JOURNAL OF WATER AND LAND DEVELOPMENT
e-ISSN 2083-4535
Polish Academy of Sciences (PAN) Institute of Technology and Life Sciences – National Research Institute (ITP – PIB)

JOURNAL OF WATER AND LAND DEVELOPMENT
DOI: 10.24425/jwld.2025.156996
2025, Special Issue: 84–96

# Flood prediction for the Vaal River Basin (South Africa) using machine learning techniques

Adebayo T. Baker*, Megersa O. Dinka, Sophia S. Rwanga

University of Johannesburg, Faculty of Engineering & the Built Environment, Department of Civil Engineering Science, Auckland Park Kingsway Campus, 2092, Johannesburg, South Africa

* Corresponding author

RECEIVED 12.06.2025 ACCEPTED 21.10.2025 AVAILABLE ONLINE 12.12.2025

**Abstract:** Flood prediction is a critical tool for disaster management and risk mitigation. Machine learning models are viable alternatives to the traditional techniques of flood prediction and analysis, which often fail in capturing the complex nonlinear relationship among meteorological parameters. This study evaluated the performance of an artificial neural network (ANN) to predict the flooding indicator (surface volume) in the Vaal River Basin using the key meteorological parameters: historical records of rainfall, wind speed, humidity, and maximum temperature. A 30-year (1994–2024) dataset was collected from the South African Weather Service and preprocessed using standard techniques. Hyper-parameter optimisation of the models was carried out using a grid-search method. The ANN model was developed by testing different topologies, training algorithms and activation functions at both the hidden and output layers. The performance of the models was evaluated using relevant statistical metrics, namely root mean square error (*RMSE*), mean absolute percentage error (*MAPE*), mean absolute error (*MAE*), value accounted for (*VAE*) and *R*-value. The ANN model with tansig-tansig activation function and Levenberg Marquardt training algorithms outperformed other architectures with *RMSE* of 6.245, *MAPE* of 25.95%, *MAE* of 4.656, *VAE* of 7.843 and *R*-value of 0.823 at the training. This research demonstrated the viability of machine learning-based flooding predictions based on weather variables, contributing to flood risk management strategies.

**Keywords:** artificial neural networks, flood prediction, machine learning, root mean square, Vaal River Basin

## INTRODUCTION

The Vaal River Basin (South Africa) is susceptible to flood events, which are exacerbated by climatic variations and anthropogenic activities (Akanbi, Davis and Ndarana, 2020). Along the Vaal River, one of the biggest rivers in South Africa, flooding has a long history. Floods are common in the Vaal River watershed for a number of reasons. Seasonal rainfall occurs in the Vaal River watershed, mostly in the summer (November to March). Climate change has also led to increasingly unpredictable weather patterns and extreme weather events, which have contributed to more frequent and intense flooding events (Chen, Chen and Lin, 2020). The basin, serving as a critical water source for domestic, agricultural, and industrial use, experiences frequent and intense flooding, leading to extensive damage and disruption. Inadequate maintenance, structural defects, or severe weather can cause levees and dams intended to regulate water flow to malfunction, resulting in floods. Flooding is a recurring natural disaster that significantly impacts communities, economies, and ecosystems globally (Mashaly and Fernald, 2020). Floods are among the most devastating natural disasters, causing loss of life, economic damage, and environmental degradation. The Vaal River Basin, spanning multiple provinces in South Africa, plays a critical role in the region's water supply and agriculture. However, its susceptibility to flooding necessitates advanced predictive tools to minimise adverse impacts.

Mohamadi, Ehteram and El-Shafie (2020) reported that effective flood management and mitigation require accurate prediction and modelling of flood occurrences. Traditional hydrologic models, which rely on physical and statistical approaches, have been instrumental in understanding flood dynamics. Machine learning algorithms, in particular, have been

shown to be incredibly adept at understanding these complex linkages and enhancing prediction accuracy in artificial intelligence (AI) models. Use of artificial neural networks (ANNs) has been found to be the preferred machine learning (ML) technique, as these techniques outperform most customary approaches (Kocher and Kumar, 2021). Flood estimation and prediction methods are essential for mitigating risks and managing water resources. These methods rely on hydrologic models and, increasingly, artificial intelligence (AI) tools to simulate and forecast flood events. Hydrologic models simulate the movement and distribution of water in a watershed. Key models include HEC-HMS (Hydrologic Engineering Centre – Hydrologic Modelling System), a widely used model for rainfall-runoff simulation and flood forecasting (Jain, Singh and Seth, 2000), and the storm water management model (SWMM), designed for urban flood modelling, which simulates surface runoff and drainage systems (Farina *et al*., 2023).

Artificial intelligence (AI) systems improve flood prediction by examining vast datasets and spotting intricate patterns (Liu *et al*., 2025). Artificial neural networks (ANNs) are important techniques that are useful for predicting rainfall-runoff and capturing nonlinear interactions (Mishra and Dwivedi, 2025). The long short term memory (LSTM) network can precisely predict floods and is ideal for time-series data, such as river flow (Li, J. *et al*., 2024). Flood prediction and estimation have been improved with the use of AI technologies and hydrologic models. Although hydrologic models offer a tangible foundation, AI tools improve precision and effectiveness, facilitating improved flood risk management.

The Vaal River Basin flood modelling difficulties may thus be fully addressed by fusing the advantages of both methodologies. Using hydrological modelling powered by artificial intelligence can assist handling difficult water management issues with improved precision, efficacy, and efficiency (Mashaly and Fernald, 2020). The term AI refers to a wide variety of computer-related fields that are concentrated on developing intelligent models capable of performing tasks that were previously completed by people (Chen, Chen and Lin, 2020). Few studies have specifically targeted the Vaal River Basin. This research aims to bridge the gap by applying machine learning techniques tailored to the basin's unique hydrological and geographical characteristics. The study aimed to explore the potential of machine learning techniques for flood prediction in the Vaal River Basin, using artificial neural networks machine learning algorithms.

Flood prediction in river basins has become increasingly critical due to climate change and urbanisation impacts on hydrological systems (Bibi and Kara, 2023). The Vaal River Basin, as one of South Africa's most important water resources, faces significant flood risks that require advanced prediction methodologies. This literature review examines the current state of ML applications in flood prediction, with a specific focus on the Vaal River Basin context. The review synthesises research on traditional hydrological modelling approaches, emerging ML techniques, and their integration for enhanced flood forecasting capabilities. Key findings indicate that while traditional statistical methods have been employed in the Vaal River system, there is substantial potential for ML-enhanced prediction systems to improve accuracy and provide cost-effective solutions for flood risk management. Floods represent one of the most destructive natural disasters globally, with complex mathematical expressions governing their physical processes (Mishra *et al*., 2022). The Vaal River Basin, spanning approximately 196,000 km$^2$ and serving as a critical water source for South Africa's economic heartland, experiences periodic flooding events that cause significant socioeconomic impacts (Masindi, 2021). Recent flooding events in 2025 have highlighted the urgent need for improved prediction capabilities, with residents and businesses facing years of recovery from economic damage (Cvetković *et al*., 2024). The advancement of machine learning techniques over the past two decades has contributed significantly to flood prediction systems, offering better performance and cost-effective solutions compared to traditional approaches (Jeba and Chitra, 2024). This literature review examines the application of ML techniques to flood prediction, with particular emphasis on the Vaal River Basin context and the broader South African hydrological environment.

The Vaal River Basin represents South Africa's most economically important catchment, supporting major urban centres including Johannesburg and Pretoria (Remilekun *et al*., 2021). The basin's hydrology is significantly influenced by the Lesotho Highlands Water Project, launched in 1997, which augments water supply through a three-phase construction involving four major dams (Sayedi, 2023). This infrastructure development has altered the natural flow regimes, creating complex hydrological conditions that challenge traditional flood prediction methods.

Statistical analysis of historical flood flows in the Vaal River has revealed critical patterns for flood risk assessment. Mamphwe (2021), identified approximately a 3% annual exceedance probability for major flood events based on historical flow data. However, the continued development of the catchment with urban expansion and infrastructure development has modified these risk profiles, necessitating updated prediction methodologies that can account for non-stationary conditions. Recent flooding events, including the 2025 incidents that affected crops and necessitated house evacuations, demonstrate the ongoing vulnerability of the basin to extreme hydrological events (Boboye and Dorasamy, 2025). The complex interplay between natural variability, climate change impacts, and anthropogenic modifications requires sophisticated modelling approaches that can capture these multi-scale interactions.

Historical approaches to flood prediction in the Vaal River Basin have relied primarily on statistical analysis of flood flows (Baloyi, 2022). These methods utilise frequency analysis, extreme value distributions, and regression techniques to establish relationships between meteorological inputs and flood outcomes. While these approaches provide valuable baseline capabilities, they are limited in their ability to capture non-linear relationships and changing basin conditions. Physical-based hydrological models, such as MIKE-11, have been employed to simulate flood processes through mathematical representation of physical laws governing water movement (Anuruddhika *et al*., 2025). These models require detailed parameter calibration and extensive data inputs, making them computationally intensive and challenging to implement in data-scarce regions.

Machine learning methods have demonstrated significant potential in advancing flood prediction systems through their ability to model complex, non-linear relationships in hydrological data (Kumar *et al*., 2023b). These techniques can process large datasets, identify patterns in multi-dimensional data spaces, and provide probabilistic forecasts that support decision-making processes. The ANNs have been widely applied in flood

prediction due to their ability to approximate complex non-linear functions (Tabbussum and Dar, 2021). The ANNs can process multiple input variables, including precipitation, temperature, soil moisture, and antecedent flow conditions, to predict flood events. Their universal approximation capabilities make them suitable for capturing the complex relationships inherent in hydrological systems.

The LSTM networks represent a significant advancement in time series prediction for hydrological applications (Choi *et al*., 2022). These recurrent neural networks can capture long-term dependencies in sequential data, making them particularly suitable for flood prediction, where antecedent conditions significantly influence current responses. Recent comparative studies have shown LSTM models among the most effective approaches for water level prediction in river systems (Li H. *et al*., 2024).

Random forest (RF) algorithms have demonstrated strong performance in flood susceptibility mapping and prediction tasks. The RF methods can handle high-dimensional datasets, provide feature importance rankings, and offer robust performance across different hydrological conditions (Cappelli *et al*., 2023). Their ensemble nature helps reduce overfitting and provides uncertainty estimates for predictions. Advanced gradient boosting techniques, including LightGBM and CatBoost, have shown promising results in flood risk assessment applications (Xu *et al*., 2023). These methods can capture complex interactions between variables and provide high accuracy in flood susceptibility mapping tasks.

Support vector machines (SVM) offer robust performance in flood prediction through their ability to handle high-dimensional data and provide good generalisation capabilities (Haddad and Rahman, 2020). The SVM methods are particularly effective in scenarios with limited training data, making them suitable for data-scarce regions. The integration of satellite-based observations offers significant potential for enhancing flood prediction capabilities in the Vaal River Basin (Masindi, 2021). Remote sensing products can provide spatially distributed information on precipitation, soil moisture, vegetation conditions, and flood extent, supplementing ground-based observations (Schoener and Stone, 2020).

The transition from research applications to operational flood forecasting systems requires consideration of computational efficiency, data latency, and system reliability (Kumar *et al*., 2023a). The ML models must be capable of processing real-time data streams and providing timely predictions to support emergency response activities. Integration of ML-based flood prediction with early warning systems requires careful attention to communication protocols, stakeholder needs, and decision support tools (Khan *et al*., 2025). The development of user-friendly interfaces and clear communication of prediction uncertainty is essential for effective implementation. Operational ML systems require robust computational infrastructure capable of handling data processing, model execution, and result dissemination (Matthew, Joshua and Philip, 2025). Cloud-based platforms and distributed computing approaches offer scalable solutions for operational flood forecasting.

The "black box" nature of many machine learning algorithms presents challenges for hydrological applications where process understanding is important (Lange and Sippel, 2020). Explainable artificial intelligence techniques and hybrid approaches that combine ML with physical understanding are needed to address these concerns. Machine learning models trained on specific basins or time periods may have limited transferability to different conditions (Ma *et al*., 2024). Transfer learning approaches and domain adaptation techniques offer potential solutions for improving model generalisation.

The application of machine learning techniques to flood prediction in the Vaal River Basin represents a significant opportunity to enhance current forecasting capabilities and improve flood risk management (Antwi-Agyakwa, Afenyo and Angnuureng, 2023). While traditional statistical approaches have provided valuable baseline capabilities, the complex, non-linear nature of hydrological processes in the basin requires more sophisticated modelling approaches. Current research demonstrates that machine learning techniques, particularly long short-term memory networks, random forest algorithms, and hybrid approaches, offer substantial improvements in prediction accuracy and computational efficiency (Sun *et al*., 2021). However, successful implementation requires careful attention to data quality, model validation, and operational considerations. The unique characteristics of the Vaal River Basin, including its economic importance, complex infrastructure, and transboundary components, present both challenges and opportunities for machine learning applications. Recent flooding events have highlighted the urgent need for improved prediction capabilities, creating a compelling case for investment in machine learning-based forecasting systems (Kumar *et al*., 2023a).

This section explores the utilisation of AI techniques in flood modelling within the Vaal River Basin. Kumar *et al*. (2023a) reported that recent floods in several parts of southern India caused significant harm to both persons and property. South Africa's Vaal River Basin has exceptional flood risks because of its distinct hydrological and climatic features. Accurate flood forecasting models are necessary for efficient flood mitigation and management. The results showed that the projected model provided high precision in projecting flood flow and successfully assisted in building a rainfall–runoff model. In order to anticipate the daily erratic stream flow of Thrace County, which is located in northwest Turkey, Sharma and Srivastava (2021) used an artificial neural network (ANN), an adaptive-network-based fuzzy inference system (ANFIS), and an SVM. They also compared the results with those of local linear regression (LLR) and dynamic log-likelihood ratio (DLLR). Results indicated that when estimating daily sporadic stream flow, ANN, ANFIS, and SVM performed better than the LLR and DLLR models.

When it came to forecasting the dam water level, SVM performed better with various input combinations based on performance evaluation factors. In their 2001 study, Thakur and Konde (2001) illustrated several aspects of flood forecasting, including the usage of models that had already been utilised, the development of input gathering techniques, and the display of results, uncertainties, and flood warnings. The goal of this study is to investigate how well an ANN model can forecast flood discharge.

Conventional approaches, including hydrological models and statistical methods, have been widely used for flood prediction. These models rely on physical and empirical relationships among hydrological parameters, but often struggle to capture non-linear dependencies. Recent advances in machine learning have introduced data-driven methods that excel in handling large datasets and uncovering complex patterns. Techniques such as ANN, SVM, and random forests have been applied to flood prediction with promising results.

## MATERIALS AND METHODS

### STUDY AREA

The interior of South Africa is home to the economically important Vaal River Basin, where there is a high concentration of mining, industrial, residential, and agricultural activity. It is the largest tributary of the Orange River in South Africa. The river has its source near Breyten in Mpumalanga province, east of Johannesburg and about 30 km north of Ermelo and only 241 km from the Indian Ocean. It then flows westwards to its conjunction with the Orange River, southwest of Kimberly in the North Cape. It is 1,458 km long, and forms the border between Mpumalanga, Gauteng and North West provinces on its north bank, and Free State in its South. The Vaal River Basin's geographical coordinates vary significantly along its length, as it is a large basin. The source of the Vaal River is near Breyten in Mpumalanga, approximately at 26°17'59.7"S 29°09'12.7"E. The Vaal River then flows southwest, eventually meeting the Orange River near Douglas, with coordinates roughly 29°4'15"S 23°38'10"E. The map of the Vaal River Basin with different land, as indicated by Masindi and Abiye (2018) in their study, is shown in Figure 1. Vaal River has the following characteristic according to Akpotu (2021): discharge 125 $m^3 \cdot s^{-1}$. It has its source from the Drakensberg in the city of Johannesburg, Kimberly, with a basic size of 196,438 $km^2$ and etymology of 1 Hai "pale" + 1 Arib "river". Water supplies, agriculture, and industry all depend on the Vaal River Basin, one of the most important river systems in South Africa. It covers an area of around 196,438 $km^2$ and crosses multiple provinces, including Gauteng, the Free State, Northwest, and Mpumalanga. The Vaal River Basin, located in South Africa, exhibits a semi-arid to subtropical climate, with significant regional variations due to differences in altitude, topography, and latitude. Average climate characteristics with a temperature during summer (November–February) 20–30°C (can exceed 35°C in low-lying areas), winter (June–August) 5–18°C (frost occurs in higher elevations) and rainfall of 400–800 $mm \cdot y^{-1}$ (higher in eastern highveld, lower in western regions). Summer-dominated rainfall (October–April), with thunderstorms common, evaporation is high, often exceeding rainfall (especially in the west).

### DATA COLLECTION AND PROCESSING

Flood prediction is critical for disaster mitigation in the Vaal River Basin, where variable rainfall, land-use changes, and increasing urbanisation exacerbate flood risks. Surface volume with other weather parameters (wind speed, humidity, maximum temperature) was used as a flood indicator for the flood prediction in the study area. Traditional hydrological models often struggle with real-time adaptability, making machine learning (ML) a promising alternative due to its ability to process complex, nonlinear relationships in environmental data (Jeba and Chitra, 2024). In this study, 30-year (1994–2024) data were sourced from the South African Weather Service in the period from 1994 to 2024 for meteorological parameters such as wind speed, humidity, maximum temperature and historical records of rainfall. These meteorological variables were chosen due to their potential influence on flooding dynamics. Additionally, historical flooding indicator (surface volume) data spanning from 1994 to 2013 were

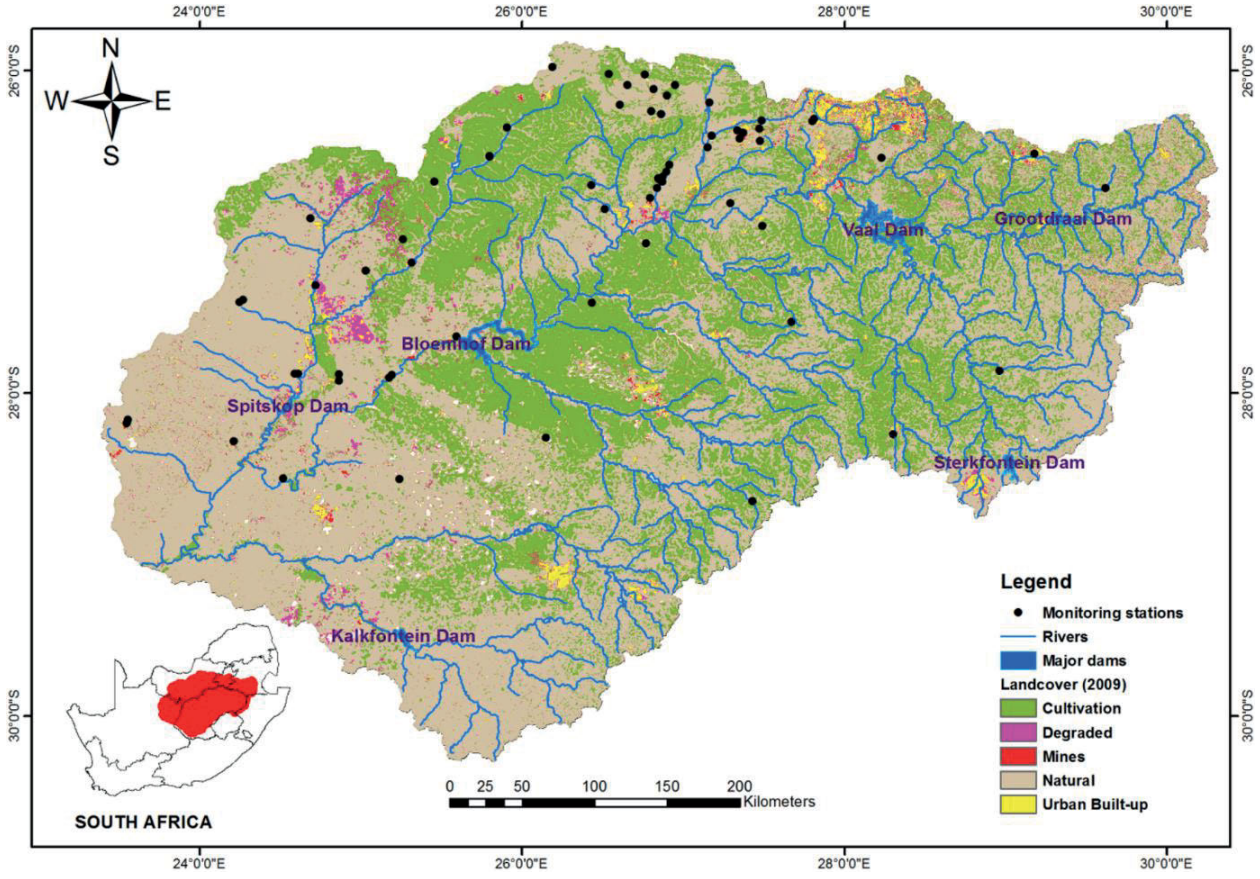


**Fig. 1.** Map showing the Vaal River Basin with different land uses; source: Masindi and Abiye (2018)

collected to complement the meteorological dataset. Surface volume serves as a fundamental physical control on flood generation and propagation in the Vaal River Basin (Baloyi, 2022). Its influence operates through multiple mechanisms, including temporary storage, hydraulic routing, and antecedent condition effects. Successful flood prediction requires a comprehensive understanding of these processes. The incorporation of surface volume as a primary parameter in flood prediction models for the Vaal River Basin is scientifically justified through multiple theoretical, empirical, and practical considerations (Funke, 2025). This justification stems from the fundamental role surface storage plays in hydrological processes, the unique characteristics of the Vaal River Basin, and the demonstrated improvement in predictive accuracy when surface volume is explicitly considered.

In flood prediction, artificial neural networks (ANNs) serve as a predictive model rather than a direct "performance indicator". However, their predictive accuracy (e.g., root mean square error (*RMSE*), Nash–Sutcliffe efficiency (*NSE*)) can be used as a key performance indicator (*KPI*) to evaluate flood forecasting systems. The ANNs are trained to map input parameters (e.g., precipitation, upstream flow, land use) to output indicators (e.g., flood occurrence, water level, inundation extent). The ANNs provide a robust KPI-driven framework for flood prediction in the Vaal River Basin, with quantifiable accuracy metrics guiding emergency response. The disparity in timeframes between the datasets was addressed through data alignment and preprocessing techniques to ensure compatibility and coherence for model development. This comprehensive dataset forms the foundation for the predictive modelling of flooding occurrence in the Vaal River Basin using meteorological parameters. In Table 1, the statistical summary and properties of the relevant variables and parameters are represented.

A correlation heatmap in Figure 2 was generated to examine the relationships between meteorological parameters and flooding parameters (surface volume). This visualisation illustrates the linear relationships among wind speed, humidity, maximum temperature, and flooding parameter (surface volume). The heatmap illustrates the degree and direction of correlations, facilitating the identification of model predictors.

## DATA PREPROCESSING

The raw data set comprising several years of multiple variables was extracted and prepared in a format ready for the model. To achieve an accurate model, the following preprocessing steps were carried out on the data.

**Table 1.** Statistical features of the weather parameters and the flooding parameter (surface volume)

| Parameters / statistical properties | Wind speed ($m \cdot s^{-1}$) | Humidity ($g \cdot m^{-3}$) | Maximum temperature (°C) | Surface volume ($m^3$) |
|---|---|---|---|---|
| Maximum | 11.500 | 100.00 | 36.0 | 3.08 |
| Minimum | 0.000 | 0.00 | 2.40 | – |
| Mean | 2.670 | 38.34 | 3.11 | – |
| Standard deviation | 1.969 | 19.215 | 5.165 | – |

Source: own elaboration.

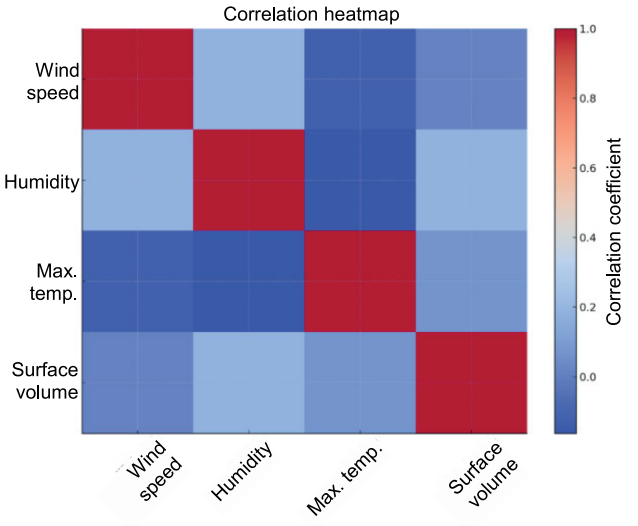


**Fig. 2.** Correlation heatmap for the model data; source: own elaboration

### Outlier removal

The dataset comprised about 30 years (1994–2024) of historical records of meteorological parameters and environmental conditions. Owing to the many years of record, many variations exist across different seasons, years, and environmental conditions, especially between the years 1996 and 1998. Over such a long period, unusual events or measurement errors may introduce outliers – data points that do not reflect typical patterns or behaviours. Over several years, the variables concerned may involve anomalous readings, possibly due to recording mistakes or extreme events. A great variation was particularly noted in the values of flooding indices (surface volume), which is the range of 8–4000 $m^3$, making the outlier critical for the machine learning model. This disparity could be attributed to factors such as floods or equipment malfunctions, which may misrepresent the true relationship between the weather parameters and flooding occurrence.

In this study, two statistical methods of outlier removal were combined logically with the "OR" operator in the MATLAB ("isoutlier" function) environment to reduce the noise in the data and achieve a reliable model. These methods are as follows.

- **Interquartile range (*IQR*) based method.** This approach uses the *IQR* between the 75$^{th}$ percentile ($Q3$) and the 25$^{th}$ percentile ($Q1$). The values in the data which do not fall within a predefined threshold are described as outliers.

$$IQR = Q3 - Q1 \tag{1}$$

The outlier threshold is defined as follows.

$$\text{Outlier}\ (x) = \begin{cases} \text{true, if } x < Q1 - 1.5\ IQR \\ \text{true, if } x > Q1 + 1.5\ IQR \\ \text{false, otherwise} \end{cases} \tag{2}$$

where: *IQR* = interquartile range.

- **Median method.** This method defines the threshold as a multiple of the mean absolute deviation as follows.

$$\text{Outlier}\ (x) = \begin{cases} \text{true, if } |x - \text{median}(X)| > k \cdot MAD \\ \text{false, otherwise} \end{cases} \tag{3}$$

where: *MAD* = mean absolute deviation.

### *Z*-score normalisation

This is another important preprocessing step for scaling the features in the range of zero (0) mean and one (1) standard deviation. This can be achieved using the following equation.

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{4}$$

where: $X_{\text{scaled}}$ = the scaled (standardised) value, $X$ = the original data point or value, $\mu$ = the mean (average) of the dataset, $\sigma$ = the standard deviation.

The target output normalisation is carried out as follows.

$$y_{\text{scaled}} = \frac{y - \text{mean}(y)}{\sigma} \tag{5}$$

where: $y_{\text{scaled}}$ = the scaled (standardised) version of $y$, $y$ = the original data variable or vector, mean($y$) = the arithmetic mean (average) of all values in $y$.

## MODEL DEVELOPMENT

### Artificial neural network

The ANN is an example of a non-linear prediction (NLP) method, which has been extensively studied and applied to a variety of problems, including meteorological simulation and forecasting (Waqas *et al*., 2023). Nourani, Paknezhad and Tanaka (2021) conducted a study on prediction interval estimation methods for artificial neural networks-based modelling of hydro-climate processes, a review. The use of artificial neural networks is a popular data-driven technique that has been frequently applied to a broad range of fields (Ma *et al*., 2023). An artificial neural network is able to handle non-linearity and automatically adjusts to new information, while generally requiring little computational effort (Jamsheed and Iqbal, 2023). The behaviour of a neural network is defined by the way its individual computing elements are connected and by the strength of those connections. These weighted connections are automatically adjusted during training of the network. Artificial neural networks with one hidden layer are commonly used in modelling since it has been found that more than one hidden layer does not yield any significant improvement in performance on a network with a single hidden layer (Uzair and Jamil, 2020).

### Artificial neural network development

The artificial neural network model was inspired by the biological nervous system and has allowed scientists and researchers to build mathematical models of neurons in order to simulate neural behaviour (Thakur and Konde, 2021). Models of a neuron were introduced in the early 1940s by McCulloch and Pitts by which they described simple logic for neural networks, and were later credited with a learning law, the perceptron learning algorithm (Sharma and Srivastava, 2021). The research on the limits to what one-layer perceptron can compute was demonstrated by Minsky and Pappert with the use of elegant mathematics (Worden *et al*., 2023). The back-propagation algorithm developed by McClelland and Rumelhart, is the most popular learning algorithm for the training of multilayer perceptron (Zhang *et al*., 2007).

The ANNs were first introduced to water resources research for their use to predict monthly water consumption and to estimate occurrences of floods. Since then, ANNs have been used for a number of different water resource applications, which include time-series prediction for rainfall forecasting, rainfall-runoff processes and river salinity. The ANNs have also been used for modelling soil and water table fluctuations, pesticide movement in soils, water table management and water quality management (Omeka *et al*., 2024). The ANN contains a large number of simple neuron-like processing elements and a large number of weighted connections between the elements. The weights of connections encode the knowledge embedded in the network. The "intelligence" of a neural network emerges from the collective behaviour of neurons, each of which performs only very limited operations. Each individual neuron finds a solution by working in parallel (Ha and Tang, 2022). In Figure 3, a flowchart for data preprocessing and model development for the study is shown.

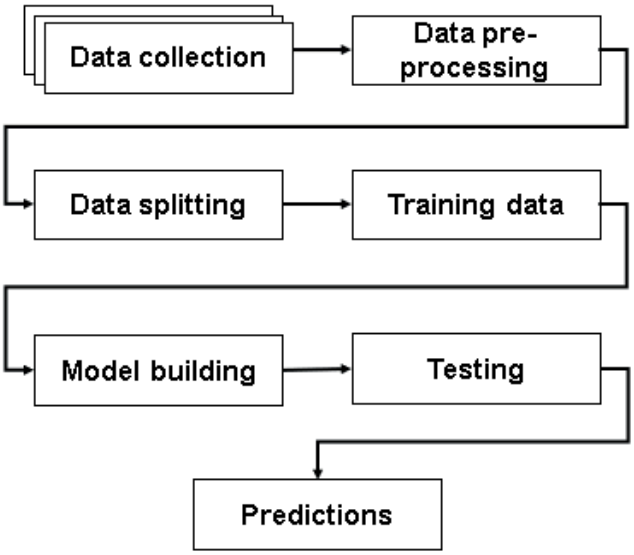


**Fig. 3.** Flowchart for data preprocessing and model development; source: own elaboration

### Model training, validation and testing

Raw data from different resources, along with datasets, were collected and preprocessed to check, clean and organise for analysis. This step involved handling missing values and converting data into a usable format. Generally, the data were split into two parts – one to train the model and the other to the trained model. After the above steps, the model is built, and different algorithms are used to check for accuracies, and depending upon those accuracies, the higher accuracy algorithm is selected for the final model. Prediction is nothing but applying a trained model to new or unseen data to generate predictions. Here, in the case of flood forecasting, it means predicting rainfall and analysing whether there is a flood or not.

## HYPER-PARAMETER OPTIMISATION

We have used a grid search method to obtain the optimal combination of hyperparameters, such as the hidden layer configuration and the learning rate for a good performance of the ANN model. In the grid search approach, the sets of predefined hyperparameters were tested systematically, while the

best combination was selected based on the lowest error threshold. This selected network is also used for the final model training and testing. In this study, our tuning search space involved six combinations as follows.

$$Ss = (Hl : 3) \cdot (Lr : 2) = C \quad (6)$$

where: *Ss* = search space of the hyper-parameter optimisation, *Hl* = hidden layer: 3 of the hyper-parameter optimisation, *Lr* = learning rate: 2 of the hyper-parameter optimisation, *C* = combination of the hyper-parameter optimisation.

The combination of hidden layer = [23 33] [24 53] [35 53] was tested while learning rates were adjusted between 0.1 and 0.00 according to Ibrahim *et al.* (2025) as provided in the neural computing and applications publishing model. At each hidden layer architecture, different combinations of activation functions at the hidden and output layers with varying combinations of training algorithms were tested. Other specified model hyper-parameters are provided in Table 2 as suggested by Ibrahim *et al.* (2025).

**Table 2.** Hyperparameter settings of the artificial neural network architecture

| Hyper-parameter | Value |
|---|---|
| Epochs | 500 |
| Minimum gradient | $1{\cdot}10^{-6}$ |
| Data splitting | 70:15:15 |
| Regularisation terms ($\lambda$) | 0.1 |
| Transfer function | · hyperbolic tangent sigmoid (tansig)<br>· logarithmic sigmoid (logsig)<br>· pure linear (purelin)<br>· short maximum (softmax) |
| Training algorithm | · Levenberg–Marquardt backpropagation (trainlm)<br>· scaled conjugate gradient (trainscg)<br>· Bayesian regularisation (trainbr) |

Source: own elaboration based on Ibrahim *et al.* (2025).

## PERFORMANCE EVALUATION

Relevant statistical metrics such as root mean square error (*RMSE*), mean absolute error (*MAE*), mean absolute percentage error (*MAPE*), and variance accounted for (*VAF*) have been selected to assess the performance of the developed ANN model for predicting flooding based on inputs such as wind speed, humidity, and maximum temperature. These metrics were computed using Equations 7–10.

$$MAPE = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{y_k - \widehat{y_k}}{y_k}\right|100\% \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}[y_k - \widehat{y_k}]}{N}} \quad (8)$$

$$MAE = \frac{\sum_{k=1}^{N}|\widehat{y_k} - y_k|}{N} \quad (9)$$

$$VAF = 1 - \left[\frac{\mathrm{var}(\widehat{y_k} - y_k)}{\mathrm{var}(y_k)}\right]100 \quad (10)$$

where: $N$ = the number of data points (observation), $k$ = predicted (forecasted) value at time period, $y_k$ = total number of time period, $\widehat{y_k}$ = the mean of the total number of time period, var = variations of data points.

This study focuses more on novel model design rather than performance comparison. In Table 3, the literature-based comparative performance table for flood prediction methods is shown.

# RESULTS AND DISCUSSION

## ARTIFICIAL NEURAL NETWORK MODEL RESULTS: MODEL PERFORMANCE AND EVALUATION

In Table 4, the statistical metrics results of artificial neural networks at the training and testing phase of different hyperparameter combinations at an optimal hidden layer

**Table 3.** Machine learning methods performance in flood prediction

| Method | Study location / basin | *NSE* | *RMSE* | $R^2$/*R* | Reference |
|---|---|---|---|---|---|
| ANN | Upper Baro watershed, Ethiopia | 0.98 | 24 $m^3{\cdot}s^{-1}$ | 0.99 | Belina, Kassa and Masinde (2025) – integrating machine learning |
| SVM | Upper Baro watershed, Ethiopia | 0.97 | 27 $m^3{\cdot}s^{-1}$ | 0.98 | Belina, Kebede and Masinde (2024) – comparative analysis |
| PSO-SVM hybrid | Barak River Basin | 0.99334 | 0.04962 | 0.98918 | Samantaray, Sahoo and Agnihotri (2023) – prediction of flood discharge |
| SVR | Morocco (data-scarce basin) | 0.72–0.85 | 25–40 $m^3{\cdot}s^{-1}$ | 0.80–0.88 | Bargam *et al.* (2024) – evaluation of SVR and RF |
| RF | large-scale flood simulation | 0.65–0.82 | 28–52 $m^3{\cdot}s^{-1}$ | 0.75–0.87 | Sasanapuri, Dhanya and Gosain (2025) – a surrogate ML model using RF random forest evaluation |

Explanations: ANN = artificial neural network, SVM = support vector machine, SVR = support vector regression, RF = random forest, PSO = particle swarm optimisation, *RMSE* = root mean square error, *NSE* = Nash–Sutcliffe efficiency, *R* = Pearson correlation coefficient, $R^2$ = coefficient of determination.
Source: own elaboration.

architecture of [23, 38] are represented. The choice of 23 and 36 neurons in the two hidden layers appears optimal, as the best performance metrics align with this configuration in sub-model 13. No particular trend was observed in the performance metrics across all the sub-models with different hyper-parameter combinations. During training, the lowest error values, as indicated by a lowest root mean square error (*RMSE*) value, are achieved by the model using the tansig transfer function at the hidden layer, tansig at the output layer, and Levenberg–Marquardt training algorithm. The *RMSE* value of 6.245 indicates better performance in minimising errors. Further indication of the model's low error prediction is depicted in its mean absolute percentage error (*MAPE*) and mean absolute error (*MAE*), values of 25.957 and 4.656, respectively, establishing its superior accuracy. Further to this, the model delivers the highest variance accounted for (*VAF*) and *R* values of 7.843 and 0.832, indicating a stronger correlation between the actual and predicted values of the flooding parameter (surface volume). The above result was in support of the findings of Bergstra and Bengio (2012) from the random search for hyper-parameter optimisation. The *R*-value reaches its peak at 0.832 for sub-model 13, indicating the strongest linear relationship between predictions and actual values. Other models with lower *R*-values (<0.75) struggle to capture the relationships accurately, particularly those with purelin. Sub-model 13 records the lowest *MAE* (4.656), followed by other models employing trainlm. Networks with tansig and logsig transfer functions exhibit moderate *MAE* values compared to those with purelin. The above agrees with the result of Netto *et al*. (2021), where they emphasised the importance of selecting appropriate architecture, transfer functions, and algorithms to achieve such results.

During testing, a trend similar to the training phase was observed in the testing phase, but with a slightly less accurate prediction owing to the slightly higher error values compared to the training phase. The results indicate a marginal decline in performance for the identical combination (tansig-tansig with trainlm) during the training phase, with *RMSE* rising to 8.174, *MAPE* to 49.758, and *MAE* to 6.986. The pipeline automatically handles the preprocessing, training and evaluation while providing a detailed performance breakdown for different data segments. This gives concrete evidence of how well the model performs in extreme values versus the normal range. As an alternative to the high *MAPE* linked outliers and model over-generalisation, data augmentation is used in conjunction with the synthetic minority oversampling technique (SMOTE) regression, which generates synthetic samples between extreme samples and neighbours. However, the *VAF* and *R*-values of 44.407 and 0.7205 suggest acceptable generalisation. Other combinations exhibit a mix of performance, but none outperform the tansig-tansig with trainlm combination across testing metrics. Although the testing *MAPE* (49.758) and *MAE* (6.986) exceed those in training, these figures remain competitive, demonstrating the model's capacity to generalise learnt patterns to novel data. The *R*-values range from 0.6897 to 0.7205 at sub-models 1 and 13, respectively, reflecting varying degrees of correlation between predictions and actual values. Also, based on *R*-values, models with tansig at both hidden and output layers showcase better performance, indicating stronger predictive relationships. While trainlm contributes to achieving better *R*-values, trainscg struggles to match this performance. Models utilising trainlm training methods frequently get superior *VAF* values, highlighting their capacity for generalisation. The

**Table 4.** Statistical metrics result of the artificial neural network at the training and testing phase

| Transfer function | | Training algorithm | Training performance metrics | | | | | Testing performance metrics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| hidden layer | output layer | | *RMSE* | *MAPE* | *MAE* | *VAF* | *R* | *RMSE* | *MAPE* | *MAE* | *VAF* | *R* |
| softmax | logsig | trainbr | 7.153 | 31.649 | 5.023 | 12.543 | 0.763 | 9.758 | 49.657 | 8.112 | 47.116 | 0.6897 |
| tansig | purelin | trainbr | 6.905 | 25.162 | 4.263 | 9.812 | 0.696 | 9.054 | 50.076 | 7.856 | 47.913 | 0.7035 |
| tansig | tansig | trainbr | 7.624 | 32.688 | 5.689 | 10.325 | 0.734 | 9.382 | 51.346 | 7.168 | 48.521 | 0.7176 |
| logsig | logsig | trainscg | 8.167 | 30.904 | 5.262 | 8.919 | 0.689 | 8.811 | 49.045 | 8.635 | 46.705 | 0.7013 |
| logsig | logsig | trainlm | 8.455 | 28.053 | 6.128 | 11.012 | 0.745 | 9.848 | 53.023 | 7.812 | 45.914 | 0.6326 |
| logsig | tansig | trainscg | 7.725 | 27.554 | 5.605 | 8.362 | 0.823 | 9.632 | 49.756 | 8.065 | 45.513 | 0.6721 |
| tansig | purelin | trainscg | 6.854 | 29.562 | 4.521 | 9.065 | 0.804 | 8.906 | 50.325 | 7.164 | 44.906 | 0.6875 |
| tansig | purelin | trainlm | 7.611 | 30.736 | 5.824 | 7.993 | 0.758 | 9.908 | 52.316 | 7.357 | 47.345 | 0.7137 |
| purelin | softmax | trainbr | 7.279 | 33.689 | 6.438 | 8.124 | 0.816 | 9.975 | 51.045 | 8.996 | 46.052 | 0.7032 |
| tansig | logsig | trainscg | 6.457 | 35.175 | 5.066 | 9.445 | 0.768 | 9.561 | 50.987 | 7.289 | 45.844 | 0.7113 |
| softmax | tansig | trainbr | 7.418 | 27.997 | 5.257 | 10.843 | 0.805 | 9.935 | 50.765 | 8.065 | 46.147 | 0.6975 |
| purelin | logsig | trainlm | 7.783 | 28.183 | 5.209 | 7.904 | 0.762 | 8.875 | 51.564 | 7.623 | 47.182 | 0.7044 |
| tansig | tansig | trainlm | 6.245 | 25.957 | 4.656 | 7.843 | 0.832 | 8.174 | 49.758 | 6.986 | 44.407 | 0.7205 |
| softmax | softmax | trainlm | 7.521 | 27.883 | 4.978 | 9.543 | 0.811 | 9.355 | 49.765 | 7.858 | 47.746 | 0.7234 |
| softmax | tansig | trainscg | 6.972 | 27.533 | 5.107 | 12.205 | 0.751 | 9.643 | 50.726 | 8.345 | 44.986 | 0.6864 |
| logsig | tansig | trainbr | 7.8155 | 25.983 | 5.723 | 11.235 | 0.787 | 9.246 | 51.353 | 7.618 | 45.579 | 0.6742 |

Explanations: *MAPE* = mean absolute percentage error, *MAE* = mean absolute error, *VAF* = variance accounted for, *RMSE*, *R* as in Tab. 3.
Source: own study.

*MAE* ranges from 6.986 to 8.635, signifying variations in absolute predictive accuracy. Models trained with trainlm often get a lower *MAE* than those trained with trainbr or trainscg.

The flood categorised boundary conditions in artificial neural network-based flood prediction in the Vaal River Basin used hydrological thresholds with the discharge rates 2.801 $m^3 \cdot s^{-1}$ based on return periods of 10-year floods. Temporal boundaries with a lead time of 72 hours for operational flood prediction and data input periods using historical data spanning 30 years were used for training. Seasonal variations with wet season (October–March) against dry season patterns were used. Spatial boundaries of the upstream catchment limit with major tributaries and extend downstream until the confluence with the Orange River. These threshold values were calibrated using historical flood records and validated against observed flood events in the Vaal River Basin to categorise when the river basin is flooded and not flooded.

## LINEARITY BETWEEN EXPERIMENT AND PREDICTED DATA

In addition to the evaluation of statistical metrics, performance during the training and testing phases is illustrated by a scatter plot comparing the predicted and actual values of the flooding parameter (surface volume), as shown in Figure 4. The scatter plot demonstrates that the artificial neural network (ANN) model has learned the overall trends and patterns in the training data with a strong *R*-value of 0.832. The model performs well in capturing the central trend and variability of the training data, as evidenced by the strong *R*-value and the alignment of many data points along the fit line. The scatter plot suggests the ANN model is moderately effective in predicting flooding occurrence based on the meteorological variables. The data points show a positive correlation with the experimental values, but are scattered around the fit line. For lower experimental values (<20), the predictions align closely with the actual values. As the experimental values increase (>30), the scatter becomes more pronounced, and the model shows some deviation from the ideal relationship. The data points exhibit significant dispersion around the fit line, which denotes the optimal relationship. This indicates that although the ANN model forecasts trends, there exists variability and a certain level of inaccuracy in specific predictions.

Beyond the statistical metrics evaluation, the testing phase performance is demonstrated using a scatter plot of the predicted and actual values of the flooding parameter (surface volume). The scatter plot suggests the ANN model is moderately effective in predicting flooding based on the weather variables. The *R*-value of 0.720 measures the magnitude of the linear relationship and exhibits a moderate positive correlation between the actual and predicted surface volume values.

## ANALYSIS OF RESIDUAL

In Figure 5a, the testing comparison plot of the actual and predicted values of the flooding parameter (surface volume) is represented. The predicted values are concentrated around the central range of the experimental values (approximately 15–25). The experimental values show a wide range from approximately 5 to over 50. The predicted values, however, are compressed into a narrower band, indicating that the ANN model does not fully capture the variability in the training data. For extreme experimental values (e.g., >30 or <10), the model does not adequately predict the corresponding values, suggesting that the ANN might have overgeneralised during training. The predicted values agree with the overall trend of the actual values, although they display discrepancies at some points. This could be attributed to overestimation and underestimation of some model hyperparameters (Adeleke and Jen, 2022).

In Figure 5d, the testing comparison plot of the actual and predicted values of flooding indices (surface volume) is represented. In Figure 5d, the actual surface volume values exhibit a significant variability, reflecting the inherent complexity of the dataset. The predicted values agree with the overall trend of the actual values, although they display discrepancies at some points. Moreover, substantial disparities between actual and predicted flooding indices (surface volume) are evident in the outliers when surface volume diverges substantially from the mean. For most of the samples, the predictions are adequately consistent with the actual values, indicating that the model effectively reflects the underlying patterns of the dataset. In addition to the comparison of actual and predicted flooding indices (surface volume), the histogram has a relatively symmetric distribution centred at zero,

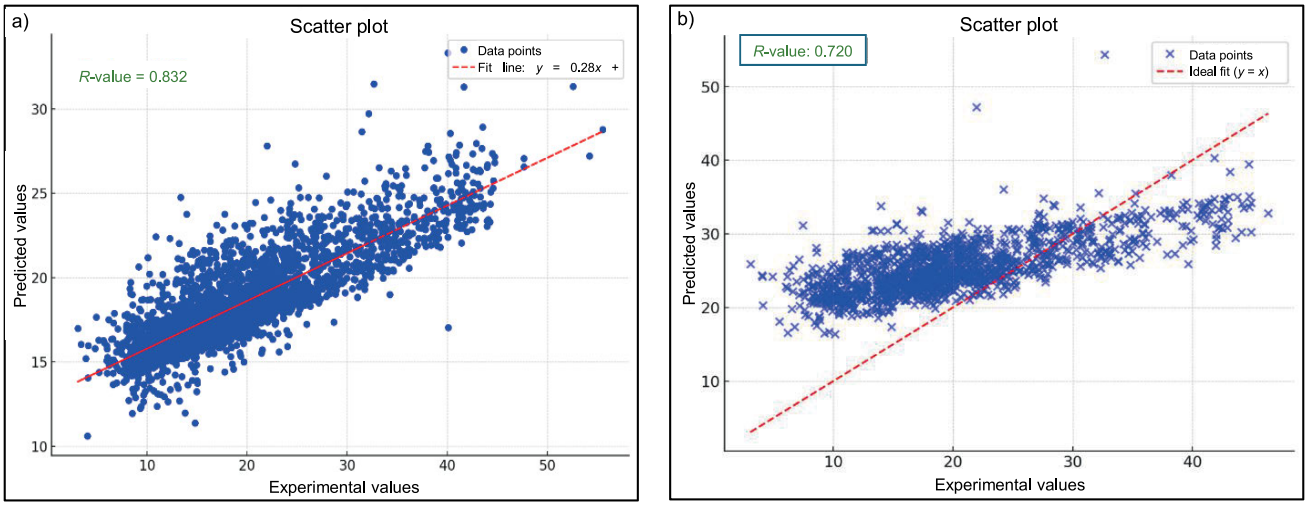


**Fig. 4.** Scattered plot of actual and artificial neural network predicted flooding indicator (surface volume) at the: a) training, b) testing; source: own study

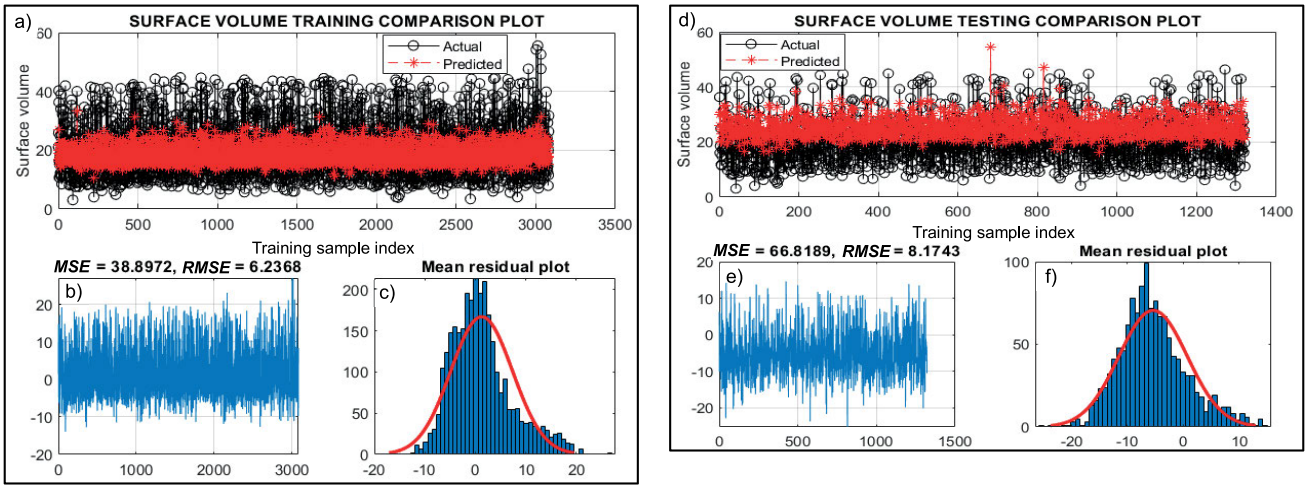


**Fig. 5.** Artificial neural network testing comparison plot and error diagram of the actual and predicted flooding indicator (surface volume) during: a), b), c) training; d), e), f) testing phases; *MSE* = mean square error, *RMSE* = root mean square error; source: own study

suggesting the model lacks substantial systematic bias (e.g., continuous overprediction or underprediction) (Figs. 5c and 5f). Nevertheless, the residuals exhibit a little dispersion, signifying variability in prediction errors. The *MSE* and *RMSE* in Figures 5b and 5e provide a more interpretable metric in the same units as the target variable. The histogram has a very symmetric distribution centred around zero, indicating that the model does not include significant systematic bias (e.g., persistent overprediction or underprediction). Nonetheless, the residuals display some dispersion, indicating diversity in prediction mistakes. It presents the *MSE* and *RMSE*, offering more comprehensible statistics in the same units as the target variable. The close correlation between actual and expected values in the mid-range indicates high predictive accuracy for standard data points.

The histogram of the residual plot in Figure 6 gives insight into the performance of the model during training. In Figure 6a, the scattered plot of the actual and predicted flooding indicator is shown. The residuals exhibit a symmetrical distribution centred around 0, with the greatest concentration of residuals occurring close to the zero residual line. This suggests that the ANN model does not demonstrate substantial consistent bias in overestimating or underestimating values during the training phase. The predominant residuals fall within the interval from −10 to 10, with a pronounced peak at 0. This indicates that most of the predictions are near the actual values, showcasing commendable accuracy throughout training. The residuals range from around −12 on the left to +22 on the right, exhibiting lower frequencies at the extremes. These outliers signify instances where the ANN failed to precisely forecast the flooding parameter (surface volume). The maximum frequency is observed at residuals around zero, with more than 500 samples exhibiting little prediction errors. This verifies that the model effectively identifies the overarching patterns of the training data.

To further understand the predictive behaviour of the ANN at the testing phase, the histogram of residuals in Figure 6B is critical. The histogram has considerable symmetry, suggesting that the residual errors are uniformly distributed around zero. This symmetry indicates that the model does not exhibit a bias towards overprediction or underprediction. The distribution of residuals indicates a certain level of variance in the errors, suggesting opportunities for enhancing the model's generalisation capacity. The majority of residuals fall within a range from −10 to +10, signifying that the model's predictions are often near the actual values. The peak of the histogram is near zero residuals,

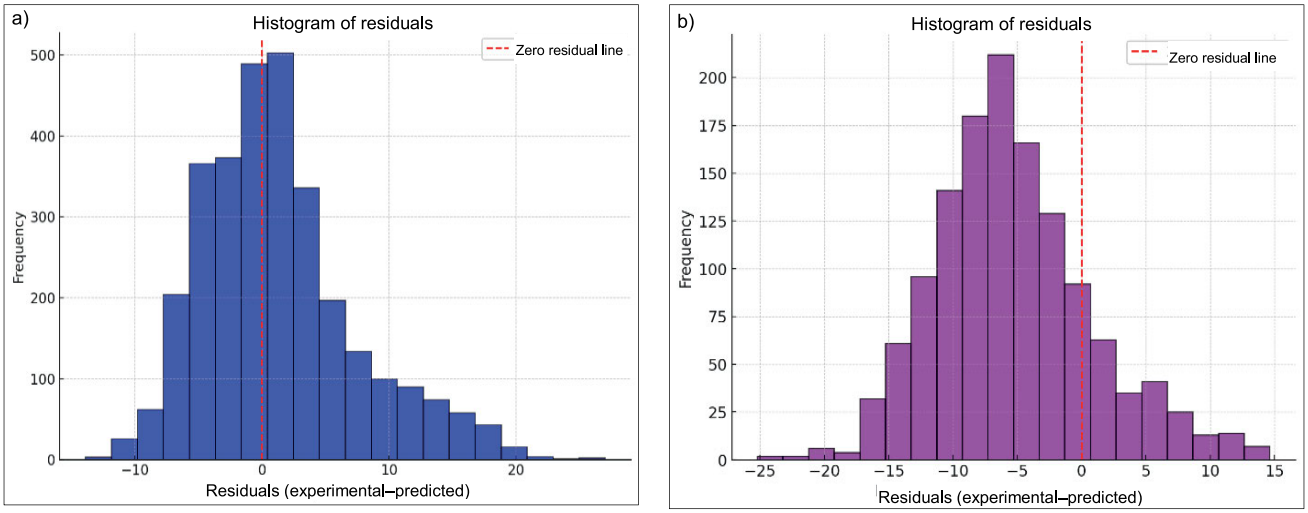


**Fig. 6.** Histogram of residuals at the phases: a) training, b) testing; source: own study

suggesting that the ANN model is unbiased and does not consistently overpredict or underpredict the flooding indices (surface volume). Although the majority of residuals are concentrated around zero, outliers are reaching –25 and +15. These signify outliers where the model had difficulties in precisely forecasting the flooding indices (surface volume).

Furthermore, the box plot in Figure 7 gives more insight into the prediction trends of the ANN model at the testing phase for flooding indices (surface volume) predictions. The red line in each box denotes the median (50th percentile) value of the dataset. The medians of both actual and projected values are closely aligned, indicating that the ANN model well represents the data's central tendency. The blue boxes denote the interquartile range (25th to 75th percentile).

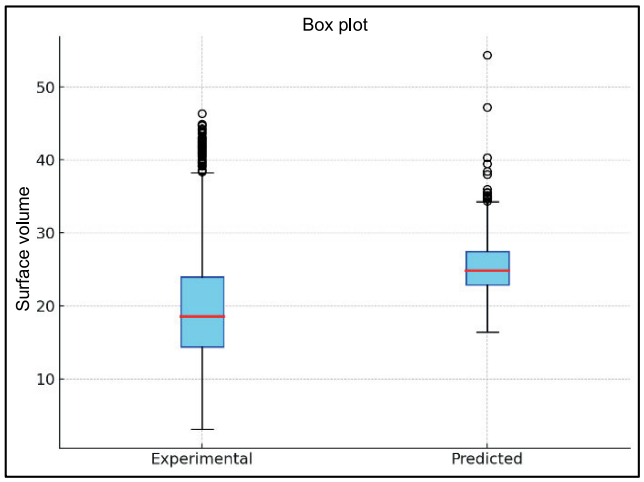


**Fig. 7.** Box plot of the actual and predicted flooding indicator (surface volume) during the testing; source: own study

In modelling complex, non-linear scenarios such as flooding occurrence involving multiple variables based on meteorological parameters, the observed trends in the performance metrics align with known theoretical and practical dynamics of the neural network model. In this research, ANN maintains lower *RMSE* and *MAE*, showcasing consistent performance across data representations. The ANNs learn representations of the input data through hidden layers, enabling them to identify and emphasise relevant features. This is particularly useful when input variables interact in non-trivial ways, as in meteorological data (Li, J. *et al.*, 2024). In addition, ANN consistently performs well in both training and testing phases, despite changes in data distribution. The ANNs can scale with large datasets, adapting through iterative learning processes like backpropagation, which minimises error across the entire dataset. This adaptability makes them particularly suited for dynamic and complex datasets like weather scenarios (Kocher and Kumar, 2021).

## CONCLUSIONS

This study demonstrated the effectiveness of machine learning in predicting floods in the Vaal River Basin. Its goal was to create a computational intelligence model that would use artificial neural networks to better anticipate floods, and it was successful in doing so. In order to anticipate the likelihood of a flood, we looked at the use of artificial neural networks, a non-linear auto-regressive machine learning technique trained on patterns of previous rainfall levels. A 30-year (1994–2024) dataset was collected from the South African Weather Service and preprocessed using standard techniques. Hyper-parameter optimisation of the models was carried out using a grid-search method. The artificial neural network (ANN) model was developed by testing different topologies, training algorithms and activation functions at both the hidden and output layers. The performance of the models was evaluated using relevant statistical metrics, namely root mean square error (*RMSE*), mean absolute percentage error (*MAPE*), mean absolute error (*MAE*), value accounted for (*VAE*) and *R*-value. The ANN model with tansig-tansig activation function and Levenberg–Marquardt training algorithms outperformed other architectures with *RMSE* of 6.245, *MAPE* of 25.95%, *MAE* of 4.656, *VAE* of 7.843 and *R*-value of 0.823 at the training. This research demonstrated the viability of machine learning-based flooding predictions based on weather variables, contributing to flood risk management strategies.

Since *RMSE* imposes a greater penalty on substantial errors compared to *MAE*, the proximity of *RMSE* and *MAE* suggests that significant errors are comparatively few in both phases. A high *MAPE* during testing indicates that the model has difficulty making accurate predictions for certain severe or high-variability instances in the testing set. The decline in *R*-value signifies a reduction in the model's predictive efficacy on testing data, perhaps attributable to overfitting or inadequate representation of testing circumstances in the training dataset.

Overall, the study indicates that the model does not exhibit a bias towards overprediction or underprediction. The majority of residuals fall within a range from –10 to +10, signifying that the model's predictions are often near the actual values.

Future research should focus on:

- taking climate change scenarios into account;
- increasing the forecasts' geographic resolution;
- creating comprehensible resources for local communities and policymakers;
- establishing a performance baseline for comparison.

## ACKNOWLEDGEMENTS

The authors would like to thank the South African Weather Service (SAWS) and the Department of Water and Sanitation (DWS) for providing the data used in this study. Civil Engineering Science Department of the University of Johannesburg, South Africa.

## CONFLICT OF INTERESTS

All authors declare that they have no conflict of interests

## REFERENCES

Adeleke, O. and Jen, T.-C. (2022) "A FCM-clustered neuro-fuzzy model for estimating the methane fraction of biogas in an industrial-scale bio-digester," *Energy Reports*, 8(15), pp. 576–584. Available at: https://doi.org/10.1016/j.egyr.2022.10.265.

Akanbi, R.T., Davis, N. and Ndarana, T. (2020) "Climate change and maize production in the Vaal catchment of South Africa:

Assessment of farmers' awareness, perceptions and adaptation strategies," *Climate Research*, 82, pp. 191–209.

Akpotu, S.O. (2021) *Analysis of temporal and spatial changes in major dissolved salts in the Vaal River system over a 40 year period*. MSc Thesis. University of Pretoria. Available at: http://hdl.handle.net/2263/83472 (Accessed: April 10, 2025).

Antwi-Agyakwa, K.T., Afenyo, M.K. and Angnuureng, D.B. (2023) "Know to predict, forecast to warn: A review of flood risk prediction tools," *Water*, 15(3), 427. Available at: https://doi.org/10.3390/w15030427.

Anuruddhika, M. *et al.* (2025) "A review of river flood models: Methods and applications for forecasting and simulation," *Ceylon Journal of Science*, 54(1), pp. 317–338. Available at: https://doi.org/10.4038/cjs.v54i1.8286.

Baloyi, L. (2022) Assessing river-aquifer interaction for sustained water abstraction, Lower Vaal Catchment, South Africa. MSc Thesis. University of the Western Cape.

Bargam, B. *et al.* (2024) "Evaluation of the support vector regression (SVR) and the random forest (RF) models accuracy for streamflow prediction under a data-scarce basin in Morocco," *Discover Applied Sciences*, 6, 306. Available at: https://doi.org/10.1007/s42452-024-05994-z.

Belina, Y., Kassa, A.K. and Masinde, M. (2025) "Integrating machine learning and physical models for rainfall-runoff prediction in the Upper Baro Akobo River Basin, Ethiopia," *Hydrological Sciences Journal*, 70(12), pp. 2129–2146. Available at: https://doi.org/10.1080/02626667.2025.2518197.

Belina, Y., Kebede, A. and Masinde, M. (2024) "Comparative analysis of HEC-HMS and machine learning models for rainfall-runoff prediction in the upper Baro watershed, Ethiopia," Hydrology Research, 55(9), pp. 873–889. Available at: https://doi.org/10.2166/nh.2024.032.

Bergstra, J. and Bengio, Y. (2012) "Random search for hyperparameter optimization," *The Journal of Machine Learning Research*, 13, pp. 281–305.

Bibi, T.S. and Kara, K.G. (2023) "Evaluation of climate change, urbanization, and low-impact development practices on urban flooding," *Heliyon*, 9(1), e12955. Available at: https://doi.org/10.1016/j.heliyon.2023.e12955.

Boboye, O. and Dorasamy, N. (2025) "Contingency planning and flood disaster management in Nigeria: A critical study," *e-BANGI Journal*, 22(2), pp. 657–671. Available at: https://doi.org/10.17576/ebangi.2025.2202.53.

Cappelli, F. *et al.* (2023) "Feature importance measures to dissect the role of sub-basins in shaping the catchment hydrological response: A proof of concept," *Stochastic Environmental Research and Risk Assessment*, 37(4), pp. 1247–1264. Available at: https://doi.org/10.1007/s00477-022-02332-w.

Chen, L., Chen, P. and Lin, Z. (2020) "Artificial intelligence in education: A review," *IEEE Access*, 8, pp. 75264–75278. Available at: https://doi.org/10.1109/ACCESS.2020.2988510.

Choi, J. *et al.* (2022) "Learning enhancement method of long short-term memory network and its applicability in hydrological time series prediction," *Water*, 14(18), 2910. Available at: https://doi.org/10.3390/w14182910.

Cvetković, V.M. *et al.* (2024) "Geospatial and temporal patterns of natural and man-made (technological) disasters (1900–2024): Insights from different socio-economic and demographic perspectives," *Applied Sciences*, 14(18), 8129. Available at: https://doi.org/10.3390/app14188129.

Farina, A. *et al.* (2023) "A simplified approach for the hydrological simulation of urban drainage systems with SWMM," *Journal of Hydrology*, 623, 129757. Available at: https://doi.org/10.1016/j.jhydrol.2023.129757.

Funke, N.S. (2025) Three perspectives, one problem: A multi-theoretical approach to the role of expertise in the acid mine drainage policy controversy in Gauteng, South Africa. PhD Thesis. Vrije Universiteit Amsterdam.

Ha, D. and Tang, Y. (2022) "Collective intelligence for deep learning: A survey of recent developments," *Collective Intelligence*, 1(1). Available at: https://doi.org/10.1177/26339137221114874.

Haddad, K. and Rahman, A. (2020) "Regional flood frequency analysis: Evaluation of regions in cluster space using support vector regression," *Natural Hazards*, 102(1), 489-517. Available at:

Ibrahim, M.Q. *et al.* (2025) "Optimising convolution neural networks: A comprehensive review of hyperparameter turning through metaheuristics algorithms," *Archives of Computational Methods in Engineering*, 32, pp. 5123–5160. Available at: https://doi.org/10.1007/s11831-025-10292-x.

Jain, S., Singh, R. and Seth, S. (2000) "Design flood estimation using GIS supported GIUHApproach," *Water Resources Management*, 14(5), pp. 369–376. Available at: https://doi.org/10.1023/A:1011147623014.

Jamsheed, F. and Iqbal, S.J. (2023) "Simplified artificial neural network based online adaptive control scheme for nonlinear systems," *Neural Computing and Applications*, 35, pp. 663–679. Available at: https://doi.org/10.1007/s00521-022-07760-x.

Jeba, G.S. and Chitra, P. (2024) "Exploring the power of deep learning and big data in flood forecasting: State-of-the-art techniques and insights," in R. Nagarajan *et al.* (eds.) *Intelligent Systems and Sustainable Computational Models*. New York: Auerbach Publications, pp. 14–33.

Khan, M.A.R. *et al.* (2025) "Development of a fog computing-based real-time flood prediction and early warning system using machine learning and remote sensing data," *Journal of Sustainable Development and Policy*, 1(01), pp. 144–169. Available at: https://doi.org/10.63125/6y0qwr92.

Kocher, G. and Kumar, G. (2021) "Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges," *Soft Computing*, 25(15), pp. 9731–9763. Available at: https://doi.org/10.1007/s00500-021-05893-0.

Kumar, V. *et al.* (2023a) "The state of the art in deep learning applications, challenges, and future prospects: A comprehensive review of flood forecasting and management," *Sustainability*, 15(13), 10543. Available at: https://doi.org/10.3390/su151310543.

Kumar, V. *et al.* (2023b) "Advanced machine learning techniques to improve hydrological prediction: A comparative analysis of streamflow prediction models," *Water*, 15(14), 2572. Available at: https://doi.org/10.3390/w15142572.

Lange, H. and Sippel, S. (2020) "Machine learning applications in hydrology," in D.F. Levia *et al.* (eds.) *Forest-water interactions*. Cham: Springer, pp. 233–257. Available at: https://doi.org/10.1007/978-3-030-26086-6_10.

Li, H. *et al.* (2024) "Water-level prediction analysis for the three gorges reservoir area based on a hybrid model of LSTM and its variants," *Water*, 16(9), 1227. Available at: https://doi.org/10.3390/w16091227.

Li, J. *et al.* (2024) "Optimizing flood predictions by integrating LSTM and physical-based models with mixed historical and simulated data," *Heliyon*, 10(13), e33669. Available at: https://doi.org/10.1016/j.heliyon.2024.e33669.

Liu, Z. *et al.* (2025) "Artificial intelligence for flood risk management: A comprehensive state-of-the-art review and future directions," *International Journal of Disaster Risk Reduction*, 117, 105110. Available at: https://doi.org/10.1016/j.ijdrr.2024.105110.

Ma, J. *et al.* (2023) "Towards data-driven modeling for complex contact phenomena via self-optimized artificial neural network methodology," *Mechanism and Machine Theory*, 182, 105223. Available at: https://doi.org/10.1016/j.mechmachtheory.2022.105223.

Ma, K. *et al.* (2024) "Transfer learning framework for streamflow prediction in large-scale transboundary catchments: Sensitivity analysis and applicability in data-scarce basins," *Journal of Geographical Sciences*, 34(5), pp. 963–984. Available at: https://doi.org/10.1007/s11442-024-2235-x.

Mamphwe, A. (2021) *Effects of flood dynamics on island geomorphology in a large mixed bedrock-alluvial anabranching river: A case study of the Vaal River near Parys*. PhD Thesis. University of the Western Cape.

Mashaly, A.F. and Fernald, A.G. (2020) "Identifying capabilities and potentials of system dynamics in hydrology and water resources as a promising modeling approach for water management," *Water*, 12(5), 1432. Available at: https://doi.org/10.3390/w12051432.

Masindi, K. (2021) *Water resources modelling in the Vaal River Basin: An integrated approach*. Johannesburg: University of the Witwatersrand.

Masindi, K and Abiye, T. (2018) "Assessment of natural and anthropogenic influences on regional groundwater chemistry in a highly industrialized and urban region: A case study of the Vaal River Basin, South Africa," *Environmental Earth Sciences*, 77, 722. Available at: https://doi.org/10.1007/s12665-018-7907-3.

Matthew, B., Joshua, M. and Philip, M. (2025) "MLOps and DataOps integration: The future of scalable machine learning deployment," Available at: https://www.researchgate.net/publication/391594334_MLOps_and_DataOps_Integration_The_Future_of_Scalable_Machine_Learning_Deployment (Accessed: April 10, 2025).

Mishra, A. *et al.* (2022) "An overview of flood concepts, challenges, and future directions," *Journal of Hydrologic Engineering*, 27(6), 03122001. Available at: https://doi.org/10.1061/(ASCE)HE.1943-5584.0002164.

Mishra, P.K. and Dwivedi, R. (2025) "Soft computing techniques for rainfall-runoff modeling and analysis in river basin," *Water Resources Management*, 39, pp. 3859–3881. Available at: https://doi.org/10.1007/s11269-025-04134-5.

Mohamadi, S., Ehteram, M. and El-Shafie, A. (2020) "Accuracy enhancement for monthly evaporation predicting model utilizing evolutionary machine learning methods," *International Journal of Environmental Science and Technology*, 17, pp. 3373–3396. Available at: https://doi.org/10.1007/s13762-019-02619-6.

Netto, R. *et al.* (2021) "Algorithm selection framework for legalization using deep convolutional neural networks and transfer learning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(5), pp. 1481–1494. Available at: https://doi.org/10.1109/TCAD.2021.3079126.

Nourani, V., Paknezhad, N.J. and Tanaka, H. (2021) "Prediction interval estimation methods for artificial neural network (ANN)-based modeling of the hydro-climatic processes, a review," *Sustainability*, 13(4), 1633. Available at: https://doi.org/10.3390/su13041633.

Obaid, A., Adam, E. and Ali, K.A. (2023) "Land use and land cover change in the vaal dam catchment, South Africa: A study based on remote sensing and time series analysis," *Geomatics*, 3(1), pp. 205–220. Available at: https://doi.org/10.3390/geomatics3010011.

Omeka, M.E. *et al.* (2024) "Efficacy of GIS-based AHP and data-driven intelligent machine learning algorithms for irrigation water quality prediction in an agricultural-mine district within the Lower Benue Trough, Nigeria," *Environmental Science and Pollution Research*, 31(41), pp. 54204–54233. Available at: https://doi.org/10.1007/s11356-023-25291-3.

Remilekun, A.T. *et al.* (2021) "Integrated assessment of the influence of climate change on current and future intra-annual water availability in the Vaal River catchment," *Journal of Water and Climate Change*, 12(2), pp. 533–551. Available at: https://doi.org/10.2166/wcc.2020.269.

Samantaray, S., Sahoo, A. and Agnihotri, A (2023) "Prediction of flood discharge usimg hybrid PSO-SVM algorithm in Barak river basin," *Journal of MethodsX*, 10, 102060. Available at: https://doi.org/10.1016/j.mex.2023.102060.

Sasanapuri, S.K., Dhanya, C.T. and Gosain, A.K. (2025) "A surrogate machine learning model using random forests for real-time flood inundation simulations," *Environmental Modelling and Software*, 188, 106439. Available at: https://doi.org/10.1016/j.envsoft.2025.106439.

Sayedi, S.S. (2023) *Combining expert opinions to assess risk of change in earth systems: Permafrost collapse, global wildfire, and water security*. PhD Thesis. Brigham Young University.

Schoener, G. and Stone, M.C. (2020) "Monitoring soil moisture at the catchment scale – A novel approach combining antecedent precipitation index and radar-derived rainfall data," *Journal of Hydrology*, 589, 125155. Available at: https://doi.org/10.1016/j.jhydrol.2020.125155.

Sharma, S.K. and Srivastava, S. (2021) "An overview on neural network and its application," *International Journal for Research in Applied Science and Engineering Technology*, 9(8), pp. 1242–1248.

Sun, Z. *et al.* (2021) "Hybrid model with secondary decomposition, randomforest algorithm, clustering analysis and long short memory network principal computing for short-term wind power forecasting on multiple scales," *Energy*, 221, 119848. Available at: https://doi.org/10.1016/j.energy.2021.119848.

Tabbussum, R. and Dar, A.Q. (2021) "Performance evaluation of artificial intelligence paradigms – Artificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference system for flood prediction," *Environmental Science and Pollution Research*, 28(20), pp. 25265–25282. Available at: https://doi.org/10.1007/s11356-021-12410-1.

Thakur, A. and Konde, A. (2021) "Fundamentals of neural networks," *International Journal for Research in Applied Science and Engineering Technology*, 9(8), pp. 407–426.

Uzair, M. and Jamil, N. (2020) "Effects of hidden layers on the efficiency of neural networks," *2020 IEEE 23rd International Multitopic Conference (INMIC)*. Available at: https://doi.org/10.1109/INMIC50486.2020.9318195.

Waqas, M. *et al.* (2023) "Potential of artificial intelligence-based techniques for rainfall forecasting in Thailand: A comprehensive review," *Water*, 15(16), 2979. Available at: https://doi.org/10.3390/w15162979.

Worden, K. *et al.* (2023) "Artificial neural networks," in T. Rabczuk and K.J. Bathe (eds.) *Machine learning in modeling and simulation: Methods and applications. Computational methods in engineering and the sciences*. Cham: Springer, pp. 85–119. Available at: https://doi.org/10.1007/978-3-031-36644-4_2.

Xu, K. *et al.* (2023) "Rapid prediction model for urban floods based on a light gradient boosting machine approach and hydrological–hydraulic model," *International Journal of Disaster Risk Science*, 14(1), pp. 79–97. Available at: https://doi.org/10.1007/s13753-023-00465-2.

Zhang, J.-R. *et al.* (2007) "A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training," *Applied Mathematics and Computation*, 185(2), pp. 1026–1037. Available at: https://doi.org/10.1016/j.amc.2006.07.025.