How to cite: Bîlbîe F., Zaharia L. (2024) Flash Flood Forecasting Using Machine Learning Models: A Scientometric Analysis. *2024 "Air and Water – Components of the Environment" Conference Proceedings*, Cluj-Napoca, Romania, p. 1-10, DOI: 10.24193/AWC2024_01.

FLASH FLOOD FORECASTING USING MACHINE LEARNING MODELS: A SCIENTOMETRIC ANALYSIS

Florin BÎLBÎE **[1,2]**, ***Liliana ZAHARIA*** **[1]**

DOI: 10.24193/AWC2024_01

ABSTRACT. Flash Flood Forecasting Using Machine Learning Models: A Scientometric Analysis. Hydro-meteorological hazards are a major issue in many regions of the world, including Romania. Among these, flash floods are the most frequent phenomena, generating significant annual socio-economic and environmental damages. In recent years, flash flood forecasting using machine learning algorithms has become an useful tool for data-based hydrologic modeling. Machine learning allows to create mathematical relationships between the river discharge and other climatic and physico-geographic parameters from the training dataset. This paper aims to perform a scientometric analysis using open-source programs, namely ScientoPyGui and VOSviewer. The expression 'flash flood forecasting AND machine learning' was searched in the Web of Science and Scopus databases. After merging and removing duplicates, 112 publications were retained for analysis. Their number has increased by 60% in the past three years (after 2021) with a trend towards a sub-branch of machine learning, namely deep learning. The spatial distribution of the papers showed that China is a global leader with 25% of the total. These findings highlight the increasing role of machine learning based models (particularly deep learning) in enhancing flash flood forecasting, a nonstructural measure for the flash flood risk mitigation.

Keywords: flash flood forecasting, scientometric analysis, machine learning, deep learning

1. INTRODUCTION

Hydrological disasters, particularly flash floods, have emerged as a major global concern, causing extensive damage and loss of life. According to report by Centre for Research on the Epidemiology of Disasters (CRED), hydrological disasters had the highest occurrence rate (50%) among all natural disasters between 2006-2015 (Guha-Sapir *et al.*, 2016). By 2050, damages from hydro-meteorological hazards are expected to reach one trillion dollars annually (Hartnett and Nash, 2017; Bubeck and Thieken, 2018; Tien Bui *et al.*, 2019). These alarming statistics underscore the urgent need for effective forecasting tools to mitigate the devasting impact of flash floods.

[1] Faculty of Geography, University of Bucharest, 1 Nicolae Bălcescu Blvd., 010041, Bucharest, Romania e-mail: florin.bilbie@s.unibuc.ro, liliana.zaharia@geo.unibuc.ro
[2] National Institute of Hydrology and Water Management, 97E București-Ploiești Road, 013686, Bucharest, Romania

The term ‘flash flood’ is often used interchangeably, but its precise definition remains a matter of debate within the scientific community (Gruntfest and Handmer, 2001a, 2001b; Kaiser *et al*., 2020). While torrential rains are a primary trigger, other factors such as rapid appearance of a large volume of water, or cyclones can also contribute to flooding. In general, flash floods are characterized by rapid increases in river discharge over a short period of time (2-6 hours) and typically occur in watersheds smaller than 250 km^2 (Stănescu and Drobot, 2002; WMO, 2011).

Hydrological forecasting models play a crucial role in mitigating the impacts of flash floods by providing timely warnings to vulnerable communities. These models can be broadly classified into three categories: ‘black-box’, conceptual and distributed models (Kan *et al.*, 2019). Machine learning models are part of the ‘black box’ models which are also called ‘data-driven’ models. They are based only on historical data and mathematical relationships that they develop themselves (Dazzi *et al.*, 2021). Machine learning allows to create mathematical relationships between the river discharge and other climatic and physico-geographic parameters from the training dataset. Subsequently, based on precipitation forecasts and mathematical relationships, machine learning algorithms generate river discharge forecasts.

This study aims to conduct a scientometric analysis of scientific publications that utilize machine learning models for flash flood forecasting. Scientometric analysis mining data from literature to gain insights into the development scientific research in a particular field (Mingers and Leydesdorff, 2015; Li *et al.*, 2021). Combining data mining with visualization to understand the high-level structure of a research field and collaboration networks of authors provides satisfactory results (Cobo *et al.*, 2011; Li *et al.*, 2021).

In recent years, researches based on scientometric analysis and visualizations/illustrations in different forms of the results have become frequent in scientific publications. In the field of flood risk, a scientometric comprehensive study of methods used flood risk analyse conducted by Diaconu et al. (2021) showed the increasing number of papers using methodologies based on machine learning. Such methodologies were used in different areas/regions of the world (e.g. the papers published by Alipour. *et al*., 2020; Dtissibe *et al*., 2020; Costache *et al.*, 2021, 2022; Towfiqul Islam *et al*, 2021; Li *et al.*, 2023). Our study focuses on the scientometric analysis of the publications using machine learning methods for flash flood forecasting. It completes the current information on this topic, providing original results of the analysis on a global scale.

2. DATA AND METHODS

This paper is based on the data resulted from the search for scientific publications in the Web of Science and Scopus databases, using the expression ‘flash flood forecasting AND machine learning’. On December 1st, 2023, 68 papers were identified in the Web of Science Core Collection and 88 in Scopus database. The first identified papers date from 2007. The results were exported as table. The ScientoPyGui program, was then used to merge the two tables and remove duplicate articles (Ruiz-Rosero *et al*., 2019). This resulted in a total of 112 papers. The dataset was subsequently examined and processed using the open-source programs VOSviewer 1.6.20 (https://www.vosviewer.com/download) and ScientoPyGui 2.1.3 (https://github.com/jpruiz84/ScientoPy/releases). For mapping we used QGIS 3.22.6 (https://qgis.org/en/site/), and Excel 2016 from Microsoft Office suite for creating graphs.

The VOS mapping technique is based on a similarity matrix which can be derived from a co-occurrence matrix by normalizing it, to compensate the variations in the overall number of occurrences or co-occurrences of entities. Using the association strength, the similarity s_{ij} between two items i and j is calculated as (1) where c_{ij} denotes the number of co-occurrences of items i and j or the total number of co-occurrences of these items (Van Eck and Waltman, 2010).

$$s_{ij} = \frac{c_{ij}}{w_i w_j} \tag{1}$$

3. RESULTS

The analysis of the identified papers published between 2007 and 1st December 2023 showed that the use of machine learning for flash flood forecasting has grown rapidly in recent years, with the last 3 years accounting for 60% of the total publications (Fig. 1). The processing power of computers has increased significantly due to technological advances. This allows for the running of very complex machine learning models (Kan *et al.*, 2019). Recent studies have demonstrated higher accuracy of machine learning models for river discharge (Alipour *et al.*, 2020; Dtissibe *et al.*, 2020; Hill and Schumacher, 2021; Nearing *et al.*, 2021), as well as for flash flood susceptibility (Costache *et al.*, 2021, 2022; Towfiqul Islam *et al.*, 2021).

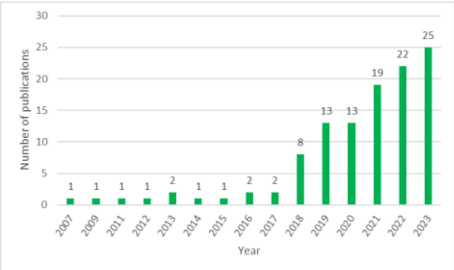

Fig. 1.* *The number of publications on the analyzed topic (source: output ScientoPyGui 2.1.3, Graph Microsoft Excel 2016)

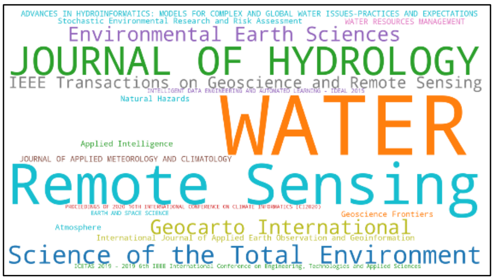

Fig. 2. Word cloud graph of identified journals (source: output ScientoPyGui 2.1.3)

A comprehensive analysis of the published literature in the considered databases revealed that the articles constitute the majority of publications (78%), followed by proceeding papers (10%), conference papers (8%) and reviews (4%). The top five journals preferred by authors are *Water* (9%), *Journal of Hydrology* (5%), *Remote Sensing* (5%), *Science of the Total Environment* (3%), and *Environmental Earth*

Science (3%) (Fig. 2). These journals are well-established and recognized for their high scientific quality and impact in the field of hydrology and water resources.

The most prolific authors who have published 3 or more articles are grouped into 7 clusters (Fig. 3). The two purple clusters represent publications published before 2016 and have authors from France (Johannet A., Dreyfus G.) and Brazil (Ueyama J., Furguim G., Pessin G.). Another two clusters, in yellow, show publications after 2022 and have authors from China (Liu C.J., Ma Q, Zanchetta A., Coulibaly P).

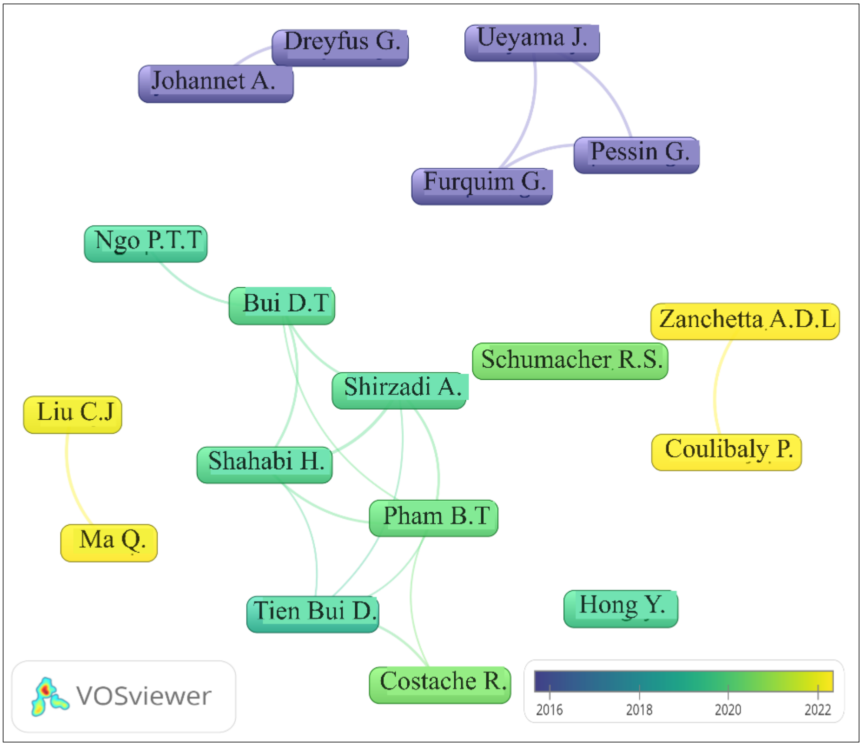

Fig. 3. Co-authorship for authors using full counting network map (source: output VOSviewer 1.6.20)

The ScientoPyGui program extracted countries from the authors affiliations. For a particular geographical region, multiple authors contributed to the paper's conception. Some of these authors were from countries outside of the analyzed region. There are 41 countries with publications on flash flood forecasting using machine learning. The spatial distribution by country (Fig. 4) shows that the top three countries with the most publications are China (28), USA (22), and India (16). All the countries in the European continent total 42 articles. The top five countries of Europe are United Kingdom (7), Norway (6), France (5), Romania (4) and Italy (3). At the global level, Romania ranks tenth.

The institutions the authors belong to were extracted from their affiliations. As shown in Figure 5 most institutions are from China. In this country there is a strong interest in this field in universities, such as North China University Water Resources and Electric Power, Beijing Normal University, National Cheng Kung University. There is also interest in other Chinese institutions, such as China Institute Water Resources and Hydropower Resources. In US, the main institutions with publication on the analyzed topic are Oklahoma University and Colorado State University.

The VOSviewer program was used to extract keywords from all papers. Keywords that appeared more than 5 times were then selected. The results illustrated in Figure 6 show a transition in research topics from rain-runoff model optimization (purple boxes) to machine learning (light green boxes), followed by a trend towards deep learning, a subfield of machine learning (yellow boxes).

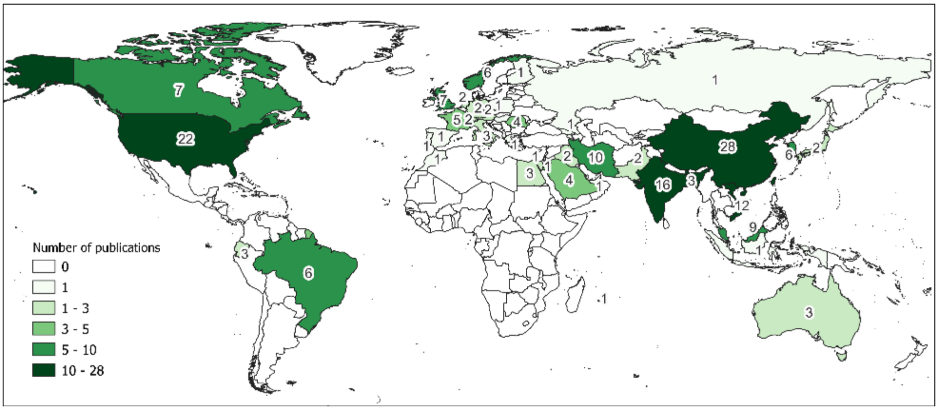

Fig. 4. Spatial distribution of identified publications by country (source: output ScientoPyGui 2.1.3, maps QGIS 3.22.6)

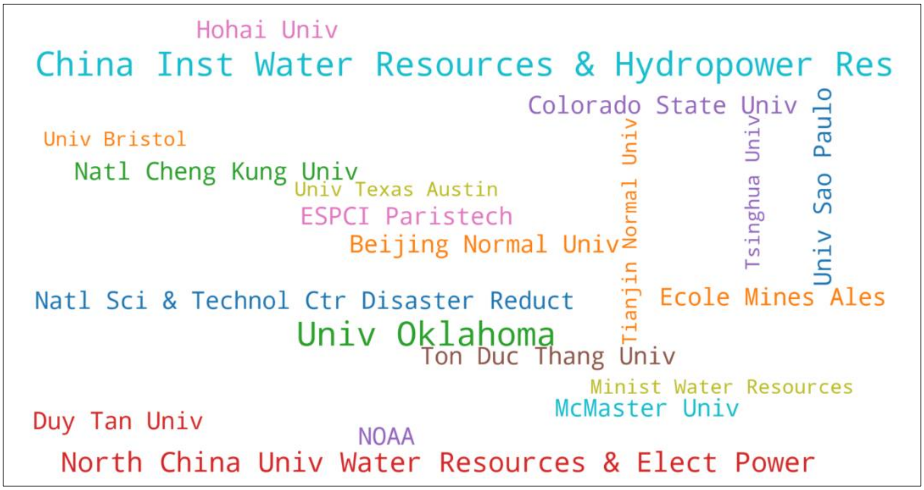

Fig. 5. Word cloud graph of the author's institutional affiliation (source: output ScientoPyGui 2.1.3)

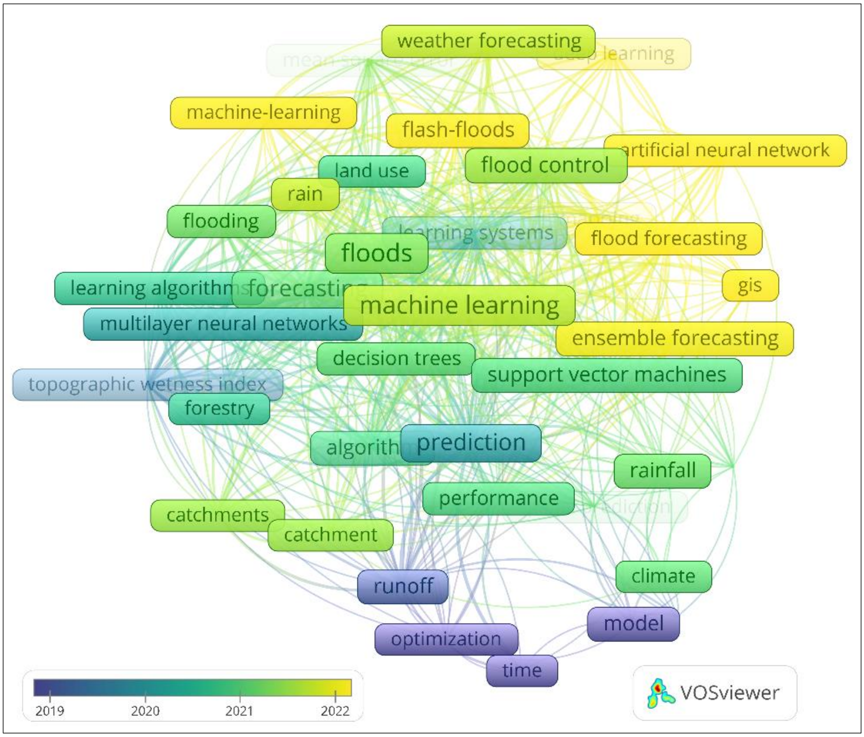

Fig. 6. Co-occurrence for index keywords using full counting network map (source: output VOSviewer 1.6.20)

4. DISCUSSIONS

This study explored the use of machine learning for flash flood forecasting by conducting a scientometric analysis of relevant papers. Our findings highlighted the rapid growth in this field, with a significant increase in publications after 2021. This underlines the growing recognition of machine learning's potential to improve flash flood prediction and mitigate its devasting consequences.

While the theoretical benefits of machine learning are promising, their real-world effectiveness remains an important area of investigation. Fortunately, several recent applications demonstrate their practical value. Below are presented implementations of machine learning in the operational forecasting in countries like China, U.S.A, Brazil, Bangladesh, Sri Lanka, Colombia.

China was implemented a long short-term memory (LSTM) method for flash flood forecasting. The LSTM based approach outperformed the benchmark rainfall triggering index (RTI) and flash flood guidance (FFG) traditional methods (Zhao *et al.*, 2022).

In U.S.A, a probabilistic forecast system for excessive rainfall, known as the Colorado State University Machine Learning Probabilities (CSU-MLP) system was developed. This system represents an example of a successful research-to-operations transition (Schumacher *et al.*, 2021)

In Brazil, a feasibility study by Lima and Scofield (2021) explored the potential of using neural networks in early warning system for operational forecasting (Lima and Scofield 2021).

From 2018 Google has expanded geographically with an operational flood forecasting system. It is available in 80 countries. In a small number of cases Google utilizes local historical and real time data provided by the following governments: Bangladesh, Sri Lanka, Colombia, Brazil. Floods alerts sent to the public via different channels: web-search, smartphone push notification, Google Maps, Flood Hub platform (https://sites.research.google/floods/l/0/0/3). Romania benefits from 14 gauges integrated within the Flood Hub system. Stage forecasting is modeled with the long short-term memory (LSTM) networks and linear models (Nevo *et al.*, 2022).

Despite the advancements, machine learning models for flash flood forecasting face certain challenges. One major obstacle is data scarcity. Furthermore, computational demands can be significant, particularly for deep learning models, posing challenges for resource-constrained environments. Addressing these challenges requires continued research efforts in data acquisition, model interpretability techniques and computationally efficient algorithms (Kratzert *et al.*, 2023).

The integration of machine learning models into existing systems can significantly improve flood management practices. Early warnings based on accurate forecasts can save lives and reduce damage. Policymakers can leverage these advancements to develop more effective early warning system, risk assessments, and infrastructure development strategies.

5. CONCLUSIONS

This paper investigated the use of machine learning methods for flash flood forecasting, based on the scientometric analysis of data extracted from the Web of Science and Scopus databases at 1st December 2023. From a total of 156 papers identified, 112 were retained for processing.

The machine learning has become a more and more used tool in hydrologic modeling for flood forecasting, consequently, the number of the scientific publications with subject in this field has grown significantly in recent years (particularly after 2021). Technological advances, as well as the increasing frequency and intensity of hydrometeorological events, have played a major role in the development of machine learning based models

The largest number of publications was found in China (28), where the most institutions involved in such studies were identified, both universities and other institutes. In the top of the countries with the most publications are also the USA (22) and India (16). Romania ranks tenth in this top with four papers.

The scientific journals that published the most articles on the use of machine learning in flash flood forecasting are *Water*, *Journal of Hydrology*, *Remote Sensing*, *Science of the Total Environment*, and *Environmental Earth Science.*

In recent years, there has been a trend towards the use of deep learning (a subfield of machine learning) for flash flood forecasting.

These findings underscore the increased interest of researchers at the global level for the development of efficient and useful tools as the machine learning, in order to improve the flash flood forecasting and mitigate the flood risk.

REFERENCES

1. Alipour A., Ahmadalipour A., Abbaszadeh P., Moradkhani H. (2020), Leveraging machine learning for predicting flash flood damage in the Southeast US, *Environmental Research Letters*, 15(2), p. 024011. DOI: https://doi.org/10.1088/1748-9326/ab6edd.
2. Bubeck P., Thieken A.H. (2018), What helps people recover from floods? Insights from a survey among flood-affected residents in Germany, *Regional Environmental Change*, 18(1), pp. 287–296. DOI: https://doi.org/10.1007/s10113-017-1200-y.
3. Cobo M.J., López-Herrera A., Herrera-Viedma V., Herrera F. (2011), Science mapping software tools: Review, analysis, and cooperative study among tools, *Journal of the American Society for Information Science and Technology,* 62(7), pp. 1382–1402. DOI: https://doi.org/10.1002/asi.21525.
4. Costache R., Arabameri A., Blaschke T., Pham Q.B., Pham B.T., Pandey M., Arora A., Linh N.T.T, Costache I. (2021), Flash-Flood Potential Mapping Using Deep Learning, Alternating Decision Trees and Data Provided by Remote Sensing Sensors, *Sensors*, 21(1), p. 280. DOI: https://doi.org/10.3390/s21010280.
5. Costache R., Arabameri A., Moayedi H., Pham Q.B., Santosh M., Nguyen H., Pandey M., Pham B.T. (2022), Flash-flood potential index estimation using fuzzy logic combined with deep learning neural network, naïve Bayes, XGBoost and classification and regression tree, *Geocarto International,* 37(23), pp. 6780–6807. DOI: https://doi.org/10.1080/10106049.2021.1948109.
6. Dazzi S., Vacondio R., Mignosa P. (2021), Flood Stage Forecasting Using Machine-Learning Methods: A Case Study on the Parma River (Italy), *Water*, 13(12), p. 1612. DOI: https://doi.org/10.3390/w13121612.
7. Diaconu D., Costache R., Popa M. (2021), An Overview of Flood Risk Analysis Methods, *Water*, 13(4), p. 474. DOI: https://doi.org/10.3390/w13040474.
8. Dtissibe F.Y., Ari A., Titouna C., Thiare O., Gueroui A. (2020), Flood forecasting based on an artificial neural network scheme, *Natural Hazards*, 104(2), pp. 1211–1237. DOI: https://doi.org/10.1007/s11069-020-04211-5.
9. Gruntfest E., Handmer, J. (2001a), *Coping With Flash Floods*. Dordrecht: Springer Netherlands. DOI: https://doi.org/10.1007/978-94-010-0918-8.
10. Gruntfest E., Handmer J. (2001b), Dealing with Flash Floods: Contemporary Issues and Future Possibilities, in E. Gruntfest and J. Handmer *Coping With Flash Floods*. Dordrecht: Springer Netherlands, pp. 3–10. DOI: https://doi.org/10.1007/978-94-010-0918-8_1.
11. Guha-Sapir D., Hoyois P., Wallemacq P., Below R. (2016) Annual Disaster Statistical Review 2016. The numbers and trends, *Brussels, Belgium*.
12. Hartnett M., Nash S. (2017), High-resolution flood modeling of urban areas using MSN_Flood, *Water Science and Engineering*, 10(3), pp. 175–183. DOI: https://doi.org/10.1016/j.wse.2017.10.003.

13. Hill A.J., Schumacher R.S. (2021), Forecasting excessive rainfall with random forests and a deterministic convection-allowing model, *Weather and Forecasting*. DOI: https://doi.org/10.1175/WAF-D-21-0026.1.
14. Kaiser M., Günnemann S., Disse M. (2020), Providing guidance on efficient flash flood documentation: an application based approach, *Journal of Hydrology*, 581, p. 124466. DOI: https://doi.org/10.1016/j.jhydrol.2019.124466.
15. Kan G., He X., Li J., Ding L., Hong Y., Zhang H., Liang K., Zhang M. (2019), Computer Aided Numerical Methods for Hydrological Model Calibration: An Overview and Recent Development, *Archives of Computational Methods in Engineering*, 26(1), pp. 35–59. DOI: https://doi.org/10.1007/s11831-017-9224-5.
16. Kratzert F., Nearing G., Addor N., Erickson T., Gauch M., Gilon O., Gudmundsson L., Hassidim A., Klotz D., Nevo S., Shalev G., Matias Y. (2023), Caravan - A global community dataset for large-sample hydrology, *Scientific Data*, 10(1), p 61. DOI: https://doi.org/10.1038/s41597-023-01975-w.
17. Li J., Goerlandt F., Reniers G. (2021), An overview of scientometric mapping for the safety science community: Methods, tools, and framework, *Safety Science*, 134, p. 105093. DOI: https://doi.org/10.1016/j.ssci.2020.105093.
18. Lima G.R., Scofield G.B. (2021), Feasibility study on operational use of neural networks in a flash flood early warning system, *RBRH*, 26, 7. DOI: https://doi.org/10.1590/2318-0331.262120200152
19. Li P., Zhao Y., Sufian M., Deifalla A.F. (2023), Scientometric analysis of flood forecasting for Asia region and discussion on machine learning methods, *Open Geosciences*, 15(1), p. 20220475. DOI: https://doi.org/10.1515/geo-2022-0475.
20. Mingers J., Leydesdorff L. (2015), A review of theory and practice in scientometrics, *European Journal of Operational Research*, 246(1), pp. 1–19. DOI: https://doi.org/10.1016/j.ejor.2015.04.002.
21. Nevo S., Morin E., Gerzi Rosenthal A., Metzger A., Barshai C., Weitzner D., Voloshin D., Kratzert F., Elidan G., Dror G., Begelman G., Nearing G., Shalev G., Noga H., Shavitt H., Yuklea L., Royz M., Giladi N., Peled Levi N., Reich O., Gilon O., Maor R., Timnat S., Shechter T., Anisimov V., Gigi V., Levin Y., Moshe Z., Ben-Haim Z., Hassidim A., Matias Y. (2022), Flood forecasting with machine learning models in an operational framework, *Hydrology and Earth System Sciences*, 26(15), pp. 4013-4032. DOI: https://doi.org/10.5194/hess-26-4013-2022
22. Nearing G.S., Kratzert F., Sampson A., Pelissier C., Klotz D (2021), What Role Does Hydrological Science Play in the Age of Machine Learning?, *Water Resources Research*, 57(3), p. e2020WR028091. DOI: https://doi.org/10.1029/2020WR028091.
23. Ruiz-Rosero J., Ramirez-Gonzalez G., Viveros-Delgado J. (2019), Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications, *Scientometrics,* 121(2), pp. 1165–1188. DOI: https://doi.org/10.1007/s11192-019-03213-w.
24. Schumacher R. S., Hill A. J., Klein M., Nelson J., Erickson M. J., Trojniak S. M., Herman G. R. (2021), From Random Forests to Flood Forecasts: A Research to Operations Success Story, *Bulletin of the American Meteorological Society*, 102(9), pp. E1742-E1755. DOI: https://doi.org/10.1175/BAMS-D-20-0186.1.
25. Stănescu Al. V., Drobot R. (2002), Masuri nestructurale de gestiune a inundatiilor, pp 18, *HGA*, Bucuresti
26 Tien Bui D., Hoang N., Pham T., Ngo P., Hoa P. (2019), A new intelligence approach based on GIS-based Multivariate Adaptive Regression Splines and metaheuristic optimization for predicting flash flood susceptible areas at high-frequency tropical

typhoon area, *Journal of Hydrology*, 575, pp. 314–326. DOI: https://doi.org/10.1016/j.jhydrol.2019.05.046.
27. Towfiqul Islam A.R.M., Talukdar S., Mahato S., Kundu S., Eibek K.U., Pham Q, Kuriqi A., Linh N.T.T. (2021), Flood susceptibility modelling using advanced ensemble machine learning models, *Geoscience Frontiers*, 12(3), p. 101075. DOI: https://doi.org/10.1016/j.gsf.2020.09.006.
28. Van Eck N.J., Waltman L. (2010), Software survey: VOSviewer, a computer program for bibliometric mapping, *Scientometrics*, 84(2), pp. 523–538. DOI: https://doi.org/10.1007/s11192-009-0146-3.
29 Zhao G., Liu R., Yang M., Tu T., Ma M., Hong Y., Wang S. (2022), Large-scale flash flood warning in China using deep learning, *Journal of Hydrology*, 604, p. 127222. DOI: https://doi.org/10.1016/j.jhydrol.2021.127222
30. *** (2011), Manual on flood forecasting and warning, WMO-No 1072, pp 23, *Chairperson Publications Board*, Switzerland