 183 deaths, 27 missing people, more than 800 injuries, and around 600,000 people displaced. The state infrastructure was severely affected, with damage to roads, bridges, and communication systems, in addition to losses estimated at US$ 4 billion (Alcântara et al., 2024). The tragedy highlighted the region's susceptibility to extreme weather events and the need for effective disaster prevention, including decision supportive hydrological modeling.

There are three general approaches to hydrological modeling: empirical, conceptual, and physical-based. Empirical models, such as models based on traditional time series or Machine Learning, are trained from previous rain events/data and how they impact river discharge or flow - they have high predictive power and low generalizability.

Conceptual models are based on reservoir properties that require large amounts of meteorological and hydrological field data to calibrate the curve fitting. They are easy to implement but hard to interpret. Physical-based models are built on government equations and the spatial distribution of physical attributes with very high complexity (Sahu et al., 2023).

Traditional time series methods include autoregressive (AR) and moving average (MA) models. In AR models, the current value depends linearly on past values and a stochastic term, while in MA models, the dependence is on current and past values of the stochastic term. The combination of these models forms ARMA. For nonstationary series, the model is generalized as ARIMA, which can be extended to SARIMA for seasonal variations (De Gooijer & Hyndman, 2006).

Machine Learning (ML), a branch of Data Science, focuses on developing algorithms to analyze real-world datasets, enabling the interpretation of complex data and accurate decision-making. The increasing demand for interpretability in Machine Learning has elevated the role of Automatic Machine Learning (AutoML). AutoML enables rapid model testing with minimal human intervention, reducing bias in model selection and tuning.

By automating the entire ML pipeline, AutoML ensures a broader exploration of models, balancing interpretability and performance, often leading to more fine-tuned and interpretable solutions. AutoML automates model training, validation, and optimization using techniques like Bayesian optimization, reducing manual effort, and improving accuracy in dynamic scenarios. Although interpretability remains a challenge, user-defined constraints can enhance transparency (Eldeeb et al., 2024).

The applications of interpretability methods on ML models in the hydrology literature are still scarce. (Stein et al., 2021) employed the Accumulated Local Effects (ALE) in a large-sample study to evaluate how attributes influence flood processes. (Cappelli & Grimaldi, 2023) compared the performance of thirteen measures of the importance of features in hydrological applications and concluded that Permutation Importance generally provides perfect rankings for the importance of features of hydrological ML models in large samples. (Bian et al., 2023) applied the Permutation Importance on a hybrid model that combined Neural Network and LightGBM, identifying key features for the prediction of runoff in the Shiyang River Basin, China.

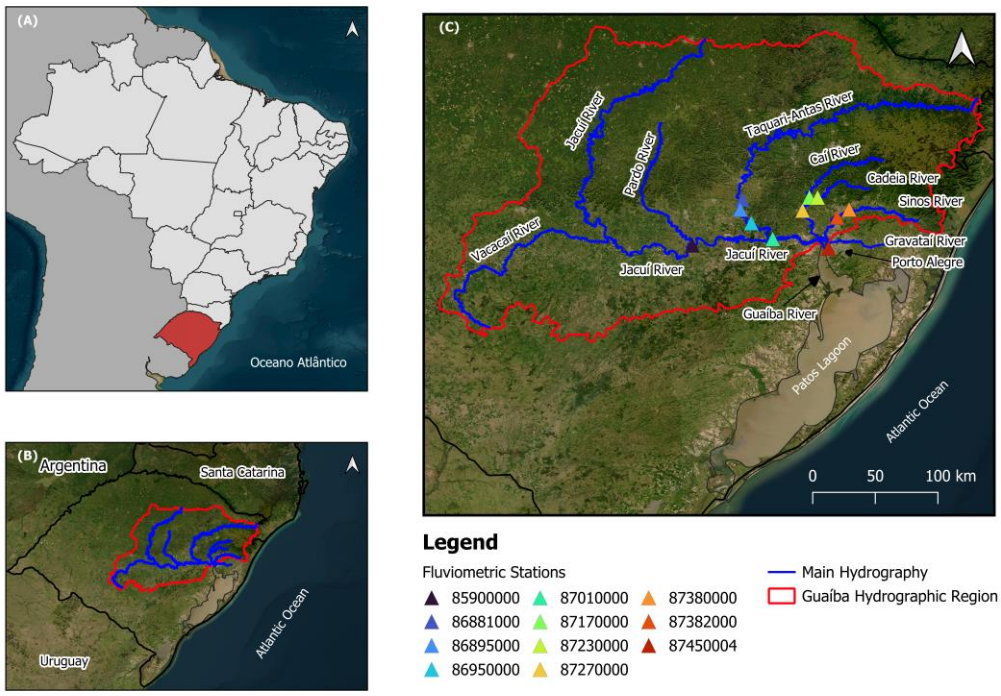
This study proposes agile Machine Learning (ML) methods for hydrological prediction, comparing them with traditional time series models (ARIMA, SARIMA). The evaluated ML models include the Bagging Regressor, Elastic Net, LASSO, and others. Training and validation were conducted using hourly data (November 11, 2023 - April 28, 2024), with testing from April 29 to May 7, 2024. Interpretability methods such as Permutation Importance and Accumulated Local Effects were applied to the top-performing models for the Guaíba Lake water level forecasts.

Few scientific studies exist on the 2024 Guaíba Lake flood (Alcântara et al., 2024). Most focus on extreme precipitation, atmospheric systems, and flood/risk mapping for mitigation. To our knowledge, no studies in the literature have explored the explicability of ML models to forecast the water level in the 2024 Guaíba Lake flood.

MATERIALS AND METHODS

The Guaíba Hydrographic Region is located in the Northeast portion of the State of Rio Grande do Sul (RS), Brazil. Sixteen islands form it and receive contributions from seven main rivers, namely Vacacaí, Jacuí, Pardo, Taquari-Antas, Caí, Sinos, and Gravataí (Figure 1). The region has an area of approximately 84,763.54 km², with 251 municipalities and an estimated population of 5.9 million people (Alcânntara et al., 2024).

Figure 1 – Location of the Guaíba Hydrographic Region, in Rio Grande do Sul, Brazil, and its main hydrography. Note that the fluviometric stations used in this work are placed as colored triangles.



Recurrent heavy rainfall has caused severe flooding in the region, with the most significant events occurring between September 2023 and May 2024. The September 2023 floods affected 107 municipalities, causing 54 deaths, while the May 2024 floods were even more severe, impacting 478 towns and 2.3 million people and resulting in 183 deaths. The Guaíba hydrographic region was the most affected, with record-breaking water levels.

To forecast Guaíba River levels, the study tested 12 machine learning models using SciKit-Learn in Python, including Bagging Regressor, ElasticNet, and XGB Regressor. The prediction accuracy, assessed using root mean squared error (RMSE), was compared with traditional univariate time series models.

In the present paper, the general AutoML framework described in (Soares et al., 2025), the so-called ML4FF, was adapted for the current prediction task. Thus, each ML method underwent a training-validation-testing phase followed by a holdout assessment.

The first phase employed nested cross-validation combined with automatic hyperparameter tuning through Bayesian optimization, aligning with standard practices to evaluate model generalization across various data splits. The second phase assesses the model's ability to generalize predictions on an unseen dataset, known as the holdout set.

The dataset is divided into two parts for each ML method: Nested Cross-Validation and Holdout. The first part (from the 1st of January of 2024 to the 28th of April of 2024 in case 1 and from the 3rd of November of 2023 to the 28th of April of 2024 in case 2) will be used in the training-validating-testing phase (nested CV loops). The second part (from the 29th of April 2024 to the 7th of May 2024 in both cases) will be used in the holdout assessment phase. The nested CV scheme considers a kouter = 30 by kinner =10-fold iteration scheme. The negative of the mean normalized Nash-Sutcliffe model efficiency coefficient (nNSE), nNSE = (2 − NSE)−1 (Nossent & Bauwens, 2012), where the mean refers to the nNSE values for each complete set of inner loops, was chosen as the loss function to optimize. By selecting a loss function based on NSE, we expect to improve the performance of hydrological forecasting for floods since this metric is sensitive to peak flows due to its quadratic formulation (Shrestha et al., 2014).

With the availability of level and precipitation at 11 fluviometric stations, we implemented fast Machine Learning hydrologic methods with a 1-hour lead time. For details of the specifications of each model, see Table S1 in the supplementary information. The forecasts of the level of Guaíba Lake (at station number 87450004 in Figure 1) at lead times of 1, 6, 12, and 24-hours were compared to the results of ARIMA (12,1,0) and SARIMA (4,1,0,12) models.

In this study, we used global black-box methods, i.e., explainers that provide a general understanding of the relationship between inputs and predictions and that can be applied to any ML model (Klaise et al., 2021). The contribution of each characteristic to the performance of the model can be evaluated using the Permutation Importance method (Breiman, 2001), (Fisher et al., 2019). Based on (Apley & Zhu, 2019), Accumulated Local Effect (ALE) is a method for computing the effects of each feature on the predictions, using the conditional distribution to average the prediction differences over other features.

The source code and data supporting this letter are publicly available in a GitHub repository, allowing other researchers to replicate the case study presented here. The code is licensed under the MIT license, and the repository can be accessed at https://github.com/jaqueline-soares/guaiba-disaster-2024.

RESULTS AND DISCUSSION

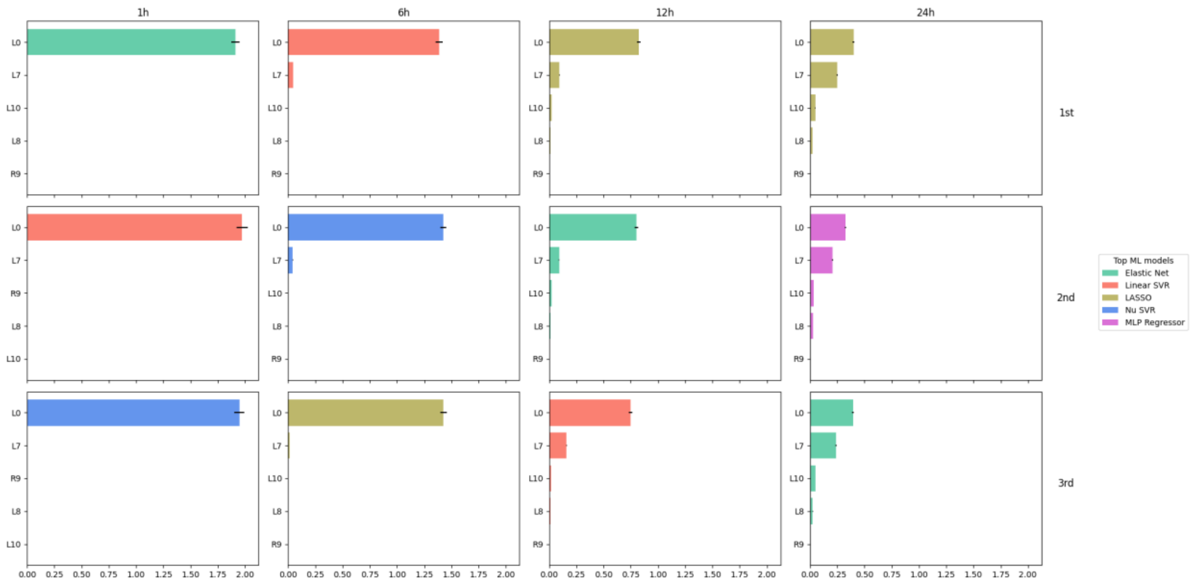
The rain that flooded Guaíba Lake began on April 25 and lasted until May 5. Over 1000 mm of rain fell over the Guaíba basin in ten days, with the headwaters of the Jacuí, Taquari, Caí, and Sinos rivers receiving the majority of the highest amounts, which led all rivers in the hydrographic basin to reach historic levels. According to the fluviometric stations data, on May 1st, the Cadeia and Caí Rivers reached the flood peak; on May 2nd, the Taquari River reached the flood peak; on May 4th, the Sinos River reached the flood peak; and on May 5th, the Jacui River and the Guaíba Lake reached the flood peak.

The SARIMA benchmark model was among the top 5 models with the lowest RMSE in all the forecasting lead times considered. Among the machine learning models, ElasticNet and LASSO were the only ones to be between the top 5 in all forecasting lead times as well. The other models ranked between the top 5 were LinearSVR (1h, 6h, and 12h forecasting), NuSVR (1h, 6h, and 24h forecasting), ARIMA (12h forecasting), and ML-Pregressor (24h forecasting). The prediction accuracy of the models compared in this study (Table S2), as well as the forecasts of the top-performing models (Figure S1), is available in the supplementary information.

To understand why this may have happened, we applied interpretability methods that provide global insights into the behavior of the ML model. Figure 2 compares the importance of the five main characteristics of R2 of the top 3 machine learning models proposed to predict the level of the Guaíba LakeRiver. It is interesting to note that although L0 (level at station number 87450004 in Figure 1) is the most crucial feature for the performance of the top models in all forecasting times considered, its importance tends to decrease as forecasting time [h] increases. However, L7 (the level at the Taquari River station), which is located 90.3 km from the outlet of the basin, gained importance as the forecast time [h] increased. Furthermore, none of the 12h accumulated precipitation at the considered stations was essential to explain the variations at the level of the Guaíba Lake.

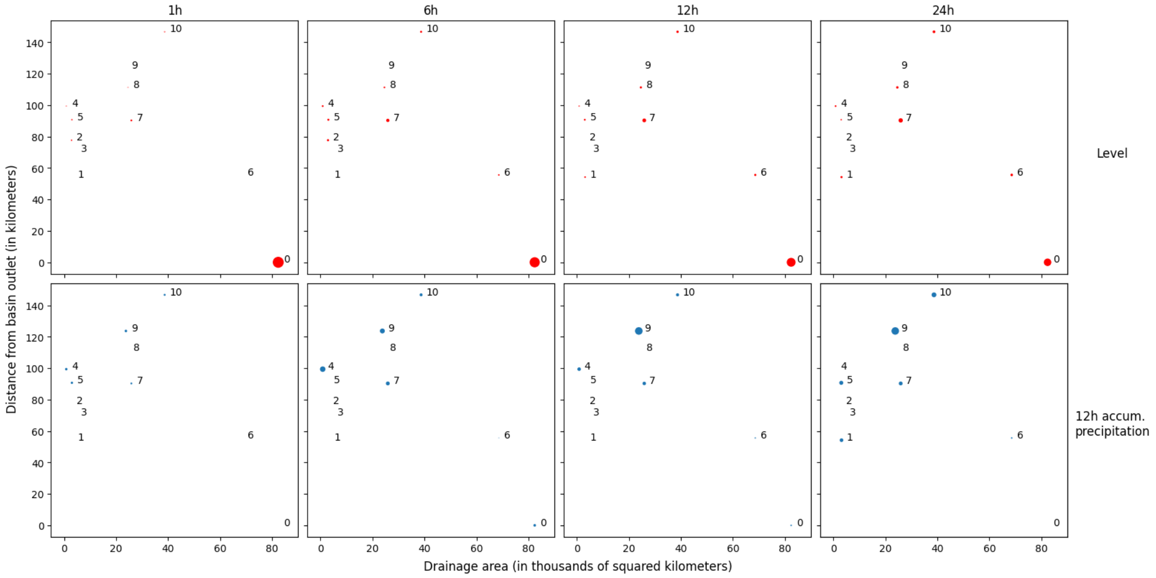
In the proposed ML models, L0 is the autoregressive term. The fact that it was the most crucial feature in the top ML models also explains the good performance of the benchmark time-series models. However, since SARIMA and ARIMA are univariate models, they could not benefit from other critical data available at different stations, as the top ML models did at higher forecasting lead times.

Figure 2 Permutation importance using R2 score. Ln represents the level at station n, and Rn represents the 12h accumulated precipitation at station n.



The sensitivity of the forecasts with respect to small changes in the value of the features can be evaluated through the Accumulated Local Effect (ALE). On average, as illustrated in Figure 3, the predictions of the top 3 machine learning models were more sensitive to L0 variations up to 6 h of forecasting time. From the 12-h forecast and later, they were also sensitive to 12-h accumulated precipitation at station number 86881000 for 12-h (see Figure 1).

Figure 3 – Top 3 machine learning models' average accumulated local effect. The circle size is proportional to the effect.



The average ALE indicates that the level in the stations with the smallest drainage area did not influence the predictions of Guaíba Lake level. However, accumulated precipitation for 12 hours further away from the basin outlet had an increasing effect on the Guaíba level forecasts.

FINAL REMARKS

This article presents experiments and analysis on the development of traditional time series (SARIMA and ARIMA) and Machine Learning (ML) methods for forecasting water levels for the Guaíba Lake watershed. SARIMA and ARIMA excelled in the short- and medium-term prediction horizons because of their autoregressive capabilities but lacked the flexibility to incorporate external spatial data. ML models (ElasticNet, LASSO, and NuSVR), while competitive in short horizons, showed the potential to leverage diverse characteristics such as precipitation and distant water levels for long-term forecasts despite challenges in converging to observed levels in extended horizons. Feature analysis highlighted the dominance of the water level at station 87450004 (L0) in short-term predictions, while features from more distant locations, such as Taquari River levels and precipitation at station 86881000, became more important over longer horizons.

Combining the temporal precision of time series models with the data integration strengths of ML models in hybrid approaches could enhance the accuracy of the prediction. Future efforts should focus on expanding monitoring networks and incorporating rainfall data from numerical models, radars, or satellites, improving predictions, and supporting better water management during extreme events.

Finally, it is essential to emphasize that during severe hydrometeorological disasters, such as the recent event in Rio Grande do Sul, the demand for rapid solutions becomes critical, and agility in modeling is essential to provide timely and accurate insights for mitigating impacts. The ML- and ST-based approaches presented in this study stand out as powerful tools in this context due to their ability to quickly adapt to dynamic flooding processes, analyze complex spatial and temporal correlations, such as relationships between upstream measurements at multiple monitoring stations and downstream river levels, and deliver reliable long-term level forecasts and flood maps that can be immediately used to support decision-making and practical actions in such rapidly evolving scenarios. The achievement of these objectives can be further enhanced by using autoML frameworks, which significantly accelerate the development, calibration, and deployment of predictive models and flood maps without compromising their accuracy.

OPEN RESEARCH SECTION

The source code and data supporting this letter are publicly available in a GitHub repository, allowing other researchers to replicate the case study presented here. The code is licensed under the MIT license, and the repository can be accessed at https://github.com/jaqueline-soares/guaiba-disaster-2024.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenacção de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by the Project 446053/2023-6 (CNPq). Finally, the authors thank the Brazilian Ministry of Science, Technology, and Innovation and the Brazilian Space Agency.

REFERENCES

ALCÂNTARA, E., MANTOVANI, J., BAIÃO, C., PAMPUCH, L., CURTARELLI, M., GUIMARÃES, Y., . . . others (2024). Unprecedented flooding in porto alegre metropolitan region (southern brazil) in may 2024: Causes, risks, and impacts. doi: https://dx.doi.org/10.2139/ssrn.4867780

APLEY, D. W., & ZHU, J. (2019). Visualizing the effects of predictor variables in black box supervised learning models. Retrieved from https://arxiv.org/abs/1612.08468

BIAN, L., QIN, X., ZHANG, C., GUO, P., & WU, H. (2023). Application, interpretability and prediction of machine learning method combined with lstm and lightgbm-a case study for runoff simulation in an arid area. Journal of Hydrology, 625 ,130091. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022169423010338

doi: https://doi.org/10.1016/j.jhydrol.2023.130091

BREIMAN, L. (2001). Random forests. Machine Learning, 45 , 5–32. doi: 10.1023/A: 1010933404324

CAPPELLI, F., & GRIMALDI, S. (2023, December). Feature importance measures for hydrological applications: insights from a virtual experiment. Stochastic Environmental Research and Risk Assessment , 37 (12), 4921–4939.

DE GOOIJER, J. G., & HYNDMAN, R. J. (2006). 25 years of time series forecasting. International Journal of Forecasting, 22 (3), 443-473. (Twenty five years of forecasting) doi: https://doi.org/10.1016/j.ijforecast.2006.01.001

ELDEEB, H., MAHER, M., ELSHAWI, R., & SAKR, S. (2024, June). Automlbench: A comprehensive experimental evaluation of automated machine learning frameworks. Expert Systems with Applications, 243 , 122877. doi: 10.1016/j.eswa.2023.122877

FISHER, A., RUDIN, C., & DOMINICI, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Retrieved from https://arxiv.org/abs/1801.01489

KLAISE, J., 259 VAN LOOVEREN, A., VACANTI, G., & COCA, A. (2021, January). Alibi explain: algorithms for explaining machine learning models. J. Mach. Learn.Res., 22 (1).

NOSSENT, J., & BAUWENS, W. (2012). Application of a normalized Nash-Sutcliffe efficiency to improve the accuracy of the Sobol'sensitivity analysis of a hydrological model. In Egu general assembly conference abstracts (p. 237).

SAHU, M. K., SHWETHA, H. R., & DWARAKISH, G. S. (2023, September). State-of-the-art hydrological models and application of the HEC-HMS model: a review. Modeling Earth Systems and Environment , 9 (3), 3029–3051. doi:10.1007/s40808-023-01704-7

SHRESTHA, R. R., PETERS, D. L., & SCHNORBUS, M. A. (2014). Evaluating the ability of a hydrologic model to replicate hydro-ecologically relevant indicators. Hydrological Processes, 28 (14), 4294-4310. doi: https://doi.org/10.1002/hyp.9997

SOARES, J. A. J. P., OZELIM, L. C. S. M., BACELAR, L., RIBEIRO, D. B., STEPHANY, S., & SANTOS, L. B. L. (2025). ML4FF: A machine-learning framework for flash flood forecasting applied to a Brazilian watershed. Journal of Hydrology - accepted for publication

STEIN, L., CLARK, M. P., KNOBEN, W. J. M., PIANOSI, F., & WOODS, R. A. (2021). How do climate and catchment attributes influence flood generating processes? a large-sample study for 671 catchments across the contiguous usa. Water Resources Research, 57 (4), e2020WR028300. doi: https://doi.org/10.1029/2020WR028300