CHEMICAL ENGINEERING TRANSACTIONS
VOL. 114, 2024
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Petar S. Varbanov, Min Zeng, Xuechao Wang, Bohong Wang

ISBN 979-12-81206-12-0; **ISSN** 2283-9216
DOI: 10.3303/CET24114171

Feasibility of Using Statistical Forecasting Method in the Marcal Catchment Area

Máté Szabó[a], Katalin Bene*[,b], Gábor Kerék[b]

[a]Veszprémi Tervező Kft, Veszprém, Hungary
[b] National Laboratory for Water Science and Water Security, Széchenyi István University, Department of Transport Infrastructure and Water Resources Engineering, Egyetem square 1., H-9026 Győr, Hungary
benekati@sze.hu

Flooding is one of the most destructive natural disasters, posing significant risks to under-construction and existing structures. It can also compromise critical infrastructure, such as roads and railways, by weakening embankments. While infrastructure damage is severe, the foremost concern remains the population's safety, making technological advancements and timely information dissemination crucial. Flood forecasting is vital in preparing communities and enabling flood defense organizations to respond effectively. This study aimed to develop a reliable flood forecasting method for the downstream sections of the Marcal River, where population density is high, using real-time data.
Accurate flood forecasting relies on a comprehensive monitoring network and precise measurements that predict water flow and other hydrological conditions over several days. Real-time data during flood events is also essential for emergency response. Key hydrological and meteorological factors, including water levels, flow rates, and precipitation, are integral to this process. The study analyzed daily water flow data from 1960 to 2018, collected from stations along the Marcal River and its tributaries, combined with precipitation data, to forecast the river's flow at its outlet in Mórichida. Multi-level regression analysis, incorporating first- and second-order polynomials, was used to predict flood peaks at this outflow. The model employed flood wave peaks and simultaneous rising or receding flows from five additional river stations. Focusing on events with peak flows exceeding 20 m^3/s, the researchers identified 68 cases, with 9-20 measurements per event. Confidence and prediction intervals confirmed the model's accuracy, predicting flood peaks within ±10 m^3/s, offering a reliable, less complex alternative to traditional models.

1. Introduction

Floods can be the most damaging and pervasive natural disasters, harming the environment and the global economy (Grigorieva and Livenets, 2022). The frequency and severity of floods have increased recently, highlighting the critical need for precise flood modeling to help disaster response and management (Ámon and Bene, 2023). Floods can cause enormous property damage, animal and human fatalities, and agriculture loss (Kumar et al., 2023).
Flood projection is a critical component of water resource management and urban planning, aiming to mitigate the adverse effects of flooding (Jain et al., 2003). Regression analyses have been extensively used to model the relationship between flood-related variables (precipitation, river discharge, and water levels) and the outcomes of flooding events (like flood extent, depth, and frequency) (Plitnick et al., 2018). Early studies often utilized linear regression models due to their simplicity and interpretability (Chen et. al., 2023). These models have been applied to predict flood levels based on upstream river flow rates and precipitation data (Wu et al., 2019). Multiple Linear Regression (MLR) models consider multiple predictors simultaneously, offering a more detailed understanding of flood dynamics. They have incorporated various hydrological and meteorological factors in flood prediction (Patel et al., 2016). These models help assess the probability of flooding given specific conditions. Given the complex nature of flood dynamics, non-linear regression models, including polynomial and spline regression, have been applied to capture the non-linear relationships between variables (Mengzhu et al., 2023). Advanced regression techniques involving machine learning, such as Random Forests and Support

Paper Received: 8 July 2024; Revised: 29 October 2024; Accepted: 26 November 2024
Please cite this article as: Szabó M., Bene K., Kerék G., 2024, Feasibility of Using Statistical Forecasting Method in the Marcal Catchment Area, Chemical Engineering Transactions, 114, 1021-1026 DOI:10.3303/CET24114171

Vector Regression, have been increasingly used because they can effectively handle non-linearities and interactions between predictors (Lange, 2020).
Regression analyses are still widely used tools for flood prediction (Plitnick et al., 2018). In regression analyses, the selection of predictor variables is crucial. Studies have highlighted the importance of incorporating various variables, including meteorological, hydrological, and land use factors, to improve model accuracy. This research used a second-order multiple regression model to predict floods on the Marcal River at Mórichida-Rábaszentmiklós. The new approach used two to five predictor variables, including the upstream peak flows in the tributaries and the simultaneously occurring flow in the other tributaries. Using concurrently occurring flows from different tributaries with peak flows can significantly improve prediction accuracy. Rainfall and land use conditions were considered as well. Since there was limited information on the change in land use conditions for the analyzed 30 y, land use was not considered. Rainfall data was evaluated but did not improve prediction accuracy.

2. Site description

The catchment is located in western Hungary, surrounded by the Bakony Mountains, the Sokoróalja hills, the slopes of the Kemeneshat, and the southern part of the Kisalföld. This area covers 3,084 km² and is characterized by diverse topography: 5 % mountainous, 25 % hilly, and 70 % flat terrain. The small streams traversing the mountainous and hilly regions are quickly responding, (flashy) streams and ephemeral. Their water flow is minimal for most of the year. However, they are subject to swift increases in flow due to climatic phenomena such as quick snow melts and intense summer rainfalls, leading to potential flash floods. The absence of designated floodplains alongside these streams means excessive floods can overwhelm channel capacities, subsequently inundating adjacent lands (Éduvizig, 2020).
The Marcal River lies on the left side of the valley, with a significant part of its tributaries flowing from the Bakony mountain. The river valley and the catchment on the left are flatlands. Its geographical location also influences the flood conditions. Therefore, it is essential to identify the timing and peak in the tributaries (Éduvizig, 2020). The most important tributaries of the Marcal River are Torna – Karako, Hajagos – Nemesszalók, Bitva – Mihályháza, and Gerence – Takácsi. The following figure shows the Marcal catchment area and the flow-measuring locations. Annual precipitation is 645 mm, and rainfall is highest between May and August, with June being the wettest month.

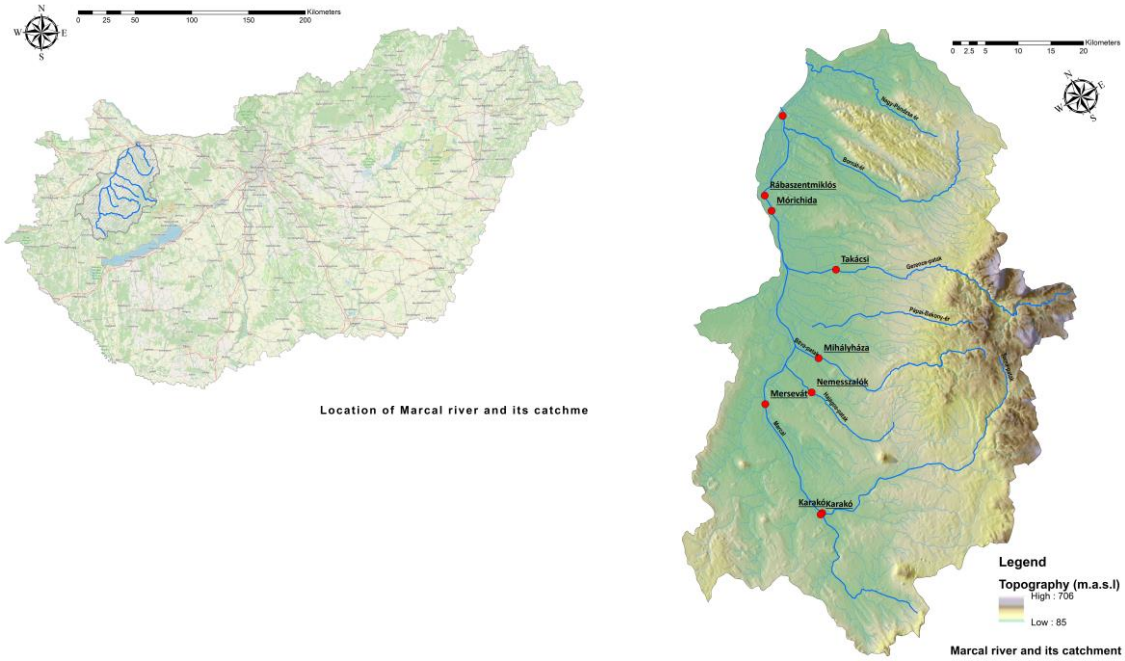

Figure 1: Watershed location, measuring stations, and tributaries

3. Methods

3.1 Data processing

The initial measurements began in 1961 at three locations: Karakó (Torna), Takácsi (Gerence), and Mórichida (Marcal). Additional measurement stations were added in 1971 at Mihályháza (Bitva) and in 1985 at Nemesszalók (Hajagos). In 2014, due to the absence of measurements at Mórichida, the Rábaszentmiklós measurements were used from January 1 to December 31. A summary of the data set can be seen in Table 1. stations.

Table 1: Flow measurements

Location	Measurement	Min	Max	Average	Number of data
Karakó (Torna)	1960-2018	0.076	49.8	1.78	191,228
Nemesszalók (Hajagos)	1985-2018	0	11.5	0.47	4,220
Mihályháza (Bitva)	1971-2018	0	9.29	0.37	106,610
Takácsi (Gerence)	1960-2018	0	29.8	0.87	82,667
Mórichida (Marcal)	1960-2018	0	146	5.74	136,662

The dataset for statistical analysis was developed through several stages. Initially, predictor events were chosen, followed by concurrent peak flows in tributaries and, finally, flows at other measuring locations that coincided. For flood forecasting purposes, events exceeding 20 m^3/s at Mórichida (Marcal) were selected, identifying 68 such events. The 20 m^3/s threshold was determined based on the experience of the water resources agency. Upstream peak flows in the tributaries were gathered for each selected peak event. Typically, these upstream events occurred 1-3 days before the Marcal outflow peak. On Karakó, Takácsi, and Mórichida, 68 peak events were collected, while Mihályháza and Nemesszalók recorded 29 and 20 events, due to their shorter measurement periods. Simultaneously occurring flows at other locations were also gathered during peak events. Figure 2 illustrates the data collection strategy for one event.

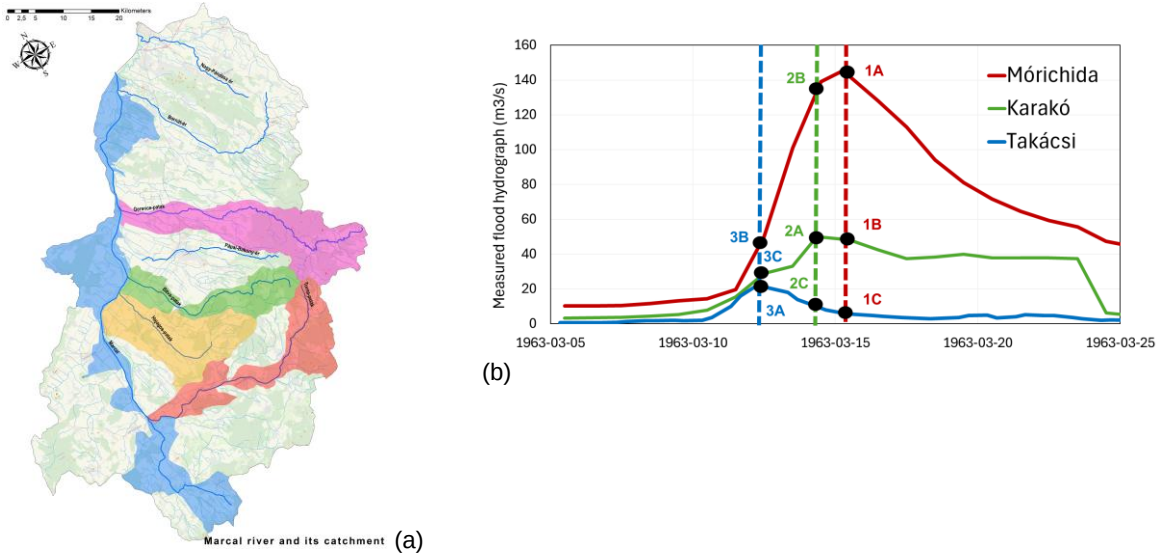

Figure 2: Measuring stations (a) and event evaluation method (b)

The peak at Mórichida (1A) was identified first, followed by the collection of two additional peaks (later expanded to five) that occurred earlier in the same event (2A and 3A). For instance, as shown in Figure 2, during the March 1963 event, the Mórichida peak (1A) was identified, along with the Karakó peak (2A) and the Takácsi peak (3A). For the Mórichida peak (1A), flows from Karakó (1B) and Takácsi (1C) were also recorded. The peak from Karakó (2A) and concurrent flows at Mórichida (2B) and Takácsi (2C) were added next, followed by the Takácsi peak (3A), with corresponding flows at Mórichida (3B) and Karakó (3C). Before 1971, this process resulted in nine data points per event, while after 1971, the number increased to 25 per event due to the addition of more measurements.

3.2 Statistical methods

Regression analyses were applied to determine the relationship between the measured flows in the tributaries and the predicted downstream peak flow. The variables were screened using cross-correlation analyses to identify the most important parameters. Added variable plots further reduced the number of independent variables. Finally, backward analyses were applied to determine the best model. Multiple regression analyses were used to predict flood peaks, and first- and second-order polynomial equations were considered during the study. The second-order multiple polynomial regression equation can be written as:

$$y_i = b_0 + b_{i1}x_{i1} + b_{i2}x_{i2}^2 \ldots\ldots + b_{n1}x_{n1} + b_{n2}x_{n2}^2 + \varepsilon_i \qquad where\ i = 1, \cdots n \tag{1}$$

Where y is the dependent variable, x_i is the independent variable, and b_i is the unknown parameter.
The adjusted coefficient of multiple determination, R^2_{adj} was used to measure regression fit. R^2_{adj} adjusts R^2 by dividing each sum of squares by its associated degrees of freedom. Consequently:

$$R^2_{adj} = 1 - \frac{(1-R^2)\times(n-1)}{n-k-1} \tag{2}$$

Collinearity diagnostics and residual analyses were performed to evaluate the underlying assumptions of the regression models. The gvlma package in R (Edsel et al., 2019) was utilized to test for skewness, kurtosis, and heteroscedasticity, ensuring the validity of the linear regression assumptions. These diagnostic metrics drove model refinement and selection to optimize performance. The dataset was not partitioned, allowing the models to capture a wide spectrum of hydrological conditions to enhance model accuracy and generalizability. This approach increases the robustness of the models, with the potential for recalibration as more data becomes available. A 90 % confidence interval was calculated for the selected models to quantify the forecasting uncertainty and assess model accuracy.

4. Results

Two locations with shorter data sets were not included in the analysis. Cross-correlation analysis was used to identify the most important predictors for the 68 events, with nine predictors available for each event. Backward elimination was applied to determine the four most appropriate models, each with four independent variables. In every model, the Karakó peak flow (KT) was included, with the Mórichida flow at the Karakó peak (KTRP) being the second most important predictor. Other key variables were the Takácsi flow at the Karakó peak (KTTP) and the Mórichida flow at the Takácsi peak (TTRP). Ultimately, two forecasting models were selected: one with four independent variables and another with two. In Model 1, the independent variables are KT, KTRP, TTRP, and TT, while in Model 2, they are KT and KTRP. The coefficients of the models, along with the adjusted R^2 values, are presented in Table 2.

Table 2: Regression results for Model 1 and Model 2, and reduced models Model R1 and Model R2

Coefficients	Model 1	Model 2	Model R1	Model R2
b0	6.087	11.016	7.732	12.360
b11 (KT)	1.302	1.424	1.101	1.118
b12 (KT)	-0.028	- 0.030	-0.026	-0.028
b21 (KTRP)	0.157	0.305	0.241	0.318
b22 (KTRP)	0.006	0.005	0.006	0.006
b31 (TTRP)	0.491		0.300	
b32 (TTRP)	-0.005		-0.003	
b41 (TT)	0.107		0.355	
b42 (TT)	-0.005		-0.011	
R^2_{adj}	0.802	0.801	0.933	0.930

Based on the adjusted R^2, the differences between the two models are insignificant. Model predictions (Model 1 and Model 2) and the observed peaks at Mórichida are shown in Figure 3(a). When comparing the peak flow predictions at the Mórichida location, Model 1 provides a more accurate prediction than Model 2. In Model 2, just two variables—the Mórichida flow during the Karakó peak discharge and the Karakó peak discharge—are sufficient to predict the peak at Mórichida. Residual analysis and multicollinearity tests were performed for both models to assess the suitability of the regression models. Partial regression plots showed a linear relationship between KT and KTRP, while the other two independent variables had little relationship. Collinearity for both models was moderate. From the normal probability standardized residuals plot (Q-Q plot, Figure 3b), most points lie close to the line, except for three events, indicating a non-normal error distribution. Both models met the heteroskedasticity assumption. However, skewness and kurtosis statistics showed that the distribution of residuals is not normal. In events 19, 28, and 35, the difference between measured and predicted peak flows

was particularly large, and these outliers were evident in the error analysis. These three events were closely investigated to understand the unusual behavior. During these events, the peak flow at Mórichida ranged from 60 to 76 m³/s, while the tributary peak flows at Takácsi and Karakó were between 12 and 28 m³/s. The peaks at the tributaries were significantly smaller than at Mórichida, whereas, in most cases, the differences are much smaller. This discrepancy could be due to measurement errors or another hydraulic influence on the Marcal River, such as a backwater effect caused by simultaneous flooding of the Danube. Since no clear explanation for these outliers was found, these three events were removed from the dataset, and the regression coefficients were recalculated.

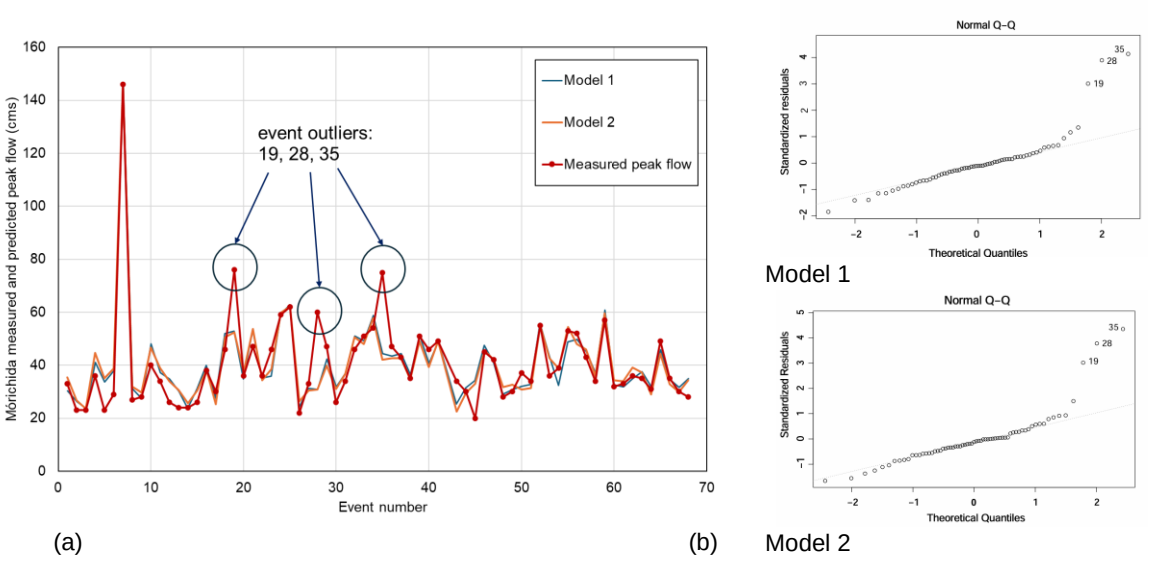

Figure 3: (a) Regression analysis results predicted and measured peak flow at Mórichida, (b) error statistics for Models 1 and 2

Table 2 shows the calculated regression coefficients for the reduced Models R1 and R2. The adjusted R2 improved, and its value is close to 0.93; the differences between the two models are insignificant. The collinearity and error statistics assumptions for both Models were acceptable. Figure 4 shows the measured and predicted peak flows and the 95 % confidence intervals for Model R1.

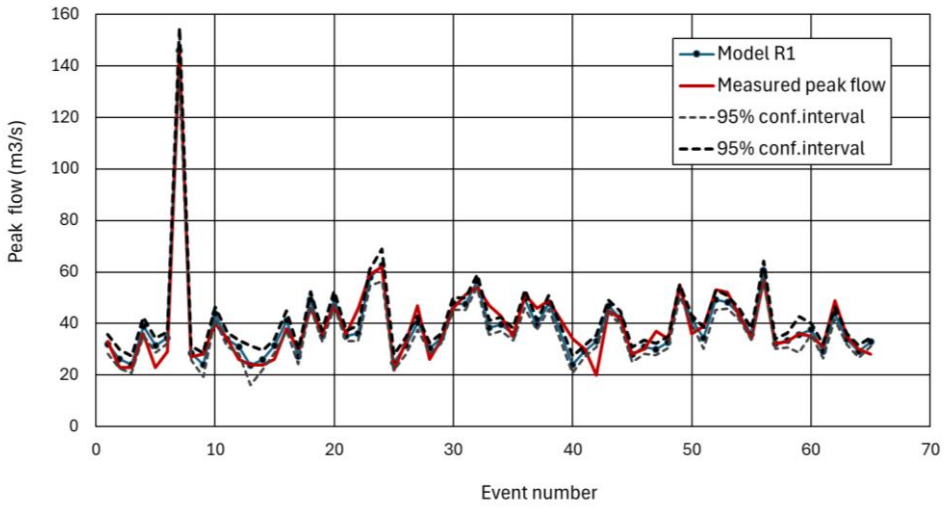

Figure 4: Measured and predicted peak flow at Mórichida for Model R1 with a 95% confidence interval

The figure shows little difference between the predicted and measured peaks at most events, and most predicted events are within the 95 % confidence interval. The impact of rainfall and whether including rainfall measurement increases prediction accuracy were investigated. There are six rainfall measuring stations on the Marcal watershed, but only at two locations (within 6 km distance) where the combined measurements were available

between 1960-2018. The added variable in the model did not change the adjusted R^2 value in either model. The location of the rain gauge can explain the small impact of the rainfall addition in the models since it is close to the Mórichida measuring gauge. Better rainfall measurement availability could have improved model prediction.

5. Conclusion

This study developed a second-order multiple regression model to predict flood peaks on the Marcal River at Mórichida-Rábaszentmiklós, using two to five predictor variables, including upstream peak flows and simultaneous flows from other tributaries. Incorporating concurrent flows improved the accuracy of the flood predictions. Two primary models were refined: Model 1, with four independent variables, and Model 2, with two. Model 1 offered slightly higher accuracy, and both models demonstrated strong predictive capabilities supported by high adjusted R^2 values and acceptable error margins. These models are effective tools for forecasting peak flows in the Marcal River basin, with future enhancements focusing on improving data accuracy and rainfall measurements. The statistical approach offers a simplified and efficient alternative to traditional numerical models, which are often resource-intensive. It is particularly suited for real-time forecasting, enabling quicker decision-making during flood events. The method also supports data-driven decision-making by identifying the most significant predictors using multi-level regression and backward elimination. The method ensures that the forecasts are grounded in reliable statistical analysis, helping water management authorities make informed decisions. The model's adaptability allows using various hydrological variables, making it applicable to different river systems. Its scalability makes it an attractive option for regions where numerical models may not be feasible due to limited resources or data availability. Overall, this approach provides an accessible, efficient, and scalable method for flood forecasting, improving disaster preparedness and resource allocation.

References

Ámon G., Bene K., 2023, Impact of Different Rainfall Intensity and Duration on Flash-Flood Events on a Steep Sloped Ungauged Watershed. Chemical Engineering Transactions, 107, 175-180, DOI: 10.3303/CET23107030.

Chen M., Papadikis K., Jun Ch., Macdonald N., 2023, Linear, nonlinear, parametric and nonparametric regression models for nonstationary flood frequency analysis. Journal of Hydrology, 616, 128772, DOI: 10.1016/j.jhydrol.2022.128772.

de Castro J.T., Salistre G.M. Jr, Byun Y.-Ch., Gerardo B.D., 2013, Flash Flood Prediction Model based on Multiple Regression Analysis for Decision Support System. Proceedings of the World Congress on Engineering and Computer Science, Vol II WCECS 2013, 23-25 October 2013, San Francisco, USA, ISBN: 978-988-19253-1-2.

Edsel A., Slate P., Slate E.H., 2019, gvlma: Global Validation of Linear Models Assumptions. R package version 1.0.0.3, <https://cran.r-project.org/web/packages/gvlma/gvlma.pdf>, accessed 10.10.2024.

ÉDUVÍZIG, 2020, 1-4 Marcal water resources management plan (in Hungarian), <http://www2.vizeink.hu/files3/1_4_Marcal.pdf>, accessed 10.10.2024.

Grigorieva E., Livenets A., 2022, Risks to the Health of Russian Population from Floods and Droughts in 2010–2020: A Scoping Review. Climate, 10, 37, DOI: 10.3390/cli10030037.

Jain A., Prasad S.K.V., 2003, Comparative analysis of event based rainfall modeling techniques–Deterministic, statistical and Artificial Neural Network. J. Hydrol. Eng., 8, 93–98, DOI: 10.1061/(ASCE)1084-0699(2003)8:2(93).

Kumar V., Sharma K.V., Caloiero T., Mehta D.J., Singh K., 2023, Comprehensive Overview of Flood Modeling Approaches: A Review of Recent Advances. Hydrology, 10(7), 141, DOI: 10.3390/hydrology10070141.

Lange H., Sippel S., 2020, Machine Learning Applications in Hydrology. In: Levia, D.F., Carlyle-Moses, D.E., Iida, S., Michalzik, B., Nanko, K., Tischer, A. (Eds.) Forest-Water Interactions, Ecological Studies, 240, DOI: 10.1007/978-3-030-26086-6_10.

Mengzhu Ch., Papadikis K., Jun Ch., Macdonald N., 2023, Linear, nonlinear, parametric and nonparametric regression models for nonstationary flood frequency analysis. Journal of Hydrology, 616, 128772, DOI: 10.1016/j.jhydrol.2022.128772.

Patel Sh., Hardaha M.K., Seetpal M., Madankar K.K., 2016, Multiple Linear Regression Model for Stream Flow Estimation of Wainganga River. American Journal of Water Science and Engineering, 2(1), 1-5, DOI: 10.11648/j.ajwse.20160201.11.

Plitnick T.A., Marsellos A.E., Tsakiri K.G., 2018, Time Series Regression for Forecasting Flood Events in Schenectady, New York. Geosciences, 8, 317, DOI: 10.3390/geosciences8090317.

Wu J., Liu H., Wei G., Song T., Zhang C., Zhou H., 2019, Flash Flood Forecasting Using Support Vector Regression Model in a Small Mountainous Catchment. Water, 11, 1327, DOI: 10.3390/w11071327.