Open Access
PIAHS

Artificial neural networks applied for flood forecasting in ungauged basin – the Paranaíba river study case

Abderraman R. A. Brandão[1], Frederico C. M. de Menezes Filho[2], Paulo T. S. Oliveira[1,3], and Maria C. Fava[4]

[1]Department of Hydraulics and Sanitation, University of São Paulo, São Carlos, SP, Brazil
[2]Institute of Exact and Technological Sciences, Federal University of Viçosa, Rio Paranaíba, MG, Brazil
[3]Faculty of Engineering, Architecture and Urbanism, and Geography,
Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil
[4]Civil Engineering Department, Federal University of São Carlos, São Carlos, SP, Brazil

Correspondence: Abderraman R. A. Brandão (abderraman.brandao@usp.br)

Abstract. Flow simulation using artificial neural networks (ANNs) in the modelling has been widely applied and has gained prominence in regions lacking data. The hydrological variables are subject to the influence of morphological characteristics and urbanization in the watershed. Statistical models, such as ANNs, need to be able to identify the relationship between the hydrological inputs and outputs of the model, without explicitly considering the other relationships involved in physical processes. This work aimed to apply a Multilayer Perceptron (MLP) neural network for predicting flows in an urban basin subject to recurrent floods, using precipitation and flow data from previous periods as inputs. After model calibration and validation for the current state of the basin (2018–2019), its responses were analysed using input data before the basin urbanization (1985–1986) to identify the error behaviour at the output as a proxy for the basin changes effect. Its efficiency was evaluated using hydrographs, showing satisfactory results in both periods. In the urbanization period, there is more dispersion for maximum flows. For the day 4 steps back in the current forecast, NSE $= 0.59$ was observed, whereas in the other period, NSE $= 0.70$. The evaluation of the models for the current period of basin urbanization showed that the model could capture the basin's physical dynamics within the established static relationship. Also, the result found in the statistical relationships for the inputs showed once again the impact of urbanization on the basin.

1 Introduction

Hydrological modelling has proven to be a reliable solution for many problems. However, in traditional models, the representation of hydrological processes translates into high complexity, due to the large number of input parameters required for model setup. In parallel with the development of traditional models that aim to represent the physical processes involved in the hydrological cycle, since the 1950s, scholars in the area have been developing data-driven models for the description of complex hydrological processes without the need for several input parameters, given the difficulty in obtaining these data (Fatichi et al., 2016).

According to Kermani et al. (2020), given the computational efficiency and flexibility of applying machine learning, this approach has been applied to solve several challenges in the field of hydrological sciences and has great potential to provide more accurate and reliable predictions when compared to traditional statistical models, stochastic methods, and empirical formulations.

Artificial Neural Networks (ANNs) are part of the set of machine learning techniques based on artificial neurons. Currently, the ANN Multilayer Perceptron (MLP) architecture is widely used. Its array is divided into input neurons, which store the input vectors, and output vectors, which receive the processing response. There is a hidden layer between the input and output vectors, which may contain one or more divi-

sions where data processing takes place, with the entire network being connected (Dazzi et al., 2021).

Considering the multilayer perceptron application perspectives, the objective of this work was to evaluate the performance of the artificial neural network to simulate average daily flows in a basin in Minas Gerais State in Brazil with input data under the effect of changes caused by the urbanization of the basin, evaluating its effectiveness according to statistical criteria. The case study was chosen due to the recurrence of Paranaíba river overflow in the rainy season, causing damage to the municipality of Patos de Minas and its population. Another factor to be highlighted is the evidence of the importance of the river for the entire Alto Paranaíba region since it supplies cities and local agricultural activities. The forecast of flows in the basin is paramount for productive environmental zoning policies, the safety of the population living in the neighbourhoods surrounding the floodplain area. According to Nogueira (2017), the area is highly susceptible to floods, while Amaral (2021) describes that floods periodically reach the urban perimeter.

2 Case Study and Problem Statement

The study was carried out in a sub-basin of the Paranaíba River, one of the main tributaries of the Paraná River basin, with the outlet defined at the fluviometric station (Code 6001100) located in the urban perimeter of Patos de Minas – MG. The drainage area of the basin is 3791.83 km^2 (Fig. 1), corresponding to 11.02 % of the entire basin of the Paranaíba River, with an extension of the main watercourse of 106.6 km.

The predominant land use and occupation class is agriculture and livestock, but there is also an urbanized area. The region has a low slope in the central region and is mountainous at the extremes and the predominant soil class is Red Latosol, according to Santos et al. (2018). The time of concentration is about 12.93 h, calculated by the U.S. Army Corps of Engineers equation (Collischown and Dornelles, 2021).

According to the Köppen-Geiger climate classification, the type of climate present in the basin is mostly of the Aw type, evidencing a tropical climate with temperatures above 18 °C, a dry season, and high annual precipitation.

3 Material and Methods

3.1 Data description and preparation of inputs

The precipitation series were corrected using the Simple Linear Regression method, data consistency was performed using the Double Mass method developed by the Geological Survey (USA), and the average rainfall in the basin was calculated using Thiessen Polygons, all as described by Collischonn and Dornelles (2021).

The basin underwent intense urbanization from the 2000s onwards (Bessa and Soares, 2002). Thus, the rainfall and
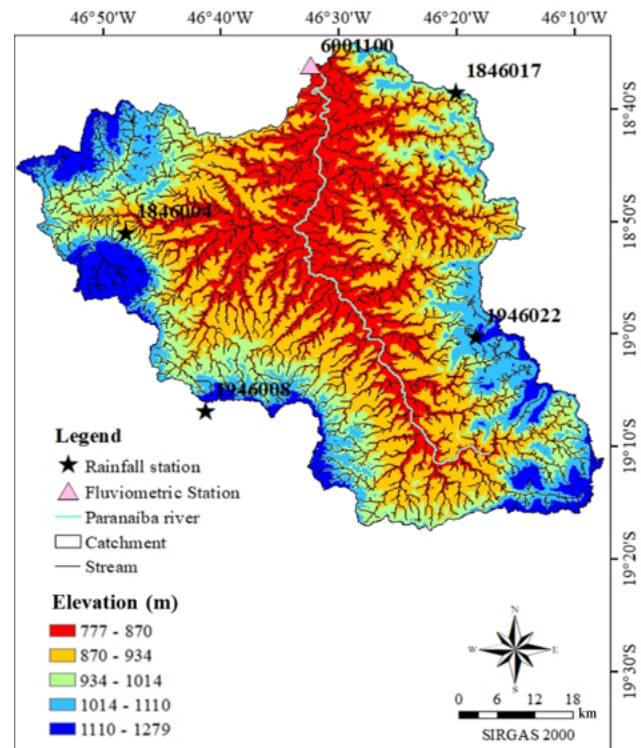

Figure 1. Location of Paranaíba River sub-catchment.

flow series for training the model before urbanization were from 1976 to 1984 (3287 daily data), and the series from 1985 to 1986 were used as a set for the test (729 daily data). For the post-urbanization, the training data selected were from 1 January 2008 to 31 December 2016 (3287), and data from 2018 to 2019 were used as a set for the test (729 daily data).

The values of past rainfall and flow to be included in the model took into account the physical characteristics of the basin to find the reaction time. We used correlation analysis to examine the interdependence between rainfall and flow for both periods. Meanwhile, autocorrelation helped identify past flow values influencing the current flow, as depicted in Fig. 3.

From the correlation graph, it can be seen that for the current period of the basin the degree of correlation between rainfall and flow has dropped in relation to the other period, requiring a smaller number of rain delays when compared to the period prior to urbanization, this being still one of the indications that the anthropological changes in the basin influenced its time of concentration. The same happens when comparing the influence of previous flows with current flows using autocorrelation with a confidence level of 99 %; there is a greater need for past flows for the current period than for the previous one, in addition to in both cases demonstrating the strong correlation of bottom recharge to the river.
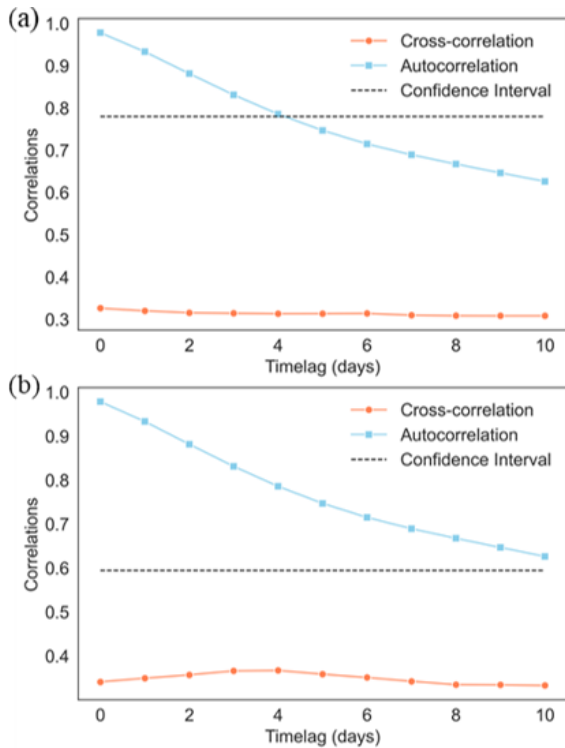
Figure 2. Correlation analysis to period **(a)** pre-urbanization and **(b)** post-urbanization.

The positive and insignificant correlations between rainfall and flow for the current (0.32) and previous period (0.36) suggest the influence of other variables on the observed flows, something already expected, under the concepts of water balance. In addition, the type of soil and its occupation, combined with the low slopes in the central area favor the recharge of the groundwater and groundwater flows; both directly influence the flow (Mendonça et al., 2021).

3.2   Model Structure definition

The training of ANNs can be supervised or unsupervised (Raschka, 2015). According to Cristaldo (2020), supervised training is the most used in hydrological forecasting problems. After training, it is expected that the network can generalize data and finds an answer in vectors different from the training. The propagation capacity of the network is related to its topology.

This study trained and validated ANNs using the Waikato Environment for Knowledge Analysis (WEKA) software, which uses a General Public License (GPL) written in Java, and has a collection of machine learning algorithms, including the MLP architecture. The software uses neurons in the hidden layer equal to half the sum of arrays and classes with the sigmoid function.

Using as neurons of the hidden layer equal to half the sum of attributes and classes, several epochs equal to 1000, a mo-
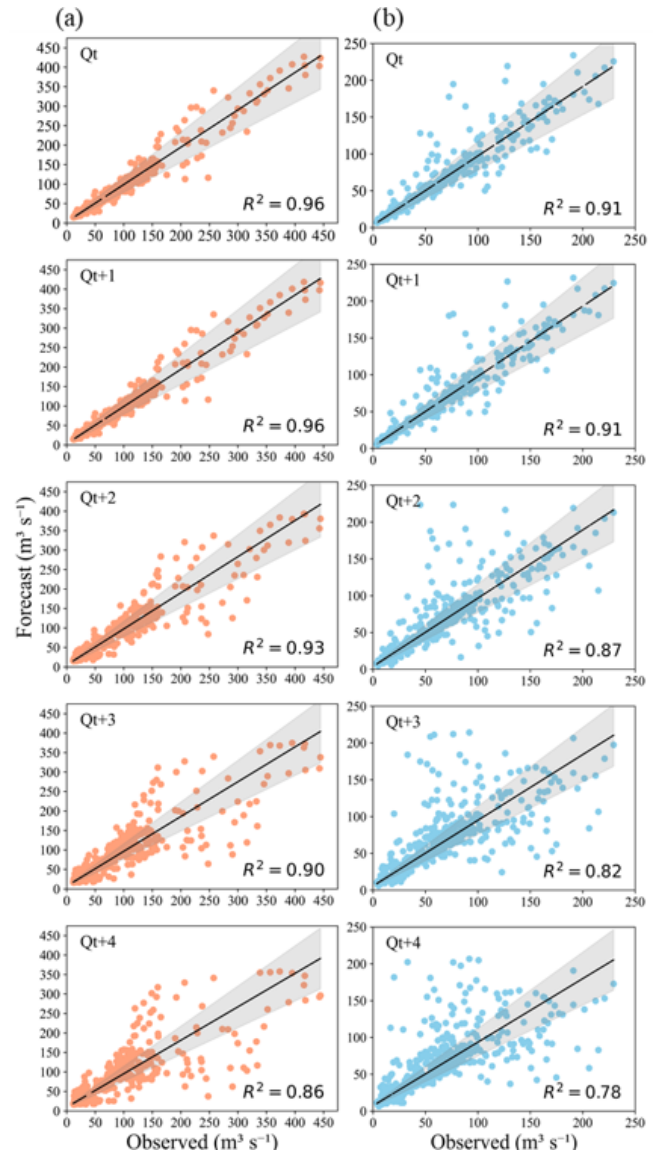

Figure 3. Observed vs. forecasted flow in Paranaíba River for **(a)** 1984 and **(b)** 2019. Comparisons include current time and 1–4 d lags. Shaded area: $\pm 20\%$ error.

mentum of 0.2, and a learning rate of 0.1. The software library's perceptron model works with backpropagation, providing supervised training. In order to avoid the underfitting and overfitting phenomena, cross-validation k-fold type was applied ($k = 4$). A summary of the trained ANNs is presented in Table 1.

3.3   Metrics for Model Evaluation

The performance of the model evaluation was assessed using the determination coefficient (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Correlation Coefficient (CC) and the Nash-Sutcliffe Coefficient of Efficiency

Table 1. ANN experiment summary.

Out variable	1984 Input variables	Hidden nodes	2019 Input variables	Hidden nodes
Q_t	$Re_t, Re_{t-1}, Re_{t-2}, Re_{t-3}, Q_{t-1}$	3	Re_t, Re_{t-1}, Q_{t-1}	2
Q_{t+1}	$Re_t, Re_{t-1}, Re_{t-2}, Q_t, Q_{t-1}$	3	$Re_t, Re_{t-1}, Q_t, Q_{t-1}$	2
Q_{t+2}	$Re_t, Re_{t-1}, Re_{t-2}, Q_t, Q_{t-1}$	2	$Re_t, Re_{t-1}, Q_t, Q_{t-1}$	2
Q_{t+3}	$Re_t, Re_{t-1}, Re_{t-2}, Q_t, Q_{t-1}$	2	$Re_t, Re_{t-1}, Q_t, Q_{t-1}$	2
Q_{t+4}	$Re_t, Re_{t-1}, Re_{t-2}, Q_t, Q_{t-1}$	2	$Re_t, Re_{t-1}, Q_t, Q_{t-1}$	2

Re_{t-i} Effective rainfall (mm), $i =$ lag time (0, 1, 2,..., day). Q_{t-k} Discharge at time (m^3 s^{-1}), $t - k$, $k = 0, \ldots, 4$.

(NSE). These are some of the criteria adopted in the studies of Mendonça et al. (2021), Dazzi, Vancondio and Mignosa (2021). RMSE and MAE criteria are better as closer to 0. The CC can range from -1 to 1, with values closer to the extremes indicating stronger correlations, and NSE should be closer to 1 for the best fit (Dazzi et al., 2021).

4 Results and discussion

At the outputs of the two networks, the evaluation parameters are similar, showing the decay of forecasting efficiency over the horizon. The best values were found in the pre-urbanization period (Table 2). As expected, the COE and NSE coefficients indicate an almost perfect fit as they approach 1 for the first days. The NSE still serves as an indicator of credibility for the model. According to the criteria established by Moriasi et al. (2015), the NSE values obtained for the two periods range from "very good" for the initial days to "Satisfactory" for the last day.

Despite the existence of discordant points in the dispersion results (Fig. 3), Pearson's coefficient was greater than 0.86 and 0.78 for the first and second periods respectively in the less accurate result and as expected were found for the most distant forecast day, while the other performance criteria CC, RMSE, MAE and NSE also resulted in plausible answers.

ANNs significantly influenced by the quantity and quality of the datasets that are trained. In Fig. 4, the data predominantly concentrates on lower flow values, and the model appropriately replicates these. However, the higher variability observed for intermediate values can be attributed to a lack of information within this range, leading the model to struggle in learning and accurately representing these values. Solomatine et al. (2003) support this observation, noting similar challenges in their studies. Moreover, hydrological variables, like flow, are substantially affected by anthropogenic activities (Wu et al., 2015). The replacement of native forests with agricultural practices, for instance, can intensify flood events (Housspanossian et al., 2023). In the current period, specifically in 2014/2015, a severe drought event occurred (Marengo et al., 2015). Hydrologic signatures also affect the
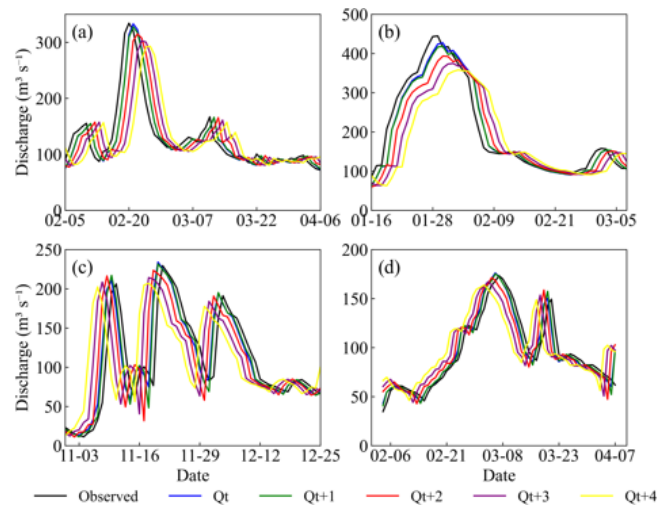

Figure 4. Flow hydrographs estimated by the MLP-ANN. **(a, b)** pre-urbanization period (1986–1985) and **(c, d)** After urbanization (2019–2019).
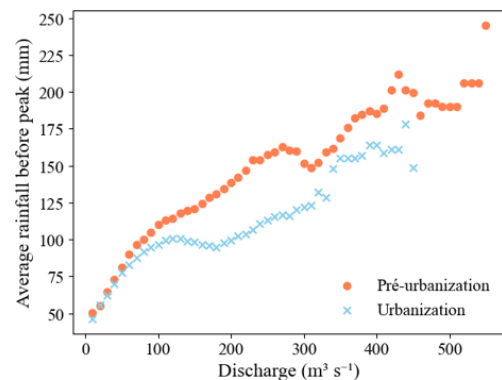

Figure 5. Average rainfall before peak of discharge (10 d) for both periods.

models (McMillan et al., 2023), and other natural mechanisms influence floods (Sharma, 2018).

These results, consequently, were also reflected in the hydrographs, where deviations between the simulated and ob-

Table 2. Comparison of errors for the verification dataset.

Prediction	Test – 1984				Test – 2019			
	CC	MAE	RMSE	NSE	CC	MAE	RMSE	NSE
		$(\mathrm{m^3\,s^{-1}})$				$(\mathrm{m^3\,s^{-1}})$		
Q_t	0.97	6.89	14.81	0.95	0.95	6.04	13.28	0.90
Q_{t+1}	0.97	6.77	14.72	0.95	0.95	5.26	12.23	0.91
Q_{t+2}	0.93	11.55	24.05	0.88	0.88	9.38	20.89	0.76
Q_{t+3}	0.88	15.76	31.81	0.79	0.81	12.53	25.57	0.64
Q_{t+4}	0.83	18.66	37.92	0.70	0.78	14.23	27.49	0.59

served curves are almost not perceptible (Fig. 4), demonstrating the coherence of the model and the respective numerical capture of the basin bias.

Analyzing past flow peaks (Fig. 5) revealed that a higher average of previous precipitation was necessary to trigger the flow peak compared to current conditions. This discrepancy in peak flow time is partly due to urbanization process. Soil waterproofing reduces rainwater infiltration and accelerates surface runoff. Thus, the water reaches the outlet faster in urbanized areas, bringing forward the peak flow. The results indicate that urbanization has sped up the flood arrival time, as already demonstrated in other studies (e.g., Lu et al., 2023; Nardi et al., 2018).

5 Conclusions

The study effectively simulated average daily flows using ANN, producing results aligned with existing literature. Statistical relationships in input data were crucial for selecting influential input variables and eliminating less impactful ones. The research also points out that anthropic change in the basin, altering its natural dynamics and showing better results in the pre-urbanization period. Moreover, it can aid decision-making in extreme events due to its versatile methodology, offering simpler replication than traditional models with their intensive data and parameter needs.

Author contributions. ARAB conceived and presented idea, wrote the paper in consultation with MCF, FCMdMF, PTSO. All authors provided critical feedback and helped shape the research, analysis and manuscript.

References

Agência Nacional de Águas e Saneamento Básico (ANA): HIDROWEB v3.3.7413.0, ANA [data set], https://www.snirh.gov.br/hidroweb/serieshistoricas, last access: 10 February 2023.

Amaral, E. A., Nascimento, A. R. T., Silva, C. R., Oliveira, A. P., and Silva, G. R.: Avaliação de impactos ambientais na APP do Rio Paranaíba e inferências para mitigação, Revista Ibero Americana de Ciências Ambientais, 12, 572–584, ISSN 2179-6858, 2021.

Collischonn, W. and Dornelles, F.: Hidrologia para engenharia e ciências ambientais, Porto Alegre: Associação Brasileira de Recursos Hídricos, ISBN 978-85-8868-634-2, 2021.

Bessa, K. C. F. O. and Soares, B. R.: Considerações Sobre A Dinâmica Demográfica Na Região Do Triângulo Mineiro/Alto Paranaíba, Caminhos de Geografia, Uberlândia, MG, 3, 22–45, https://doi.org/10.14393/RCG3615293, 2002.

Cristaldo, M. F., Jesus, L., Oliveira, P. T., Souza, C. C., Viganó, H. H. G, and Padovi, C. R.: Redes Neurais Artificiais Aplicadas À Previsão De Enchentes Para Região Do Pantanal No Mato Grosso Do Sul, Geociências, 39, 191–201, https://doi.org/10.5016/geociencias.v39i1.13644, 2020.

Dazzi, S., Vacondio, R., and Mignosa, P.: Flood Stage Forecasting Using Machine-Learning Methods: A Case Study on the Parma River (Italy), Revista Water, 13, p. 1612, https://doi.org/10.3390/w13121612, 2021.

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., Downer, C. W., Camporese, M., Davison, J. H., Ebel, B., Jones, N., Kim, J., Mascaro, G., Niswonger, R., Restrepo, P., Rigon, R., Shen, C., Sulis, M., and Tarboton, D.: An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, J. Hydrol., 537, 45–60, https://doi.org/10.1016/j.jhydrol.2016.03.026, 2016.

Houspanossian, J., Giménez, R., Whitworth-Hulse, J. I., Nosetto, M. D., Tych, W., Atkinson, P. M., Rufino, M. C., and Jobbágy, E. G.: Agricultural Expansion Raises Groundwater and Increases Flooding in the South American Plains, Science, 380, 1344–1348, https://doi.org/10.1126/science.add5462, 2023.

Kermani, M. Z, Matta, E., Cominola, A., Xia, X., Zhang, Q., Liang, Q., and Hinkelman, R.: Neurocomputing in surface water hydrology and hydraulics: A review of two decades retrospective, current status and future prospects, J. Hydrol., 588, 125085, https://doi.org/10.1016/j.jhydrol.2020.125085, 2020.

Marengo, J. A., Nobre, C. A., Seluchi, M. E., Cuartas, A., Alves, L. M., Mendiondo, E. M., Obregón, G., and Sampaio, G.: A Seca e a Crise Hídrica de 2014–2015 em São Paulo, Rev. USP, 106, 31–44, https://doi.org/10.11606/issn.2316-9036.v0i106p31-44, 2015.

McMillan, H., Coxon, G., Araki, R., Salwey, S., Kelleher, C., Zheng, Y., Knoben, W., Gnann, S., Seibert, J., and Bolotin, L.: When good signatures go bad: Applying hydrologic signatures in large sample studies, Hydrol. Process., 37, e14987, https://doi.org/10.1002/hyp.14987, 2023.

Mendonça, L. M., Gomide, I. S., Sousa, J. V., and Blanco, C. J. C.: Modelagem chuva-vazão via redes neurais artificiais para simulação de vazões de uma bacia hidrográfica da Amazônia, Revista de Gestão de Água da América Latina, Porto Alegre, 18, 2021, https://doi.org/10.21168/rega.v18e2, 2021.

Moriasi, D. N., Gitau, M. W., Pai, N., and Daggupati, P.: Hydrologic and water quality models: Performance measures and evaluation criteria, T. ASABE, 58, 1763–1785, https://doi.org/10.13031/trans.58.10715, 2015.

Lu, M., Yu, Z., Hua, J., Kang, C., and Lin, Z.: Spatial dependence of floods shaped by extreme rainfall under the influence of urbanization, Sci. Total Environ., 857, 159134, https://doi.org/10.1016/j.scitotenv.2022.159134, 2023.

Nogueira, T. P. N.: Mapeamento Da Suscetibilidade À Inundação Na Bacia Hidrográfica Do Ribeirão Da Fábrica, Município De Patos De Minas – MG. 123 f., Master's Thesis, Environmental and Environmental Quality, Federal University of Uberlândia, Uberlândia, https://doi.org/10.14393/ufu.di.2017.304, 2017.

Nardi, F., Annis, A., and Biscarini, C.: On the impact of urbanization on flood hydrology of small ungauged basins: The case study of the Tiber river tributary network within the city of Rome, J. Flood Risk Manag., 11, S594–S603, https://doi.org/10.1111/jfr3.12186, 2018.

Raschka, S. and Mirjalili, V.: Python Machine Learning, 2 edn., Packt, Birmingham, ISBN 9781787125933, 2015.

Santos, H. G. dos, Jacomine, P. K. T., Anjos, L. H. C. dos, Oliveira, V. A. de, Lumbreras, J. F., Coelho, M. R., Almeida, J. A. de, Araujo Filho, J. C. de, Oliveira, J. B. de, and Cunha, T. J. F.: Brazilian System of Soil Classification, 5th edn. rev. and expanded, Brasília, DF, Embrapa, http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1094003 (last access: 15 February 2023), 2018.

Sharma, A., Wasko, C., and Lettenmaier, D. P.: If precipitation extremes are increasing, why aren't floods?, Water Resour. Res., 54, 8545–8551, 2018.

Solomatine, D. P. and Dulal, K. N.: Model trees as an alternative to neural networks in rainfall-runoff modelling, Hydrolog. Sci. J., 48, 399–411, https://doi.org/10.1623/hysj.48.3.399.45291, 2003.

Wu, J., Yin, J., Hao, Y., Liu, Y., Fan, Y., Huo, X., Liu, Y., and Yeh, T. C. J.: The Role of Anthropogenic Activities in Karst Spring Discharge Volatility, Hydrol. Process. 29, 2855–2866, https://doi.org/10.1002/hyp.10407, 2015.