Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765527

COMPARATIVE ANALYSIS OF MACHINE LEARNING BASED HYBRID MODELS IN FLOOD FORECASTING

Deelip Ananda Patil1, Dr. K. M. Alaskar2

1 Research Scholar, Bharati Vidyapeeth (Deemed To Be University), Institute of Management Kolhapur

2 Professor, Bharati Vidyapeeth (Deemed To Be University), Institute of Management Kolhapur

Abstract

Floods are a type of natural calamity that can harm infrastructure, socioeconomics, and human lives. To offer citizens a sustainable flood risk management system, flood forecasting is crucial. This paper suggests a straightforward machine learning (ML) method that consists of two or more generic algorithms that work in conjunction to answer problems that they were not intended to. Since the majority of machine learning algorithms are tailored for a specific dataset or task, merging different ML algorithms can significantly enhance the end result by assisting in either tuning one another, generalisation, or adaptation to new tasks. The purpose of this paper is to provide an understanding and comprehensive review of machine learning-based hybrid models used in long-term and short-term flood forecasting. It entails researching machine learning-based hybrid models used for flood forecasting and conducting a comparative assessment of the models' parameters, pre-processing methods, and performance measurements. According to this review, machine learning-based hybrid models have been widely used for short-term and long-term flood forecasting. As predictors, various parameters or flood variables have been used. The hybridization of the model has been found to improve forecast performance. The findings of this study will benefit future researchers by providing information on current progress in the use of machine learning-based hybrid models in short-term and long-term flood forecasting.

Keywords: Flood Forecasting, Hybrid Models, Machine Learning Models, Long-term and Short-term Flood Forecasting.

1. Introduction

Floods are one of the most destructive natural disasters, and modeling them is very challenging. The goal of developing a flood prediction model is to reduce the risk and minimise the loss of human life as well as property damage. There have been numerous mathematical models proposed to reduce the risk of flooding. The development of technology has made it possible to predict floods in order to reduce flood damage. In this review, we highlight a machine learning-based hybrid models for predicting flood. Floods can be predicted with a certain amount of lead time. According to [27], the forecast lead time can be classified as short-term (up to 2 days), medium-term (up to ten (10) days), long-term (more than ten (10) days), and seasonal (it takes several months).

Maier et al. [51] suggests that Water level, river flood, soil moisture, rainfall-discharge, precipitation, river inflow, peak flow, river flow, rainfall-runoff, flash flood, rainfall, streamflow, seasonal stream flow, flood peak discharge, urban flood, plain flood, groundwater level, rainfall stage, flood frequency analysis, flood quantiles, surge level, extreme flow, storm surge, typhoon rainfall, and other flood resource variables can be used to categorise the applications in flood prediction.

Hybrid models enhance forecasting ability as well as usable lead time, indicating the possibility of operationalizing complementary model structures. While more data is required to provide actual value with the hybrid method to impact-based translation of flood forecasts, the data requirements that indicate to the need for systematic data collection of flood impacts during and after flood events are identified.

Although physical-based models and data-driven models are excellent tools for predicting hydrological events, they each have their own limitations, such as complicated calculation and longer processing times in physical-based models and model performance instability in data-driven models. Because of these constraints, researchers are increasingly employing sophisticated data-driven models such as machine learning [52]. Forecasting in machine learning is done using historical data and does not require understanding of the underlying physical processes [63]. Machine learning has achieved high predictive potential with less complexity, less development time, and minimal inputs when compared to completely distributed models [53][39]. The use of machine learning has proven its ability to outperform conventional physical models in producing acceptable forecasts while accommodating the non-linearity of hydrological events [39].

Machine learning models such as Decision Tree (DT), Multilayer Perceptron (MLP), Adaptive Neuro-Fuzzy Inference System (ANFIS), Wavelet Neural Network (WNN), Ensemble Prediction Systems (EPSs), Artificial Neural Networks (ANN) [54][55][56], Random Forest [57], Extreme Learning Machine (ELM) [58], Support Vector Machine (37), Support Vector Regression (SVR) [59] and Neuro Fuzzy [60]. While individual machine learning algorithms have yielded significant results in flood forecasting, the performance of such models can be improved by combining them with other machine learning methods. Hybridization not only accelerates the learning process and improves generalization ability [58][60], but also improves prediction accuracy [60][61][62].

When data is scarce, a machine learning model's ability to learn from previous data can be a deficiency that affects the model's performance. Therefore, optimization of data preprocessing is important to solve this problem. To improve prediction performance and accuracy, pre-processing methods such as the Particle Swarm Optimization algorithm (PSO) [43], Stochastic Gradient Descent (SGD), Wavelet Transform [44] [45], Ensemble Empirical Mode Decomposition (EEMD) [46] and the Genetic Algorithm used (GA) [47] and Correlation Analysis [30] used.

In this review, we present insight of machine learning-based hybrid models used in long-term and short-term flood forecasting.

2. Research Methods and Outline

A literature review was carried out in order to identify and provide a thorough understanding of the machine learning-based hybrid model in flood forecasting. This review is based on Kitchenham and Brereton's [50] guidelines, which allow us to discover and synthesise information in published materials in a systematic manner. It was founded on pre-defined research questions in order to provide concise analysis and valuable information to the research community.

This SLR is conducted by adhering to a formal procedure with defined stages, making it more objective and repeatable. The acceptance and appreciation of this study's conclusion depend heavily on this procedure. The three major stages of SLR are plan the review, conduct the review, and report the review. The actions taken for each stage of this research are listed in Figure 1 [48].

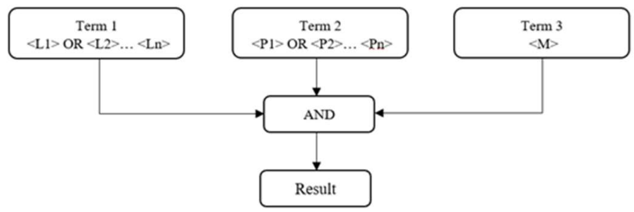

Fig. 1 Systematic literature review Stages

In this review, it is essential to conduct a formal, structured search procedure. It assists in locating all pertinent and related literature using the online digital resources that fit the search criteria.

The keyword search query for the automatic search contained three main search terms. The term most frequently used to express flood prediction is Term 2 (P1-Pn>), while the lead time of the forecast is regarded to be Term 1 (L1-Ln>). The terms "short term," "daily," "hourly," "long term," "monthly," "seasonal," "yearly," and "annually" were included in Term 1; "flood forecasting," "flood prediction," "flood estimation," or "flood analysis" were included in Term 2; and "Hybrid" was included in Term 3. The search query process was shown in Figure 2.

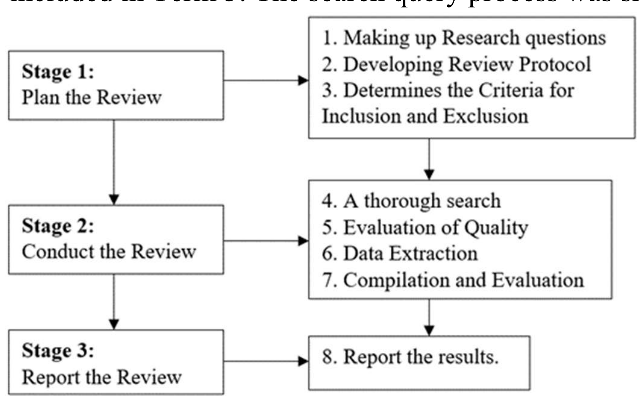

Fig. 2 Method to generate search query automatically

Following automatic and manual searching, primary papers were included in this study based on the initial search results. All primary papers are quality checked to determine their value.

The analysis results and discussion are presented in the following section. Section 3 presents the machine learning-based hybrid models developed for short- and long-term flood forecasts, while Section 4 presents a comparative assessment of the machine-learning-based hybrid model for short- and long-term flood forecasts. including flooding variables, pre-processing techniques, and performance measurement methods used. The study's conclusion is found in Section 5.

3. Machine Learning based Hybrid Models in Flood Forecasting

Hybrid models based on machine learning are essential for flood forecasting to support water resource management and reduce the impact of flooding. It helps in communicating with people about possible flooding. Researchers are now interested in investigating and exploring the realm of long-term and short-term flood forecasting through the use of machine learning-based hybrid models.

The broad view of developing the machine learning based hybrid model can be summarized as shown in Figure 3. This flow represents the general flow, and difference in result can be found among researchers that use the original data sets without any pre-processing and involve the validation process between training and testing. The development of a machine learning-based hybrid model for short- and long-term flood forecasts collects data from reliable resources. Then the data is pre-processed by converting the original datasets into a new format and preparing it as input to the model. Preprocessing is crucial as it determines the input possibilities for the model. Hybrid model development phase in which tools or programming languages are used to turn the algorithm into a working model. This model is then evaluated during training and testing of the model.

The machine learning-based hybrid model is developed in short-term and long-term flood forecasting, where the hybrid model uses different machine learning methods in the form of an integration or an ensemble.

Integration has included integrating a machine learning model with another machine learning model, a data-driven model, a physics-based model, or other traditional methods. While an ensemble can be achieved by combining multiple models through aggregation, bagging, boosting, shuffling, or stacking techniques.

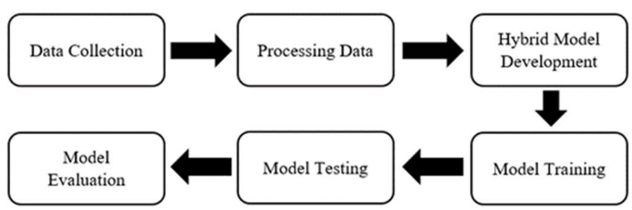

Fig. 3 Flow of Machine Learning based Hybrid Model development for flood forecasting

3.1 Machine Learning based Hybrid Models in Short-Term Flood Forecasting

There is a growing tendency to develop hybrid machine learning (ML) methods to improve prediction quality in terms of accuracy, generalization, uncertainty, longer lead time, speed, and computational cost. There are many different hybrid techniques, including more popular ones like ANFIS[24] and WNN[22], as well as more innovative ones like SVM-FR[20], SAINA-LSTM[10], Transformer Neural Network[15], LSTM-seq2seq[11], Wavelet-based NARX[13], WBANN[22], LSTM-KNN[21], CAGANet[23], RNN-SVR[14], RSVRCPSO[14], MLR-ANN, FFRM-ANN[16] and EPSs[19]. These methods are presented in Table 1; An overview of the methods and their applications is followed by a discussion on ML methods.

Self-activated and Internal Attention LSTM, or SAINA-LSTM, is a novel attention-based Long-Short Term Memory (LSTM) Cell Deep Learning (DL) model for streamflow simulation post-processing presented by [10]. In order to increase the focus on the more important time points and to improve the information flow of the cell, they built improved self-awareness mechanisms into the inner structure of the LSTM cell in this model. The SAINA performance

LSTMs are then compared to those of the EnsPost streamflow prediction ensemble postprocessor currently used by the National Weather Service, a recently created multiscale alternative, gradient boosting, two other deep learning algorithms (LSTM and Gated Recurrent Unit (GRU)) and several other machine learning algorithms.

Three machine learning models and four deep learning models are compared to the LSTM-seq2seq model proposed by [11]. The model's prediction accuracy and convergence rate are evaluated using the RMSE and NSE metrics. The accuracy of the predictions of the LSTM-seq2seq, LSTM-BP, LSTM and BP models on data sets with different characteristics is examined. The results show that on the dataset with stationarity and trend, LSTM-seq2seq and LSTM-BP models provide better predictions than LSTM and BP models.

Using a dataset of level-based precipitation data for the years 2000 to 2010, reference [13] proposed a wavelet-based NARX (WNARX) model for daily precipitation forecasting. WNARX proved better when predictive performance was also compared to ANN, WENN, ARMAX, and NARX models. To accurately estimate rainfall, reference [14] proposed a hybrid forecasting method called RSVRCPSO. RNN, SVR and a chaotic particle swarm optimization method are combined in RSVRCPSO (CPSO). This dataset, which contained information on nine typhoon events, was collected by three rain gauges between 1985 and 1997. The results showed that the proposed model performed superiorly for predicting precipitation. The RSVRCPSO model produced less RMSE learning and testing than the SVRCPSO, resulting in superiority in prediction.

Castangia et al. [15] used the Transformer neural network to predict potential flooding with a day's lead time. To achieve this, they trained a model to predict the water level of a river based on previous observations from upstream stations. According to the results, the transformer outperforms both LSTM and GRU recurrent neural networks in making predictions.

To predict hourly flow phases, Hsu et al. [16] proposed a hybrid model called FFRM-ANN model that combined the Flash Flood Routing model (FFRM) and ANN. In this research, the FFNN and FBNN-ANN algorithms were used. Precipitation and flow rate data from eight typhoon events between 2004 and 2005 were selected to train the model. The results showed that the FFRM-ANN hybrid model provided an effective FFRM for accurate flood forecasts. The effectiveness of the proposed method was demonstrated by comparing the hybrid method with each algorithm used in the study.

Le et al. [17] emphasized that the complexity of the StackedLSTM and BiLSTM models does not come with an increase in performance, since the comparison results show that their respective performances are no greater than that of the two standard models LSTM and GRU. The results of this study demonstrate that LSTM-based models can produce impressive predictions even in the presence of upstream dams and reservoirs. The LSTM and GRU models with a simple architecture (one hidden layer) are adequate to produce highly reliable predictions while minimizing the computation time for the power flow prediction problem.

Pan et al. [18] proposed a monsoon rain enhancement (AME) based on ANNs that is a mixture of linear regression and a state-space neural network (SSNN). A benchmarking was performed between the performance of the proposed model and the hybrid MLR-ANN approach. Based on 371 rain gauge stations for six typhoons, this dataset included measurements of total rainfall, wind, and humidity from 1989 to 2008. The results showed the technique was highly reliable and improved prediction accuracy for R2, peak runoff, and total volume. Reference [19] used daily flood data from the 2013-2014 storm season to build an EPS model of six ANNs for the daily stream flow forecast. The proposed model had a short development time and offered probabilistic predictions to remove prediction uncertainties. According to reports, the ensemble prediction method is very reliable and useful.

In reference [20] an advanced ensemble model for flood forecasting was proposed by combining FR and SVM. The results were compared to DT. An inventory map with flood forecasts for various locations was included in this data set. During construction of the model, up to 100 flood sites were used for training and validation. The results of the evaluation showed that the ensemble model had a high success rate. The results demonstrated the effectiveness, accuracy and speed of the model in determining flood vulnerability.

The results of [21] showed that the LSTM with internal memory can learn and maintain long-term input-output relationship dependencies in a variety of climates. The results show that LSTM provides results comparable to the conceptual XAJ model and that it provides more robust results than the simple RNN model. Comparisons between the coupled model and the LSTM model show that the ANN algorithm can improve the accuracy of the LSTM model in predicting runoff in three catchments.

Table 1. Short-term flood prediction using Machine Learning based hybrid models

Reference	Modelling Techniques	Flood Resource Variables	Estimation Type	Region
[10]	SAINA-LSTM vs LSTM vs GRU	Streamflow Water Level	1 to 7 days	United States
[11]	LSTM-seq2seq vs DeepAR vs ARIMA vs LR vs RF	Water Level	Hourly	China
[12]	ANN–NLPM vs. ANN	Rainfall–runoff	Daily	China
[13]	Wavelet-based NARX vs. ANN, vs. WANN	Streamflow forecasting	Daily	India
[14]	RNN–SVR, RSVRCPSO	Flash flood: rainfall forecasting	Hourly	Taiwan
[15]	Transformer Neural Network vs LSTM and GRU	water level	1 Day ahead	Russia
[16]	FFNN vs. FBNN vs. FFRM–ANN	Flash floods	Hourly	Taiwan
[17]	StackedLSTM vs BiLSTM vs LSTM vs GRU vs FFNN vs CNN	Streamflow	1-2 Day	Vietnam

[18]	AME and SSNN vs. ANN	Rainfall forecasting	Hourly	Taiwan
[19]	EPS of ANNs	Flood	Daily	Canada
[20]	SVM-FR vs. DT	Rainfall–runoff	Real-time	Malaysia
[21]	LSTM-KNN vs XAJ vs LSTM vs RNN	Rainfall	Real-time	China
[22]	WBANN vs. WANN vs. ANN vs. BANN	Flood	Hourly	India
[23]	CAGANet vs LSTM vs AM-LSTM SVM	Rainfall	Daily	China
[24]	ANN vs. ANFIS	Daily flow	Daily	Iran
[25,26]	ANFIS vs. ANN	Water level	Hourly	Taiwan

A hybrid wavelet, bootstrap technique and ANN model known as WBANN has been proposed in reference [22]. It increased the accuracy and reliability of the ANN model's short-term flood forecast. The efficiency of WBANN was compared to that of WNN, bootstrap-based ANNs, and BANNs. The ANN models have been greatly improved by wavelet decomposition. In addition, the bootstrap resampling provided reliable insights. [23] suggested a combined neural network model CAGANet based on data augmentation to predict daily runoff in Sichuan province's Qingxi basin. When forecasting on a data collection without using data augmentation methods, the proposed CAGANet model has greater prediction accuracy than a single SVM model, neural network models LSTM and AM-LSTM, and its NSE can reach 0.854.

Using ANN, ANFIS, MLR and MNLR, Rezaeianzadeh (2014) [24] showed a number of forecasting systems for daily flow prediction. In addition, RMSE and R2 were used to determine model performance. Precipitation information from various meteorological stations was included in this dataset. According to the assessment, ANFIS, MLR, and ANN models performed worse than MNLR models with lower RMSE values. MNLR has also been recommended as a low-cost, effective model for daily flow prediction. For 1-3 hours before high tide, Chang and Chang [25] created an accurate water level prediction algorithm based on ANFIS. The ANFIS correctly and accurately predicted the water level. Hourly water levels from five sensors from 1971 to 2001 were used. They concluded that the ANFIS model could effectively handle a large dataset through fast learning and accurate predictions [26].

3.2 Machine Learning based Hybrid Models in Long-Term Flood Forecasting

Machine learning algorithms for long-term flood forecasting are also known as hybrid models. Hybrid models are developed by integrating, compositing, or combining multiple machine learning methods to generate forecasts with acceptable performance and accuracy. To improve model performance, some researchers combine machine learning with other traditional methods, such as physical techniques.

Various hybrid models are created in flow forecasting by combining input optimization methods with machine learning models. Researchers are interested in optimization techniques such as Discrete Wavelet Transform (DWT)[45], Empirical Mode Decomposition (EMD)[46], Ensemble Empirical Mode Decomposition (EEMD)[46], and Genetic Algorithm (GA)[47]. Using hybrid models such as DWT-ANN[40], DWT-RBFNN[44] and DWT-SVR[33] has been shown to provide accurate monthly runoff forecasts.

Meshram et al. [28] described three AI methods (ANFIS, GP and ANN) to predict discharge into India's Shakkar watershed (Narmada Basin). According to the results, for all AI methods (ANFIS, GP, and ANN), the model with cyclic terms performed better than models that did not take into account the periodic nature and were only applied considering the previous current flow. Araghinejad [29] demonstrated the use of ensembles for probabilistic flood forecasts in real scenarios. He used K-nearest-neighbor regression to combine individual networks to improve prediction accuracy. The hybrid model of K-NN was proposed as an EPS of ANNs to improve the generalization ability of neural networks, and the results were compared to those obtained using MLP, MLP-PLC, and ANN. Hourly reservoir water level records from 132 typhoons from 1971 to 2001 were used. The proposed EPS demonstrated potential generalization and predictive accuracy.

Compared to traditional multilayer perceptron, hybridization of fuzzy neural network and least squares method (fuzzy MLP) has been shown to provide higher quality flow prediction [30]. Tantanee et al. [31] proposed WARM, a hybrid of wavelet and autoregressive models that was better suited to long lead times.

WLGP is a composite model that predicts monthly current flow using discrete wavelet transform and linear genetic programming. Using WLGP with multi-resolution time-series subsignals as inputs has improved prediction accuracy over single models such as Linear Genetic Programming (LGP), WaveletANN (WHEN), ANN, and Multi Linear Regression MLR [32]. By developing an EEMD-ANN model for monthly forecasts, reference [33] contributed to the improvement of decomposition ensemble forecast models. There was a significant increase in accuracy compared to SVM, ANFIS and ANNs.

ANFIS-FFA has been shown to outperform traditional ANFIS in terms of prediction accuracy with less input. This is due to the robustness of the FFA, which helps to optimize the membership function parameters in each input pair [34]. This prediction model uses climate signals as predictors and compares the results to the standard ANN and POAMA models. This research found that ANFIS outperforms traditional models in predicting springtime precipitation when accurate predictors are used. When forecasting monthly precipitation, the empirical mode decomposition of the ensemble integrated into the Support Vector Machine (EEMD-SVR) has been shown to outperform ANN, ARIMA and SVR [35].

Prasad et al. [36] proposed a hybrid model involving WNN and iterative input selection (IIS). The hybrid model was called IIS-W-ANN and was tested against the M5 model structure. Their data set comprised 40 years of water level measurements in streams. The M5 tree was beaten by the IIS-W-ANN hybrid model. After this research, the novel IIS-W-ANN method should be considered as an excellent flood forecasting model.

Zhu, Zhou, Ye and Meng [37] added the integration of ML with time series decomposition to predict monthly current flow by estimation and comparison of model accuracy. They also combined SVM with Discrete Wavelet Transform (DWT) and EMD. DWT-SVR and EMD-SVR were the names for the hybrid versions. Results showed that decomposition improved the accuracy of predicting current flow, but DWT performed even better. Further comparisons of the SVR, EMD-SVR, and DWT-SVR models showed that EMD and DWT were significantly more accurate than SVR in predicting monthly discharge.

Table 2. Long-term flood prediction using Machine Learning based hybrid models

Reference	Modelling Techniques	Resource Variables	Estimation Type	Region
[28]	ANFIS vs GP vs ANN	Streamflow	Monthly	India
[29]	EPS of ANNs: K-NN vs MLP vs. MLP–PLC vs ANNE	Streamflow	Seasonal	Canada
[30]	Fuzzy MLP vs MLP	Flow	Monthly	Brazil
[31]	WARM vs. AR	Rainfall	Yearly	Thailand
[32]	WLGP vs LGP, vs ANN, vs WANN vs MLR	Stream Flow	Monthly	Iran
[33]	EEMD–ANN vs. SVM vs. ANFIS	Runoff forecast	Monthly	China
[34]	ANFIS-FFA vs ANFIS	Stream Flow	Monthly	Malaysia
[35]	EEMD-SVR vs ANN vs ARIMA vs SVR	Rainfall	Monthly	China
[36]	Hybrid WNN vs. M5-model tree	Streamflow water level	Monthly	Australia
[37]	SVR vs DWT–EMD	Streamflow	Monthly	China
[38]	DWT-RBFNN vs QP-DWT-RBFNN vs Q-RBFNN vs QP-RBFNN	Rainfall	Monthly	China
[39]	ANFIS vs ANN vs POAMA	Rainfall, Climate Signals	Seasonal	Australia
[40]	WNN vs. ANN	Rainfall–runoff	Monthly	Italy

[41]	NFNN-MKV	Rainfall, Inflow, Discharge	Annually Monthly	China
[42]	WA-ELM vs ELM	River Flow	Monthly	Iraq

An artificial neural network is used on an individual basis with runoff and precipitation data to forecast the monthly stream flow in the Jinshan River basin. To address the non-linearity problem, the technique of time series decomposition is used. To obtain suitable results for the Jinsha River basin, artificial neural networks were combined with the discrete wavelet transform technique.[38]

Mekanik et al. [39] predicted monthly precipitation using ANFIS. The performance and accuracy of the ANN model and a physical model were compared and the results for ANFIS were promising. Rainfall data from 1900 to 1999 was used for training and validation, and data from the following decade was used for assessment. ANFIS outperformed the ANN models in all cases, outperformed the Predictive Ocean Atmosphere Model for Australia (POAMA) and outperformed the climatology. The study also demonstrated the accuracy of ANFIS compared to global climate models. In addition, research proposed ANFIS as an alternative tool for long-term forecasting. ANFIS has been reported to be easy to use, with low complexity and input requirements, and requires less development time.

Canna et al. (2005) [40] previously confirmed WNN as the most accurate forecast model for monthly rain-runoff forecasts as well as for other engineering applications. This model can solve the problem of stationary or volatile strong random processes by combining the advantages of the new fuzzy neural network and the Markov prediction model. The NFNN-MKV model is used to determine and forecast the river discharge at Weijiabao on the Weihe River in China for the next 156 months (training and testing for 120-month forecast). Comparisons of the NFNN-MKV model, the WNN model, and the SVR model show that the NFNN-MKV model can significantly improve the prediction accuracy [41].

For predicting river flow in a semi-arid environment, this research highlighted a novel complementary data intelligence (DI) model called Wavelet Extreme Learning Machine (WA-ELM). Monthly flow data from 1991 to 2010 are used to calibrate and validate the applied forecast model, which was built using previous flow data as the predictor. The prediction efficiency of the developed WA-ELM model is verified using a standalone ELM model. The success of the models is evaluated using various statistical metrics and graphical analysis visualizations. The results show that combining a wavelet approach to data preprocessing with an ELM model improves river flow predictability [42]

4. Comparative evaluation of Machine Learning based Hybrid Models for Flood Forecasting

4.1 Parameters in Model Development

The selection of suitable flood variables leads to the use of a machine learning-based hybrid model for short- and long-term flood forecasting. Extensive use is made of historical data from various sources, depending on the forecast location. Precipitation, runoff, climate indices, climate signals, runoff, rainfall, peak runoff, runoff, and runoff are all commonly used parameters by researchers [48]. The Red Cross Red Crescent Climate Center (RCCC) [44], the Australian Bureau of Meteorology's Climate Data Online [36][39], the Royal Netherlands Meteorological Institute Climate Explorer, the Central Water Commission, the Indian Metrological Department's [13][22] and the Malaysian Ministry of Irrigation and Drainage [20] provide most of this historical data.

Researchers can use a mix of parameters, resulting in higher counts for certain parameters. Although a combination of parameters can in some cases have a large impact on the predicted value, a single flood variable can also give an acceptable result. According to one study, using rainfall data improves the accuracy of the model [38]. The historical datasets were used in the primary studies to build and score the model in terms of hourly, daily, weekly, monthly, or yearly.

4.2 Methods for Data Pre-processing

Data pre-processing is an iterative process that transforms raw data into understandable and usable formats. Raw data sets are typically characterized by incompleteness, inconsistencies, poor behavior and trends, and errors. Preprocessing is required to handle missing values and resolve inconsistencies. This can improve the performance of the predictive model [44][45][46]. The choice of input is critical as it affects the precision and accuracy of the forecasting model [33][37].

Preprocessing methods used in the primary studies such as Particle Swarm Optimization Algorithm (PSO) [43], Stochastic Gradient Descent (SGD), Wavelet Transform [44][45], Ensemble Empirical Mode Decomposition (EEMD) [46] and Genetic Algorithm (GA) [47] and correlation analysis [30]

The most exciting aspect of PSO is that it has a stable topology where particles can communicate with each other and increase the learning rate to reach the global optimum. The metaheuristic nature of this optimization algorithm gives us numerous possibilities as it optimizes a problem by iteratively trying to improve a candidate solution. With ongoing research into ensemble learning, its applicability will only increase [43]. The use of wavelet decomposition techniques such as Discrete Wavelet Transform (DWT) decomposed time series data into a shifted and scaled version of a wavelet known as the mother wavelet [64]. It is a useful technique for analyzing time-series variations, which provide information about a signal's time and frequency domains.

The disadvantage of EMD is that mode mixing is common. This disadvantage led the researchers to propose an Ensemble Empirical Mode Decomposition (EEMD). The application of EEMD has significantly improved the forecasting model. It was also discovered that using time series data of different lengths in EEMD can lead to different model performance, necessitating an update of the decomposition and the model as new information is added [33][46]. To select important variables for the model, the Cross Correlation Function (CCF) is used [30]. It calculates the linearity of the similarity between two signals. CCF is used in studies [30] to determine time lags and model input variables. The Genetic Algorithm (GA) is an optimization method based on genetic and natural selection principles [45]. It is based on biologically inspired operators such as mutation, crossover and selection.

4.3 Metrics used for Evaluation of Model Performance

In the field of machine learning, measuring the performance of the model is just as important as building models. Essentially, we're evaluating how accurate our model's predictions are.

Therefore, the predictive model is evaluated against performance metrics. Evaluation metrics quantify the performance of a machine learning model. It involves training a model and then comparing the predictions to the expected values. The models could be evaluated using a single measurement or a combination of measurements to determine model performance [48].

Root Mean Square Errors (RMSE) are the most commonly used performance metrics by researchers, followed by Mean Absolute Error (MAE), Nash-Sutcliffe Efficiency (NSE), Correlation Coefficient (R), Coefficient of Determination (R2), and Mean Absolute Percent Error (MAPE).

Root Mean Square Error or RMSE is one of the most popular measures to estimate the accuracy of the values predicted by our forecasting models versus the actual or observed values while training the regression models or time series models. It measures the error in our predicted values when the target or response variable is a continuous number. Thus, RMSE is a standard deviation of prediction errors or residuals. It indicates how distributed the data are around the line of best fit.

To calculate RMSE, the formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2}$$

Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of each prediction error across all instances in the test set. Each prediction error is the difference between the true value and the predicted value for the instance.

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

The Nash-Sutcliffe efficiency (NSE) is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance (Nash and Sutcliffe, 1970). The Nash-Sutcliffe efficiency indicates how well the plot of observed versus simulated data fits the 1:1 line. NSE = 1, corresponds to a perfect fit of the model to the observed data. NSE = 0 indicates the model predictions are as accurate as the mean of the observed data, Inf < NSE < 0 indicates the observed mean is a better predictor than the model.

$$E = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}$$

Pearson's correlation coefficient helps you find the relationship between two quantities. It gives you the measure of the strength of the association between two variables. The value of Pearson's correlation coefficient can range from -1 to +1. 1 means they are strongly correlated and 0 means no correlation. -1 means there is a negative correlation.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

5. Conclusion

In this article, we reviewed the machine learning-based hybrid models commonly used for short-term and long-term flood forecasting. The main objective of this paper is to explore machine learning-based hybrid models used for flood forecasting and to conduct a comparative assessment of the model parameters, pre-processing methods and performance measurements. Hybridization of the model was found to improve forecasting performance. The first was novel hybridization, either through the integration of two or more machine learning methods or the integration of one machine learning method. The second was the use of data decomposition techniques for the purpose of improving the quality of the data set, which greatly contributed to improving prediction accuracy. The third was the use of an ensemble of methods that dramatically increased the generalizability of the models and reduced prediction uncertainty. The fourth was using add-on optimization algorithms to improve machine learning quality. Flood forecasting is expected to experience significant improvements through these four key technologies for both short-term and long-term forecasting. The results of this study will benefit future researchers by providing information on current advances in the use of machine learning-based hybrid models in short-term and long-term flood forecasting.

References

[1] Cheng Q, Wu Y, Castiglione A, Narducci F, Wan S. DA-Net: Dual Attention Network for Flood Forecasting. Journal of Signal Processing Systems. 2023 Jan 24:1-2.

[2] Indra G, Duraipandian N. An improved flood forecasting system with cluster based visualization and analyzing using GK-ANFIS and CGDNN. Expert Systems with Applications. 2023 Feb 1;212:118747.

[3] Chang LC, Liou JY, Chang FJ. Spatial-temporal flood inundation nowcasts by fusing machine learning methods and principal component analysis. Journal of Hydrology. 2022 Sep 1;612:128086.

[4] Cai B, Yu Y. Flood forecasting in urban reservoir using hybrid recurrent neural network. Urban Climate. 2022 Mar 1;42:101086.

[5] Dong L, Fang D, Wang X, Wei W, Damaševičius R, Scherer R, Woźniak M. Prediction of streamflow based on dynamic sliding window LSTM. Water. 2020 Oct 29;12(11):3032.

[6] Yan L, Chen C, Hang T, Hu Y. A stream prediction model based on attention-LSTM. Earth Science Informatics. 2021 Jun;14:723-33.

[7] Feng J, Wang Z, Wu Y, Xi Y. Spatial and temporal aware graph convolutional network for flood forecasting. In2021 International Joint Conference on Neural Networks (IJCNN) 2021 Jul 18 (pp. 1-8). IEEE.

[8] Lin Y, Wang D, Wang G, Qiu J, Long K, Du Y, Xie H, Wei Z, Shangguan W, Dai Y. A hybrid deep learning algorithm and its application to streamflow prediction. Journal of Hydrology. 2021 Oct 1;601:126636.

[9] Dtissibe FY, Ari AA, Titouna C, Thiare O, Gueroui AM. Flood forecasting based on an artificial neural network scheme. Natural Hazards. 2020 Nov;104:1211-37.

[10] Alizadeh B, Bafti AG, Kamangir H, Zhang Y, Wright DB, Franz KJ. A novel attention-based LSTM cell post-processor coupled with bayesian optimization for streamflow prediction. Journal of Hydrology. 2021 Oct 1;601:126526.

[11] Dai Z, Zhang M, Nedjah N, Xu D, Ye F. A Hydrological Data Prediction Model Based on LSTM with Attention Mechanism. Water. 2023 Jan;15(4):670.

[12] Pang B, Guo S, Xiong L, Li C. A nonlinear perturbation model based on artificial neural network. Journal of Hydrology. 2007 Feb 15;333(2-4):504-16.
[13] Nanda T, Sahoo B, Beria H, Chatterjee C. A wavelet-based non-linear autoregressive with exogenous inputs (WNARX) dynamic neural network model for real-time flood forecasting using satellite-based rainfall products. Journal of Hydrology. 2016 Aug 1;539:57-73.
[14] Hong WC. Rainfall forecasting by technological machine learning models. Applied Mathematics and Computation. 2008 Jun 15;200(1):41-57.
[15] Castangia M, Grajales LM, Aliberti A, Rossi C, Macii A, Macii E, Patti E. Transformer neural networks for interpretable flood forecasting. Environmental Modelling & Software. 2023 Feb 1;160:105581.
[16] Hsu MH, Lin SH, Fu JC, Chung SF, Chen AS. Longitudinal stage profiles forecasting in rivers for flash floods. Journal of hydrology. 2010 Jul 15;388(3-4):426-37.
[17] Le XH, Nguyen DH, Jung S, Yeon M, Lee G. Comparison of deep learning techniques for river streamflow forecasting. IEEE Access. 2021 May 5;9:71805-20.
[18] Pan TY, Yang YT, Kuo HC, Tan YC, Lai JS, Chang TJ, Lee CS, Hsu KH. Improvement of watershed flood forecasting by typhoon rainfall climate model with an ANN-based southwest monsoon rainfall enhancement. Journal of Hydrology. 2013 Dec 9;506:90-100.
[19] Fleming SW, Bourdin DR, Campbell D, Stull RB, Gardner T. Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific Northwest river. JAWRA Journal of the American Water Resources Association. 2015 Apr;51(2):502-12.
[20] Tehrany MS, Pradhan B, Jebur MN. Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. Stochastic environmental research and risk assessment. 2015 May;29:1149-65.
[21] Liu M, Huang Y, Li Z, Tong B, Liu Z, Sun M, Jiang F, Zhang H. The applicability of LSTM-KNN model for real-time flood forecasting in different climate zones in China. Water. 2020 Feb 6;12(2):440.
[22] Tiwari MK, Chatterjee C. Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ANN (WBANN) hybrid approach. Journal of Hydrology. 2010 Nov 26;394(3-4):458-70.
[23] Bi XY, Li B, Lu WL, Zhou XZ. Daily runoff forecasting based on data-augmented neural network model. Journal of Hydroinformatics. 2020 Jul 1;22(4):900-15.
[24] Rezaeianzadeh M, Tabari H, Arabi Yazdi A, Isik S, Kalin L. Flood flow forecasting using ANN, ANFIS and regression models. Neural Computing and Applications. 2014 Jul;25:25-37.
[25] Chang FJ, Chang YT. Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. Advances in water resources. 2006 Jan 1;29(1):1-0.
[26] Mosavi A, Lopez A, Varkonyi-Koczy AR. Industrial applications of big data: State of the art survey. InRecent Advances in Technology Research and Education: Proceedings of the 16th International Conference on Global Research and Education Inter-Academia 2017 16 2018 (pp. 225-232). Springer International Publishing. Industrial Applications of Big Data: State of the Art Survey.
[27] World Meteorological Organization, TECHNICAL REGULATIONS: Volume III - Hydrology, no. 2. 2006.

[28] Meshram SG, Meshram C, Santos CA, Benzougagh B, Khedher KM. Streamflow prediction based on artificial intelligence techniques. Iranian Journal of Science and Technology, Transactions of Civil Engineering. 2022 Jun;46(3):2393-403.
[29] Araghinejad S, Azmi M, Kholghi M. Application of artificial neural network ensembles in probabilistic hydrological forecasting. Journal of Hydrology. 2011 Sep 15;407(1-4):94-104.
[30] Araujo R, Valenca M, Fernandes S. A new approach of fuzzy neural networks in monthly forecast of water flow. InAdvances in Computational Intelligence: 13th International Work-Conference on Artificial Neural Networks, IWANN 2015, Palma de Mallorca, Spain, June 10-12, 2015. Proceedings, Part I 13 2015 (pp. 576-586). Springer International Publishing.
[31] Tantanee S, Patamatammakul S, Oki T, Sriboonlue V, Prempree T. Coupled wavelet-autoregressive model for annual rainfall prediction. Journal of Environmental Hydrology. 2005;13.
[32] Ravansalar M, Rajaee T, Kisi O. Wavelet-linear genetic programming: a new approach for modeling monthly streamflow. Journal of Hydrology. 2017 Jun 1;549:461-75.
[33] Tan QF, Lei XH, Wang X, Wang H, Wen X, Ji Y, Kang AQ. An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. Journal of Hydrology. 2018 Dec 1;567:767-80.
[34] Yaseen ZM, Ebtehaj I, Bonakdari H, Deo RC, Mehr AD, Mohtar WH, Diop L, El-Shafie A, Singh VP. Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model. Journal of Hydrology. 2017 Nov 1;554:263-76.
[35] Ouyang Q, Lu W, Xin X, Zhang Y, Cheng W, Yu T. Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction. Water resources management. 2016 May;30:2311-25.
[36] Prasad R, Deo RC, Li Y, Maraseni T. Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. Atmospheric Research. 2017 Nov 15;197:42-63.
[37] Zhu S, Zhou J, Ye L, Meng C. Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River, China. Environmental Earth Sciences. 2016 Mar;75:1-2.
[38] Tayyab M, Zhou J, Adnan R, Zeng X. Application of artificial intelligence method coupled with discrete wavelet transform method. Procedia Computer Science. 2017 Jan 1;107:212-7.
[39] Mekanik F, Imteaz MA, Talei A. Seasonal rainfall forecasting by adaptive network-based fuzzy inference system (ANFIS) using large scale climate signals. Climate dynamics. 2016 May;46:3097-111.
[40] Cannas B, Fanni A, Sias G, Tronci S, Zedda MK. River flow forecasting using neural networks and wavelet analysis. InGeophys. Res. Abstr 2005 (Vol. 7, p. 08651).
[41] Shi B, Hu CH, Yu XH, Hu XX. New fuzzy neural network–Markov model and application in mid-to long-term runoff forecast. Hydrological Sciences Journal. 2016 Apr 25;61(6):1157-69.
[42] Yaseen ZM, Awadh SM, Sharafati A, Shahid S. Complementary data-intelligence model for river flow simulation. Journal of Hydrology. 2018 Dec 1;567:180-90.
[43] Anupam S, Pani P. Flood forecasting using a hybrid extreme learning machine-particle swarm optimization algorithm (ELM-PSO) model. Modeling Earth Systems and Environment. 2020 Mar;6(1):341-7.

[44] Hagen JS, Cutler A, Trambauer P, Weerts A, Suarez P, Solomatine D. Development and evaluation of flood forecasting models for forecast-based financing using a novel model suitability matrix. Progress in Disaster Science. 2020 Apr 1;6:100076.
[45] Sahay RR, Srivastava A. Predicting monsoon floods in rivers embedding wavelet transform, genetic algorithm and neural network. Water resources management. 2014 Jan;28:301-17.
[46] Wang ZY, Qiu J, Li FF. Hybrid models combining EMD/EEMD and ARIMA for Long-term streamflow forecasting. Water. 2018 Jun 27;10(7):853.
[47] Liu Y, Wang L, Du S, Zhao L, Liu X. Flood Forecasting Method Based on Improved VMD-FOS-QR-RBL. IEEE Access. 2022 Dec 26.
[48] Khairudin NM, Mustapha NO, Aris TN, Zolkepli MA. In-Depth review on machine learning models for long-term flood forecasting. J. Theor. Appl. Inf. Technol. 2022 May 31;100:1-9.
[49] Mosavi A, Ozturk P, Chau KW. Flood prediction using machine learning models: Literature review. Water. 2018 Oct 27;10(11):1536.
[50] Kitchenham B, Brereton P. A systematic review of systematic review process research in software engineering. Information and software technology. 2013 Dec 1;55(12):2049-75.
[51] Maier HR, Jain A, Dandy GC, Sudheer KP. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environmental modelling & software. 2010 Aug 1;25(8):891-909.
[52] Bontempi G, Ben Taieb S, Le Borgne YA. dMachine learning strategies for time series forecasting. Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures 2. 2013:62-77.
[53] Galelli S, Castelletti A. Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. Hydrology and Earth System Sciences. 2013 Jul 11;17(7):2669-84.
[54] Jabbari A, Bae DH. Application of Artificial Neural Networks for accuracy enhancements of real-time flood forecasting in the Imjin basin. Water. 2018 Nov 11;10(11):1626.
[55] Berkhahn S, Fuchs L, Neuweiler I. An ensemble neural network model for real-time prediction of urban floods. Journal of hydrology. 2019 Aug 1;575:743-54.
[56] Shamim MA, Hassan M, Ahmad S, Zeeshan M. A comparison of artificial neural networks (ANN) and local linear regression (LLR) techniques for predicting monthly reservoir levels. KSCE Journal of Civil Engineering. 2016 Mar;20:971-7.
[57] Muñoz P, Orellana-Alvear J, Willems P, Célleri R. Flash-flood forecasting in an Andean mountain catchment—development of a step-wise methodology based on the random forest algorithm. Water. 2018 Oct 26;10(11):1519.
[58] Chen L, Sun N, Zhou C, Zhou J, Zhou Y, Zhang J, Zhou Q. Flood forecasting based on an improved extreme learning machine model combined with the backtracking search optimization algorithm. Water. 2018 Sep 29;10(10):1362.
[59] Gizaw MS, Gan TY. Regional flood frequency analysis using support vector regression under historical and future climate. Journal of Hydrology. 2016 Jul 1;538:387-98.
[60] Zhou Y, Guo S, Chang FJ. Explore an evolutionary recurrent ANFIS for modelling multi-step-ahead flood forecasts. Journal of hydrology. 2019 Mar 1;570:343-55.

[61] Jhong YD, Chen CS, Lin HP, Chen ST. Physical hybrid neural network model to forecast typhoon floods. Water. 2018 May 13;10(5):632.
[62] Altunkaynak A, Nigussie TA. Prediction of daily rainfall by a hybrid wavelet-season-neuro technique. Journal of Hydrology. 2015 Oct 1;529:287-301.
[63] Mosavi A, Edalatifar M. A hybrid neuro-fuzzy algorithm for prediction of reference evapotranspiration. InRecent Advances in Technology Research and Education: Proceedings of the 17th International Conference on Global Research and Education Inter-Academia–2018 17 2019 (pp. 235-243). Springer International Publishing.
[64] Tayyab M, Zhou J, Dong X, Ahmad I, Sun N. Rainfall-runoff modeling at Jinsha River basin by integrated neural network with discrete wavelet transform. Meteorology and Atmospheric Physics. 2019 Feb;131:115-25.