How to cite: Taşar B., Üneş F., Demirci M., Güzel H., Varçin H. (2024) Suspended Sediment Estimation Using Machine Learning Methods. *2024 "Air and Water – Components of the Environment" Conference Proceedings*, Cluj-Napoca, Romania, p. 105-114, DOI: 10.24193/AWC2024_10.

SUSPENDED SEDIMENT ESTIMATION USING MACHINE LEARNING METHODS

BESTAMI TAŞAR [1]***, FATIH ÜNEŞ***[2]****, MUSTAFA DEMİRCİ***[3]***, HASAN GÜZEL***[4] ***HAKAN VARÇİN***[5]

DOI: 10.24193/AWC2024_10

ABSTRACT. Suspended Sediment Estimation Using Machine Learning Methods. Suspended sediment in rivers is important for efficiently using water resources and hydraulic structures. In this study, the suspended sediment load of rivers was estimated using traditional multi-linear regression (MLR), machine learning methods such as the support vector machines (SVM) and M5 decision tree (M5T). Data on daily stream flow, daily maximum and minimum water temperature and suspended sediment concentration in the river were used as input data in all models to predict daily suspended sediment discharge. The performance of all methods is evaluated based on a statistical approach. Determination coefficient (R^2), root mean square error (RMSE) and mean absolute error (MAE) are used as comparison criteria. Overall, the machine learning approaches better predict suspended sediment discharge.

Keywords: Sediment Discharge, Prediction, Linear regression, Support Vector Machines, M5 tree.

Introduction

Accurate prediction of suspended sediment is of great importance in understanding the morphology of the river and utility water supply problems. Suspended sediment load in streams can be determined by different methods such as direct measurements at sediment observation stations, sediment rating curves, regression, artificial intelligence methods, and empirical approaches based on experimental studies. Although direct measurements from sediment observation stations are the most reliable way of determining the sediment material, it is a time-consuming, costly, and error-prone method due to the sampling procedure (Olive & Rieger, 1988; Öztürk & Apaydın,2001). Another method is artificial intelligence techniques or flexible calculation methods. Flexible calculation methods attempt to model suspended sediments using techniques such as Artificial Neural Networks (ANN), Fuzzy Logic Systems (FL), Adaptive Neural Fuzzy Systems (ANFIS), or

[1] Iskenderun Technical University, Hatay – TURKEY, e-mail: bestami.tasar@iste.edu.tr

[2]* Iskenderun Technical University, Hatay – TURKEY, e-mail: fatih.unes@iste.edu.tr

[3] Iskenderun Technical University, /Hatay – TURKEY, e-mail: mustafa.demirci@iste.edu.tr

[4] Iskenderun Technical University, Hatay – TURKEY, e-mail: hasan.guzel@iste.edu.tr

[5] Iskenderun Technical University, Hatay – TURKEY, e-mail: hakan.varcin@iste.edu.tr

Genetic Algorithm (GA) using various inputs. Empirical approaches are another method used in the literature based on experimental studies used in predicting suspended sediment (Lane and Kalinske 1941, Einstein 1950, Brooks 1963). When the literature is examined, some studies draw attention, especially under these groupings.

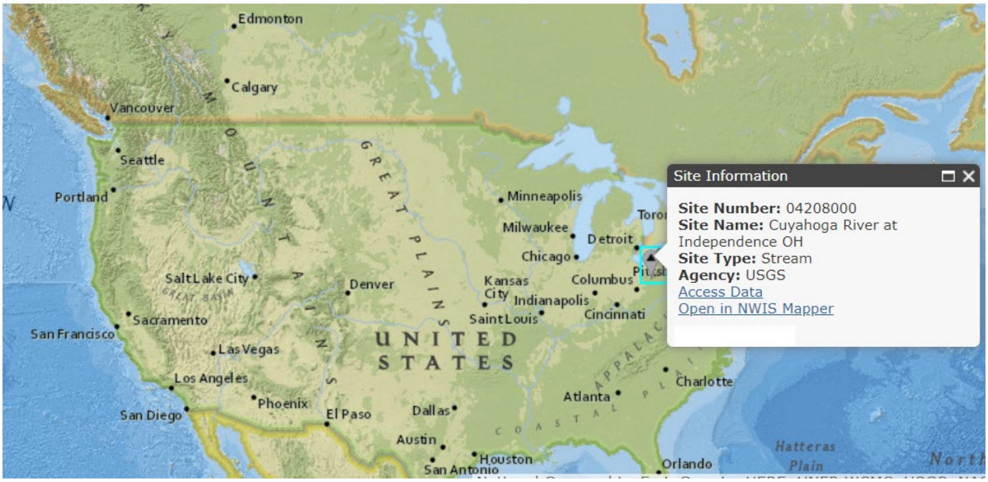
In recent years, artificial intelligence approaches have been widely used in water resource management and hydrological projects (Melesse et al. (2011); Üneş and Demirci, (2015); Üneş et al. (2020); Baek et al. (2020); Han and Morrison (2022)). Memarian et al. (2013) observed the sediment load by combining artificial neural networks with genetic algorithms. Demirci and Baltaci (2013) predicted suspended sediment in a river using fuzzy logic. Liu et al. (2013) estimate daily suspended sediment concentration in the Yellow River Basin within the borders of China with daily data covering 2193 between 1967 and 1972 using the wavelet transform-ANN method. Demirci et al. (2015) studied suspended sediment estimation using an artificial intelligence approach. Zounemat-Kermani et al. (2016) investigated the usability of artificial neural networks and support vector regression (SVR) models using an 8-year data series of three separate hydrometric stations for the suspended sediment concentration prediction. Taşar et al. (2017) forecasted suspended sediment in rivers using an artificial neural networks approach. Sari et al. (2018) estimated suspended sediment concentration from monitored data of turbidity and water level using artificial neural networks. Yadav and Satyannarayana (2020) estimated suspended sediment yield using multi-objective genetic algorithm optimization of an artificial neural network in the Mahanadi River basin, India.

This study aims to improve reliable and accurate mean-daily flow sediment load discharge models using various machine learning techniques. Suspended sediment discharge was predicted using hydro-meteorological parameters such as daily river flow, suspended sediment concentration, and water temperature (maximum and minimum) measured between 1969 and 1974 at the Ohio State Cuyahoga County Station on the Cuyahoga River. For the estimation of the amount of sediment; Multiple Linear Regression (MLR), support vector machines (SVM) and M5 Decision Tree (M5T) models were used.

Study Area

Data from Cuyahoga County Station on Cuyahoga River in Ohio (USGS Station No. 04208000, latitude 41°23'43", longitude 81°37'48"), operated by the US Geological Survey (USGS), were used in the study. The Cuyahoga River is located in the northeast of Ohio, USA, and feeds Lake Erie. Daily sediment discharge changes between the years 1969 and 1974 are shown in Fig. 2. While creating the models, 80% of the data set was used as training and 20% as a test. Total of 1659 daily data;

the Training process was done with the first 1324 days of data and the remaining 335 data were applied as testing. Test data were taken into account in model performances.



Site Number: 04208000
Site Name: Cuyahoga River at Independence OH
Site Type: Stream
Agency: USGS

Figure 1. The location of the Cuyahoga County Station in Ohio (USGS)

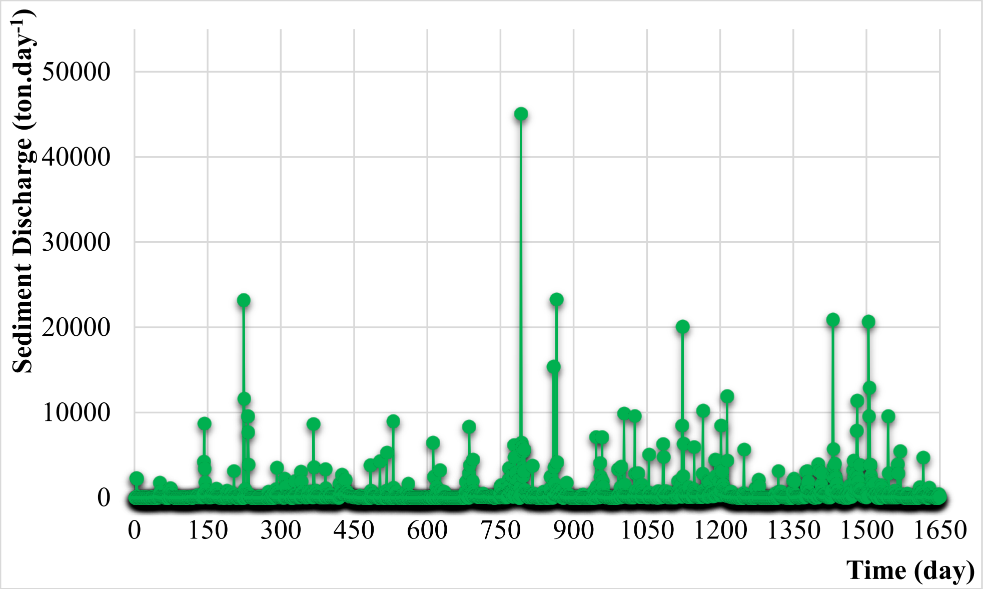


Figure 2. Sediment discharge fluctuations of Cuyahoga County Station

Methods

Multi-Linear Regression (MLR):

It is accepted that there is a relationship between the variables in problems expressed with two or more variables. The general equation of the multi-linear regression method used to determine the effect of the independent variables on the dependent variables was given in Equation 2(Güzel et al.,2023)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad (1)$$

Support vector machines (SVM)

Support vector machines (SVM), which include regression and classification forms, have been introduced by Vapnik as a robust and important learning tool (Vapnik, 1995). Since then, there has been an increasing amount of research into the application of SVMs over time. In recent years, SVMs have been used as a new learning approach in water resources. In the SVM model, sediment discharge (S_D) was predicted using the data of daily river flow (Q), suspended sediment concentration (S_c), and maximum water temperature (T_{max}), minimum water temperature (T_{min}), using the polynomial kernel function.

M5 Decision Tree Method (M5T)

The M5 decision tree algorithm was originally developed by Quinlan (1992). A detailed description of this technique can be found in Witten and Frank (2005). A brief description of this technique is as follows. The M5 algorithm generates a regression sequence by iteratively dividing the sample space using tests on a single attribute that maximizes the variance in the target space. The mathematical formula for calculating standard deviation reduction (SDR):

$$SDR = sd(T) - \sum I \frac{Ti}{T} I \; sdITiI \qquad (1)$$

where T represents a set of samples reaching the node, Ti represents the subset of samples resulting from the potential set, and sd is the standard deviation.

Results

The results of multiple linear regression (MLR), Support vector machines (SVM), and M5 Decision Tree (M5T), models for test data are given as follows. For each model, root mean squared error (RMSE), mean absolute error (MAE) and determination coefficients (R^2) between model estimates and observed values are calculated. The results are also used to compare the performance of model estimation and observation data. RMSE and MAE are determined as follows.

$$RMSE = \sqrt{\frac{1}{N} * \left(\sum_{i=1}^{N} Yi_{measure} - Yi_{prediction}\right)^2} \qquad (2)$$

$$MAE = \frac{1}{N} * \sum_{i=1}^{N} \left| Yi_{measure} - Yi_{prediction} \right| \qquad (3)$$

Where N is the data set numbers and Y_i shows the sediment discharge data. The performance of model results is shown in Table 1.

Table 1. Statistical Results for Prediction Models

Models	**MAE (ton.day^{-1})**	**RMSE (ton.day^{-1})**	**R^2**
MLR	468.79	898.05	0.841
SVM	206.89	546.76	0.944
M5 TREE	**115.81**	**346.74**	**0.976**

For all models, sediment discharge (S_D) was predicted using the data of daily river flow (Q), suspended sediment concentration (Sc), and maximum water temperature (T_{max}), minimum water temperature (T_{min}). When Table 1 was examined, all models gave different results. According to the RMSE, MAE, and R^2 criteria, the best results were obtained in M5 Tree, and SVM models. MLR model gave the worst results in all criteria.

MLR Result

For MLR analysis, linear model includes constant and linear terms. Eqs. (4) were developed for S_D prediction by linear model.

$$y = S_D = -930.34 + 50.67T_{min} - 41.86T_{max} + 22.67Q + 7.02S_c \quad (4)$$

In Figure 3a and Figure 3b, the distribution and scatter graphs of MLR are shown respectively. The determination coefficient was obtained as R^2 = 0.841 in the scatter graph. As seen in Figure 3a, it is observed that the observed daily real-time suspended discharge values in the test phase MLR give estimates far from the real values. It has been observed in the scatter and distribution plots that the MLR values are lower than the real values. When table 1 was examined, MLR models showed the worst performance. according to MAE, RMSE, and R^2 (468.79, 898.05, 0.841)

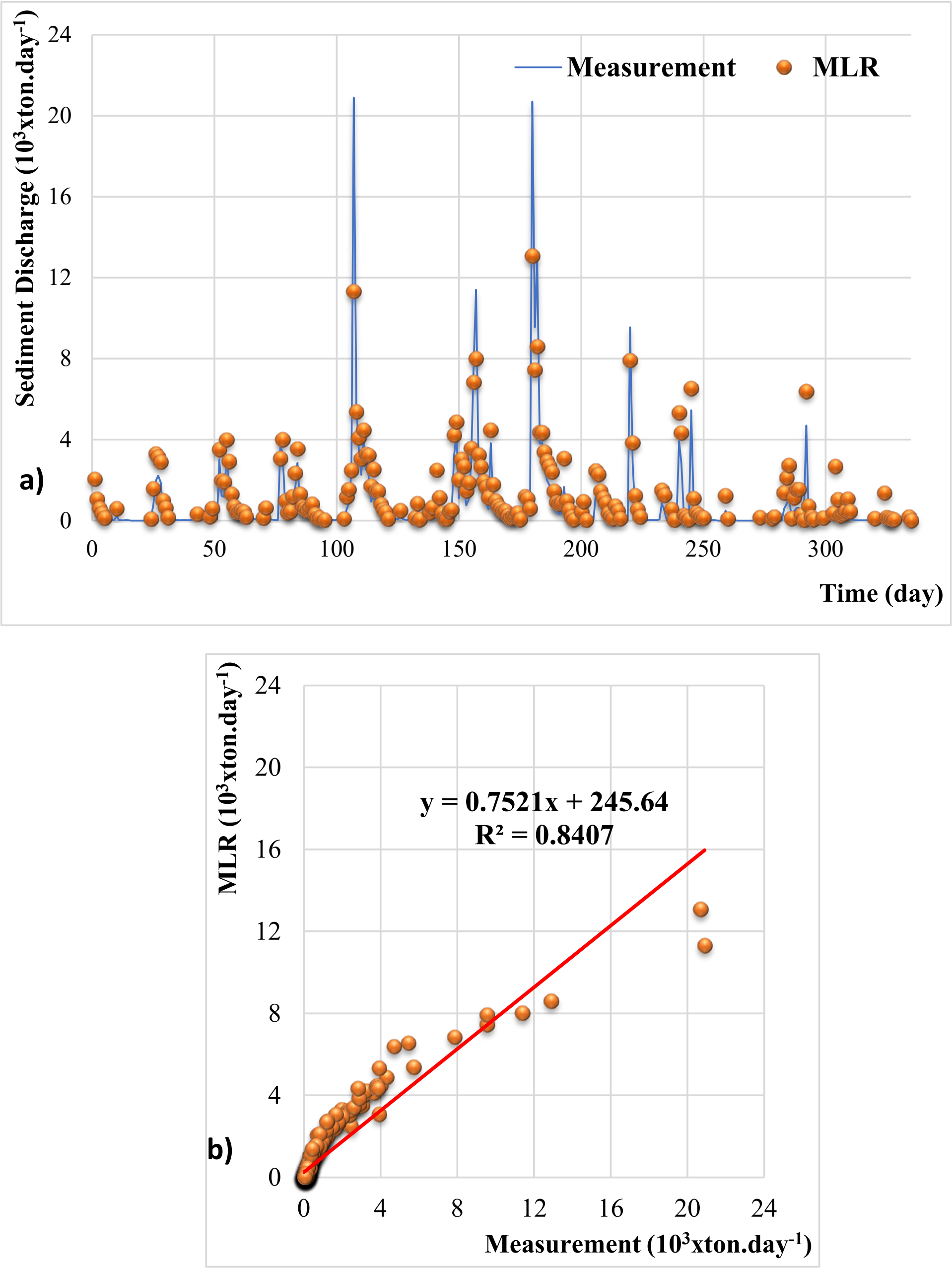


Figure 3. Measurement and MLR for testing data: a) distribution b) scatter graph

SVM Results

For the SVM model, distribution and scatter graphs are shown in Figure 4a and Figure 4b separately. As can be seen in Figure 1, the determination coefficient $R^2 = 0.944$ was obtained. Although SVM daily real-time sediment discharge values in the test phase give better results than MLR values, it is observed that they give distant predictive values to the real values.

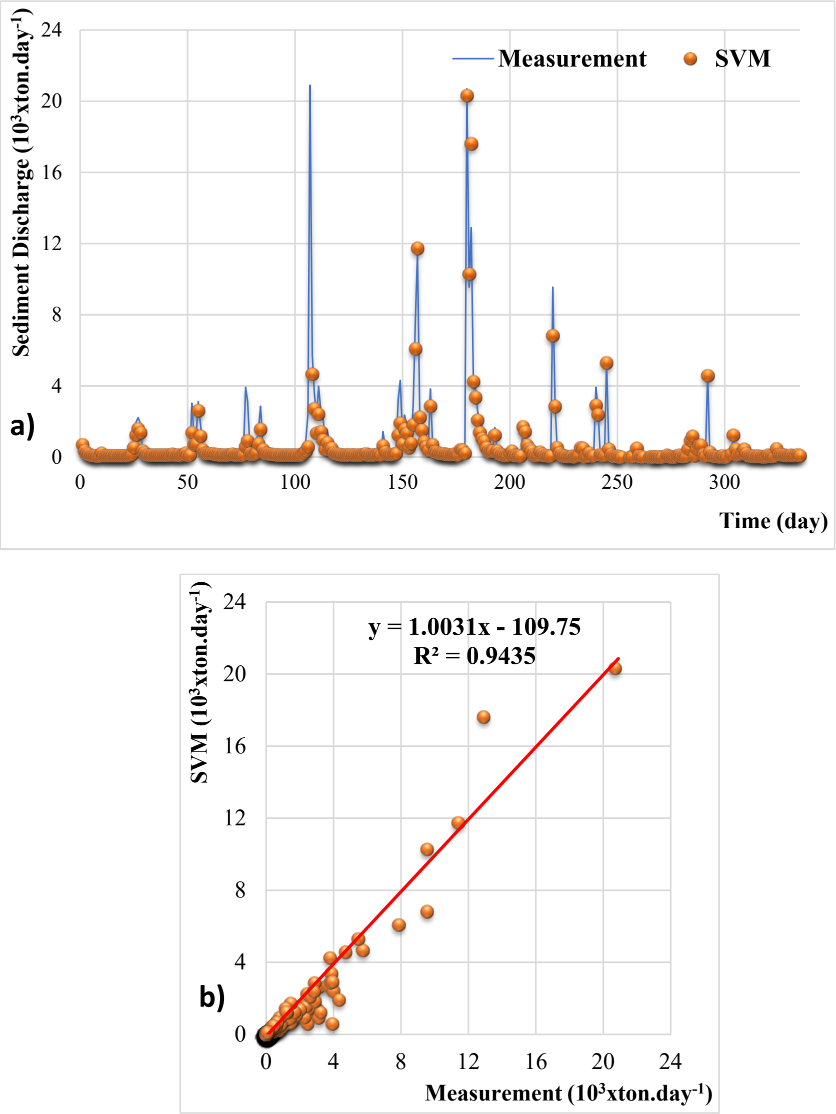


Figure 4. Measurement and SVM for testing data a) distribution b) scatter graph

M5 Tree Results:

Figure 5a. and Figure 5b shows the distribution and scatter diagrams of the estimated test results, respectively. In Figure 5a, M5T prediction values are seen near the actual values. The determination coefficient was obtained as R^2 = 0.976 as seen in Figure 14b. Results of M5T prediction values of daily real-time suspended sediment discharge are better than MLR prediction values and the good estimated results are observed according to the actual values.

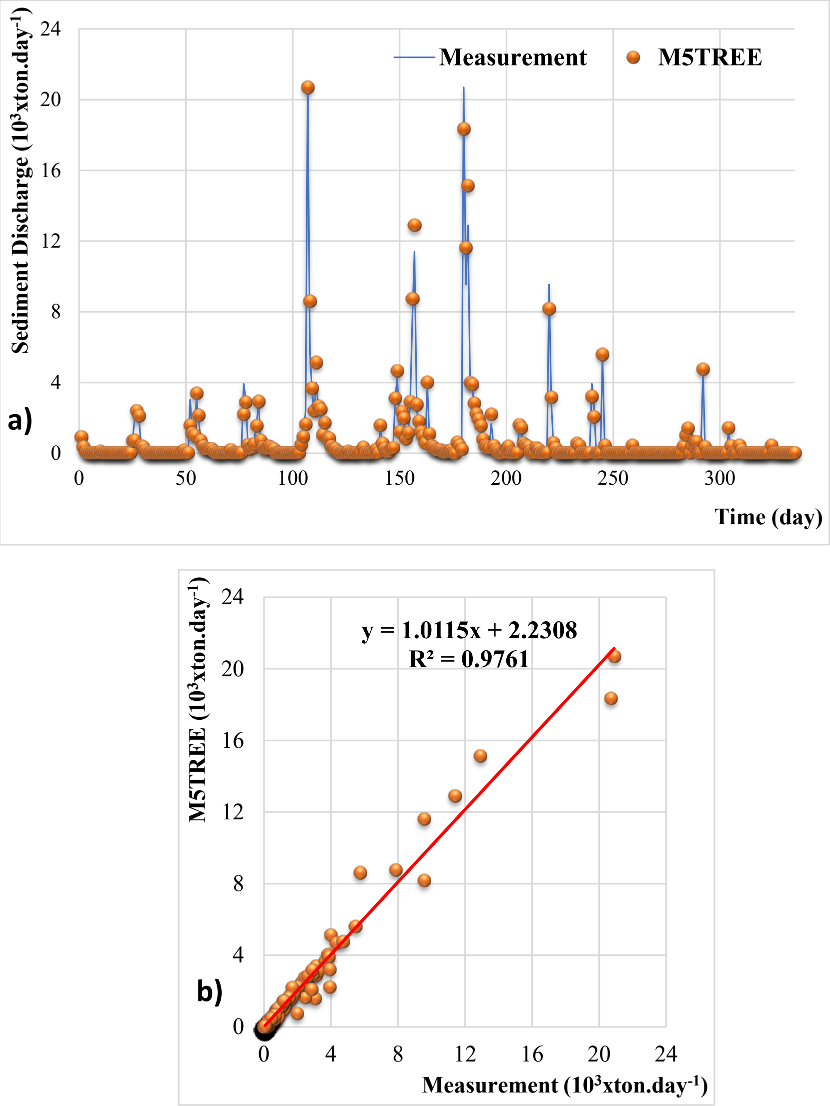


Figure 5. Measurement and M5 Tree for testing data: a) distribution b) scatter graph

Conclusion

In this study, the capabilities of regression analysis (MLR), support vector machines (SVM) and M5 Decision Tree (M5T) have been investigated in estimating sediment discharge. Daily river flow, suspended sediment concentration, sediment discharge, and water temperatures of the Cuyahoga River in the USA were used. While creating the models, 80% of the data set was used as training and 20% as a test.

When the results are evaluated, the M5 Decision Tree model approach has the best results according to the statistical criteria M5T and SVM models have also successful in sediment discharge prediction. MLR model failed compared to other models with the highest MAE, RMSE, and the smallest R^2. The worst results were obtained in the MLR method. MAE, RMSE, and R^2 criteria. In this article, it is demonstrated that M5T and SVM can be a suitable alternative for river sediment estimation in future research.

Acknowledgements

In this paper, hydrological data were taken from the website of the USGS. The authors wish to thank the USGS technical team, also USGS staff members who are involved in measuring and transferring hydrological data.

REFERENCES

1. Baek, S. S., Pyo, J., & Chun, J. A. (2020). Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. *Water*, *12*(12), 3399.
2. Brooks, N.H. (1963). Calculation of suspended load discharge from velocity and concentration parameters. Proceedings of Federal Interagency Sedimentation Conference, U.S. Department of Agriculture, Miscellaneous Publication no. 970.
3. Demirci, M., & Baltaci, A. (2013). Prediction of suspended sediment in river using fuzzy logic and multilinear regression approaches. *Neural Computing and Applications*, *23*(1), 145-151.
4. Demirci, M., Üneş, F., & Saydemir, S. (2015). Suspended sediment estimation using an artificial intelligence approach. In *Sediment matters* (pp. 83-95). Springer, Cham.
5. Einstein, H. A. (1950). *The bed-load function for sediment transportation in open channel flows* (No. 1026). US Government Printing Office.
6. Güzel, H., Üneş, F., Erginer, M., Kaya, Y. Z., Taşar, B., Erginer, İ., & Demirci, M. (2023). A comparative study on daily evapotranspiration estimation by using various artificial intelligence techniques and traditional regression calculations. *Mathematical Biosciences and Engineering*, *20*(6), 11328-11352.
7. Han, H., & Morrison, R. R. (2022). Data-driven approaches for runoff prediction using distributed data. *Stochastic Environmental Research and Risk Assessment*, *36*(8), 2153-2171.
8. Lane, E. W., & Kalinske, A. A. (1941). Engineering calculations of suspended sediment. *Eos, Transactions American Geophysical Union*, *22*(3), 603-607.

9. Melesse, A. M., Ahmad, S., McClain, M. E., Wang, X., & Lim, Y. H. (2011). Suspended sediment load prediction of river systems: An artificial neural network approach. *Agricultural Water Management*, *98*(5), 855-866.
10. Memarian, H., Balasundram, S. K., & Tajbakhsh, M. (2013). An expert integrative approach for sediment load simulation in a tropical watershed. *Journal of Integrative Environmental Sciences*, *10*(3-4), 161-178.
11. Olive, L. J., & Rieger, W. A. (1992). Stream suspended sediment transport monitoring-why, how and what is being measured. *Erosion and Sediment Transport Monitoring Programmes in River Basins*, *210*, 245-254.
12. Ozturk, F., & Apaydın, H. (2001). Suspended Sediment Loads Through Flood Events for Streams of Sakarya River Basin. Turk Journal of Engineering Environmental Science, 25, 643–650.
13. Quinlan JR (1992) Learning with continuous classes. In: Proc. of the Fifth Australian Joint Conference on Artificial Intelligence. World Scientific Singapore 343–348.
14. Sarı, E., Çağatay, M. N., Acar, D., Belivermiş, M., Kılıç, Ö., Arslan, T. N., ... & Sezer, N. (2018). Geochronology and sources of heavy metal pollution in sediments of Istanbul Strait (Bosporus) outlet area, SW Black Sea, Turkey. *Chemosphere*, *205*, 387-395.
15. Taşar, B., Kaya, Y. Z., Varçin, H., Üneş, F., & Demirci, M. (2017). Forecasting of suspended sediment in rivers using artificial neural networks approach. *International Journal of Advanced Engineering Research and Science*, *4*(12), 237333
16. Üneş, F., & Demirci, M. (2015). Generalized regression neural networks for reservoir level modeling. *International Journal of Advanced Computational Engineering and Networking*, *3*, 81-84.
17. Üneş, F., Demirci, M., Zelenakova, M., Çalışıcı, M., Taşar, B., Vranay, F., & Kaya, Y. Z. (2020). River flow estimation using artificial intelligence and fuzzy techniques. *Water*, *12*(9), 2427.
18. USGS.gov | Science for a Changing World [WWW Document]. Available online: https://www.usgs.gov/
19. Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. (New York: Springer-Verlag).
20. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
21. Yadav, A., & Satyannarayana, P. (2020). Multi-objective genetic algorithm optimization of artificial neural network for estimating suspended sediment yield in Mahanadi River basin, India. *International Journal of River Basin Management*, *18*(2), 207-215.
22. Zounemat-Kermani, M., Kişi, Ö., Adamowski, J., & Ramezani-Charmahineh, A. (2016). Evaluation of data driven models for river suspended sediment concentration modeling. *Journal of Hydrology*, *535*, 457-472.