Hybrid machine learning model based on feature decomposition and entropy optimization for higher accuracy flood forecasting

Nazli Mohd Khairudin [a,1,*], Norwati Mustapha [a,2], Teh Noranis Mohd Aris [a,3], Maslina Zolkepli [a,4]

[a] Faculty of Computer Sciences and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[1] nazmkhair@gmail.com; [2] norwati@upm.edu.my; [3] nuranis@upm.edu.my; [4] masz@upm.edu.my
* corresponding author

ARTICLE INFO

Article history
Received May 31, 2023
Revised July 1, 2023
Accepted July 8, 2023
Available online December 29, 2023

Keywords
Discrete wavelet transform
Empirical mode decomposition
Ensemble EMD
Entropy
Mutual information

ABSTRACT

The advancement of the machine learning model has widely been adopted to provide flood forecasts. However, the model must deal with the challenges of determining the most important features to be used in flood forecasts with high-dimensional non-linear time series when involving data from various stations. Decomposition of time-series data such as empirical mode decomposition, ensemble empirical mode decomposition, and discrete wavelet transform are widely used for optimization of input; however, they have been done for single dimension time-series data which are unable to determine relationships between data in high dimensional time series. In this study, machine learning models, which are Artificial Neural Network (ANN), Adaptive Neuro Fuzzy Inferences System (ANFIS), and Long-Short Term Memory (LSTM), are integrated with decomposition methods to provide a hybrid model to forecast the monthly water level using monthly rainfall data from Kelantan River Basin. To effectively select the best rainfall data from the multi-stations that provide higher accuracy, these rainfall data are analyzed with entropy called Mutual Information that measures the uncertainty of random variables from various stations. Mutual information acts as an optimization method to help the researcher select the appropriate features to score a higher accuracy for the model. The experimental evaluations using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Nash-Sutcliffe Efficiency (NSE) proved that the hybrid machine learning model based on the feature decomposition and ranked by Mutual Information can increase the accuracy of water level forecasting. This outcome will help citizens manage the risk of floods.

1. Introduction

The advancement of machine learning model has proved it potential in providing accurate forecasts for hydrological data such as rainfall. In many literatures, the development of machine learning models for long-term flood forecasting such as monthly water level can be categorized as single or hybrid model. Single models have been applied only one machine learning method to produce the forecast. Among the widely used single machine learning models are Support Vector regression (SVR), Artificial Neural Networks (ANN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). Hybrid models also often used in long-term flood forecasting.

Hybrid models have become a growing interest to the researcher in developing model for long-term flood forecast. Multiple machine learning techniques are integrated, combined, or used in ensembles to create models. Hybrid machine learning models can also be developed by combining input optimization into the models that can give forecasts with acceptable performance and accuracy. In certain studies, hybrid machine learning is used with more traditional techniques like physical approaches to improve the performance of the models.

In most machine learning model either it is single or hybrid, historical data is the main input component. Original historical data may be hard to come by, inaccurate, noisy, or unbalanced, making it inappropriate to use it exclusively for forecasting. [1], [2]. The implementation of pre-processing method has been proven to enhance the forecasting model performance [3]. The choice of input is crucial because it affects the forecast model's precision and accuracy [4]. Several pre-processing methods can be found in the literature and the most frequent use in recent studies are Empirical Wavelet Decomposition (EMD), Ensemble Empirical Wavelet Decomposition (EEMD), and Wavelet Transform (WT).

In a wavelet decomposition like Discrete Wavelet Transform (DWT), time series data were decomposed into a shifted and scaled version of a wavelet known as a mother wavelet [5]. It is a convenient method that able to analyses time-series variation that gives information on time and frequency domains of signal. DWT is more efficient than calculating wavelet coefficient at each possible scale in the continuous wavelet transform that takes more of process work [6]. Although DWT is a great pre-processing technique, it is sensitive in the selection of mother wavelet [7], [8]. The decomposition technique of DWT has been proved to improved forecast performance [5], [6], [9], However, in certain instances, the decomposition of this time-series data did not result in any appreciable differences in the forecasting precision. This is because wavelet analysis is based on the convolution of the signal and the filter [10]. The drawback of DWT is the occurrence of shift variance in which may cause by the down sampling [11].

Non-linear signal is decomposed into intrinsic mode functions (IMFs) and one residual component using the Empirical Mode Decomposition (EMD) [12]. Non-stationary and non-linear time series data can be decomposed using EMD [13]. EMD is superior to DWT since it is fully self-adaptive and doesn't require a predetermined basis function. The disadvantage of EMD is that mode mixing occurs frequently [4]. EEMD decomposing technique adds finite noise to the signal that can address the mode mixing problem of EMD [3]. The use of EEMD has effectively improved the forecasting model [14]. It is also found that the use of different length time series data in EEMD could produce various performances of the model in which the decomposition and model must be updated whenever new information is added [4].

Decomposition of input in the pre-processing phase of any modelling has a significant role in helping to select the most dominant input for the model. Pre-processing the input has the advantages in producing better forecast accuracy [15]. In many cases, decomposition of input is done within the single dimension of data, while when multi-dimensional data is considered, it is very challenging to determine which of these data are the most suitable in producing a higher forecast accuracy. Decomposing single dimension time-series data will only optimize specific input. When multi-dimensional data is considered, it is very time consuming and hard to determine which of these data can produce a high accuracy forecast. In this case, it is crucial to analyze the relationship between these data to accommodate the best combination of data to be used in the machine learning model. As time series that consist of high dimensional data can be inefficient to be use in machine learning [16], it is important that the best input are selected to provide higher accuracy.

This study aims to improve the accuracy of a hybrid machine learning model to forecast monthly water level based on decomposition of multi-dimensional input using Empirical Model Decomposition (EMD), Ensemble Empirical Model Decomposition (EEMD), and Discrete Wavelet Transform (DWT). These inputs are analyzed on their relationship and ranked using entropy called Mutual Information to find the optimize input. Mutual Information quantity the information shared between two random variables in which it expresses how the value of one variable aids in forecasting the value of the other. The performances of these three models then were assessed in terms of accuracy. An accurate monthly water level forecasting is crucial to protect the downstream cities and communities from the flood hazard, thus minimizing any possible loss and damage perceived by flood victims. Monthly water level can be a great indication to the authorities in managing the risk of flood and helping people in the evacuation process as an early warning can be assigned and disseminate to the citizen.

2. Method

2.1. Areas of Study and Dataset

Monthly average rainfall data from eight rainfall stations are gathered from Kelantan River Basin. Data are provided by the Department of Irrigation and Drainage Malaysia, an agency that responsible in managing the water resources in Malaysia. These eight stations are located along two main tributaries of Sungai Kelantan which are Lebir River and Galas River. Six of the eight stations are along Lebir River which are Gunung Gagau, Kuala Koh, Kampung Aring, Kampung Lalok, Kampung Tualang and Kuala Krai. Meanwhile the rest are along the Galas River which is Dabong and Limau Kasturi. These stations are located at the upper stream and the water level station of Kuala Krai is at the downstream. The forecasting model will analyze these data to forecast monthly water level in the Kuala Krai water level station.

Dataset used in this study is from April 2011 to November for all the stations. The data for each station is divided into two datasets. First is the training dataset which covers 75% of the whole data and the rest 25% is for the test datasets.

2.2. Feature Decomposition and Optimization Model Development

Input data contribute on how machine learning model behave. In this study, input is optimized with a proposed method based on decomposition of signals and entropy called Mutual Information. Decomposition of time-series data has been widely used as the pre-processing method of the original data. In many cases, decomposition methods have been used in single dimension data such as water level [17], runoff [18], and streamflow [15], but the effect of this method is lacked to be known when these decomposed single-series data is being part of the features of a high dimensional dataset. The effect of such conditions is an important aspect to be studied as it provides the machine learning models with the most valuable features which can determine the model performance. Entropy is firstly introduced by [19] that measures the uncertainty or variability in random variables. The entropy has been defined as :

$$H(X) = H(P) = -\sum_{i=1}^{N} p(x_i) \log [(p(x_i)] \tag{1}$$

where $H(X)$ is the entropy function for random variable X, N is the size of the time-series data, $p(x)$ is the probability density function (PDF) for variable i, and log [] is the log-part function for the PDF. In hydrological forecasting, entropy has been adopted to produce a robust and reliable flood forecasting at the upper Yangtze River was [20].

Data from eight rainfall stations are decomposed using different decomposition methods namely Empirical Mode Decomposition (EMD), Ensemble Mode Decomposition (EEMD), and Discrete Wavelet Transform (DWT). Decomposing such inputs will give in-depth knowledge about the data in term of time and space. These methods are selected due to their high adaptive nature and inherent to the characteristic of the non-linear data. They allowed the extraction of useful features from non-stationary, non-linear data, allowing machine learning algorithms to efficiently learn and represent the complex patterns and dynamics present in the data. Empirical Mode Decomposition (EMD) decomposed signal into various components called Intrinsic Mode Functions (IMF) and residual that preserve the time-spectrum [12]. EMD decomposed the original signal by using shifting process. As hydrological data such as rainfall or runoff are non-linear and non-stationary, EMD can be used to decompose the original data. The applications of EMD in hydrological forecast can be found in several studies such as in [21] that used it with Encoder Decoder LSTM model to forecast monthly streamflow for the Yangtze River. EMD also has proven to increase the performance of hydrological forecasting when coupled with Seasonal-Trend decomposition using Loess (STL) and ANN [13].

Ensemble Empirical Mode Decomposition (EEMD) is a decomposition method that address the drawbacks of mode mixing problem in EMD by adding finite noise to the signals. Mode mixing problems have the potential to generate a decomposed signal that unable to represent the original characteristics of the data [13]. As a result, by include noise in the signals, a consistent reference backdrop of the time-frequency space is created, allowing varied signal scales to be automatically projected onto the proper reference scales. The integration of EEMD with Artificial Neural Network (ANN) with machine learning model such as to forecast runoff has indicated that it can improve the accuracy during the flood season [4].

Discrete Wavelet Transform (DWT) is a method in which the signal is discretely sampled. DWT decomposed time series data into a shifted and scaled version of a wavelet called mother wavelet [5]. It provides the information for both time and frequency domain of nonstationary signal [22]. As hydrological data might be recorded at a discrete time rather than in continuous manner, DWT is a suitable method to be used in pre-processing of this kind of data [23]. The decomposition level of DWT is a subjective matter and there is no direct method to determine it [24]. As decomposition level can determine model performance, the optimal level must be selected. An empirical equation (2) adapted by [25] has been used in this study to determine the decomposition level.

$$L = int[\log(N)] \tag{2}$$

where L is the level of decomposition, N is the length of time-series data, and int [] is the integer-part function. The Daubechies wavelet is used as the mother wavelet with three vanishing moments (db3) which provided the high vanishing moments for given support width [22]. Total number of IMF and coefficient from the decomposition method for all rainfall stations are shown in Table 1.

Table 1. Total IMF and coefficient for each decomposition method

Stations	**Decomposition Method**		
	EMD	***EEMD***	***DWT***
	Total IMFs	***Total IMFs***	***Total Coefficient***
Gunung Gagau (GG)	IMF1 – IMF5	IMF1 – IMF6	cA2, cD2, cD1
Kuala Koh (KK)	IMF1 – IMF6	IMF1 – IMF6	cA2, cD2, cD1
Kampung Aring (KA)	IMF1 – IMF5	IMF1 – IMF6	cA2, cD2, cD1
Tualang (T)	IMF1 – IMF5	IMF1 – IMF6	cA2, cD2, cD1
Kampung Lalok (KL)	IMF1 – IMF5	IMF1 – IMF6	cA2, cD2, cD1
Kuala Krai (KKr)	IMF1 – IMF5	IMF1 – IMF5	cA2, cD2, cD1
Limau Kasturi (LK)	IMF1 – IMF6	IMF1 – IMF6	cA2, cD2, cD1
Dabong (D)	IMF1 – IMF6	IMF1 – IMF6	cA2, cD2, cD1

EMD and EEMD provided each of the time series data IMFs and residual components in which they may be strong or weakly correlated with the original data. To select the most correlated set of data, Spearman correlation coefficient is used. Spearman correlation method is used as it is suitable for time-series data that is not normally distributed [26]. In DWT, Spearman correlation coefficient has also been used to select the most correlated sub time series with the original data between the approximation coefficient (cA) and the details coefficient (cD) [23], [27]. Spearman cross-correlation function (p-value) can evaluate the level of relationship between the decomposed data and original data [23], [26], [27]. The function can be defined in the below equation:

$$p - value = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3}$$

where $p-value$ is the Spearman correlation coefficient, d_i is the ranking difference of each observation, and n is the number of observations.

Table 2 presents the selected IMF and coefficients with their respective $p-value$ of every station. For example, IMF4 and cA2 produce the highest $p-value$ using EMD, EEMD and DWT respectively for Gunung Gagau.

Table 2. Spearman correlation coefficient value for selected IMF and coefficient

Stations	Decomposition Method					
	EMD		*EEMD*		*DWT*	
	Selected IMF	*p-value*	*Selected IMF*	*p-value*	*Selected Coefficient*	*p-value*
Gunung Gagau (GG)	IMF4	0.573	IMF4	0.540	cA2	0.753
Kuala Koh (KK)	IMF2	0.543	IMF2	0.449	cA2	0.664
Kampung Aring (KA)	IMF1	0.674	IMF1	0.653	cA2	0.552
Tualang (T)	IMF2	0.580	IMF2	0.520	cA2	0.642
Kampung Lalok (KL)	IMF1	0.697	IMF1	0.633	cD1	0.581
Kuala Krai (KKR)	IMF1	0.568	IMF1	0.537	cA2	0.612
Limau Kasturi (LK)	IMF3	0.463	IMF3	0.524	cA2	0.711
Dabong (D)	IMF1	0.630	IMF1	0.637	cA2	0.626

There is a representative feature (selected IMF and coefficient) for every decomposition method from each station. There are 8 features in total for each decomposition method to be used in forecasting water level of Kuala Krai. This process will generate optimized datasets with the structure as in Fig. 1. Feature 1 to Feature 8 are the selected sub time series for each of the rainfall stations.

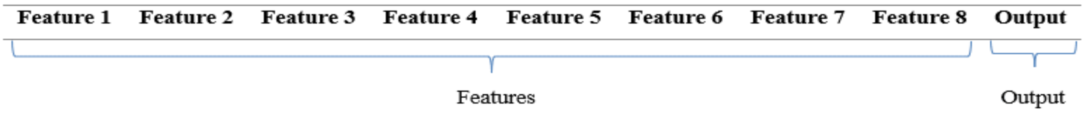

Fig. 1. Dataset structure for each decomposition method

Although the highest correlated features were selected from the decomposed data, uncertainty still can occur in forecasting due to the variability of data. Therefore, the relationship between these data must be established to select only the features that can provide higher accuracy to be used in the model. This kind of relationship can be analyzed using entropy called Mutual Information. It measured the dependency between random variables and measuring the dispersion of information. Additionally, it managed a robust nonlinear relationship between input and output that might be used to determine superior input in a machine learning model with nonlinear time-series. Mutual Information between two random variables A and B can be defined as [28]:

$$MI(A,B) = H\ (A) + H(B) - H\ (A,B) \tag{4}$$

where $H(A)$ and $H(B)$ are the entropy of A and B, and the joint entropy of $H(A,B)$ is:

$$H(A,B) = -\sum_{a\epsilon A}\sum_{b\epsilon B}{}^{PAB}(a,b)logP_{AB}(a,b) \tag{5}$$

where a and b is the specific value of A and B, respectively $p(a,b)$ is the joint probability of these values occurring together.

Table 3 presents the value of Mutual Information of each station and its rank from more to less dominant features. It shows that the most dominant features of EMD, EEMD and DWT are coming from Kuala Koh station and Gunung Gagau station, respectively. This has given us the ability to select the optimized datasets which consist of the most dominant features to be used in the machine learning model:

Table 3. Mutual Information rank of optimization methods

Optimization Method	More to Less Dominant							
Opt1*	KK	T	KL	GG	KKR	D	LK	KA
MI value	0.2996	0.2095	0.1648	0.1374	0.1167	0.0848	0.0552	0
Opt2 **	GG	KKR	KL	KK	D	LK	T	KA
MI value	0.3467	0.2632	0.232	0.2029	0.1592	0.1476	0.1388	0.0295
Opt3 ***	GG	KKR	D	T	KK	KA	KL	LK
MI value	0.4191	0.3654	0.2857	0.2559	0.1641	0.1158	0.09	0.0587

*Opt1 = optimization with integration of EMD decomposition with entropy, ** Opt2 = optimization with integration of EEMD decomposition with entropy and *** Opt3 = optimization with integration of DWT decomposition with entropy

To forecast the monthly water level of Kuala Krai, these optimized datasets were the input for the machine learning models. Hybrid machine learning models were introduced by integrating the decomposition and optimization method with three machine learning model namely Artificial Neural Network (ANN), Adaptive Neuro Fuzzy Inferences System (ANFIS), and Long-Short Term Memory (LSTM). These three type of machine learning models have been chosen as they have proved their ability to produce accurate forecast using non-linear and non-stationary hydrological time series data [29]–[31]. Fig. 2 present the details flow of the model development.

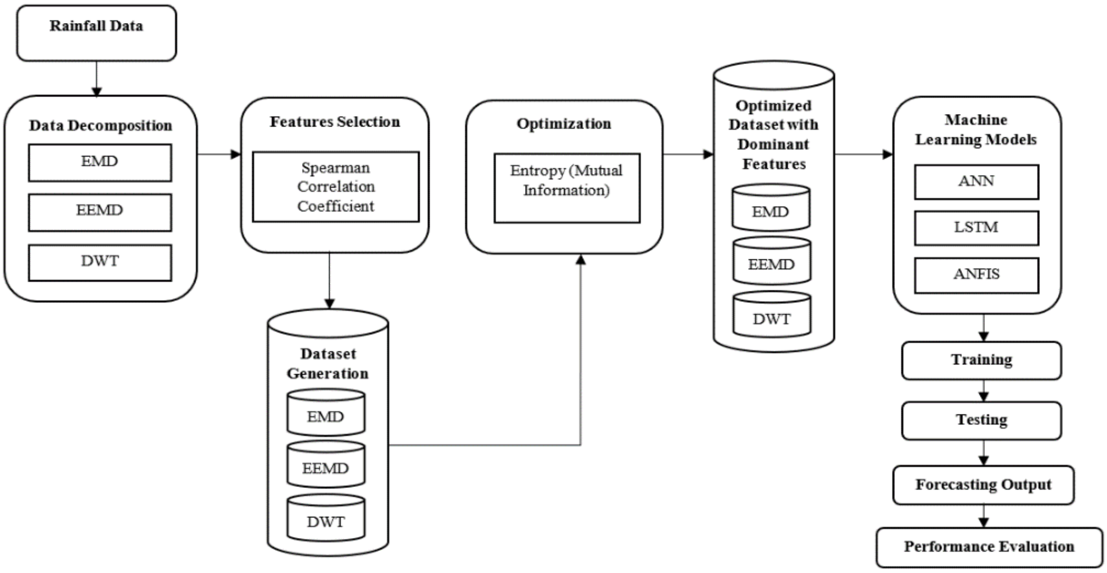

Fig. 2. Flow of the hybrid model development with optimization method

Artificial Neural Networks (ANNs) are modelled after the structure and operation of the human brain. ANNs are made up of layers of neurons, which are interconnected nodes [32]. ANNs can minimize the discrepancy between expected and actual outputs as they learn from data. After being trained, ANNs can categorize or predict new inputs using the data's discovered patterns. ANFIS works by using neural networks and fuzzy logic [33]. It uses a learning algorithm that modifies these rules' parameters in accordance with training data. ANFIS uses a forward-pass technique, where membership function degrees are computed, and input values are fuzzified. The final output is produced by aggregating the findings after these degrees have been propagated through the fuzzy inference procedures. LSTM (Long Short-Term Memory) were developed to address the vanishing gradient problem and identify long-term dependencies in sequential data [2]. LSTM uses memory cells and gating mechanisms. It has input, forget, and output gates that control the information flow. Due to its good retention and use of pertinent information over extended sequences, LSTM is well suited for sequential data processing tasks.

All model performance in this study is evaluated using statistical method often known as "goodness of fit" [34]. To comprehensively assess the model, the forecast result is evaluated against the original value using three measurements, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Nash-Sutcliffe Efficiency (NSE) as in (6), (7), and (8).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i}^{N} = 1(y_i - \hat{y})^2} \tag{6}$$

$$MAE = \frac{1}{N} \sum_{i}^{N} = 1|y_i - \hat{y}| \tag{7}$$

where y_i is the original value at period i, $\hat{y}_i$ is the forecasted value at the period of i, and N denotes the number of the sample.

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_m^t - Q_o^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q_o})^2} \tag{8}$$

where Q_0 is the original value, and Q_m is the forecasted value. Q_o^t is the original value at time t. The higher value of NSE indicate more powerful forecast model.

3. Results and Discussion

The purpose of the hybrid model is to forecast water level at Kuala Krai station in the downstream by using rainfall data from the various stations located along the river at the upper stream. As Lebir river and Galas River are the main tributaries of Sungai Kelantan, it is important to analyze the effect of the rainfall that occurred in stations alongside these two rivers towards the Kuala Krai water level. Each model is trained with 75% of the dataset and test with 25% of the dataset. The performance of the model then is assessed by using statistical methods of RMSE, MAE and, NSE. To get the best performance for every model, the models are run repeatedly by using every feature in the rank with the lowest rank is eliminate in each cycle. The cycle stopped after ranked one only feature is run through the models. The best performance results for all these models are presented in Table 4 with their respective number of dominant features. To ease the identification of the hybrid model, each model is labelled by its machine learning type and the optimization method. For example, ANN-Opt1 represent the ANN model that used optimization with EMD decomposition, while Opt2 is optimization with EEMD decomposition and Opt3 is optimization with DWT decomposition. The machine learning models is compared with each other, and comparison is extended with the model that used original data without optimization.

Table 4. Performance measurement of hybrid model with optimization

Model	RMSE	Number of Features	MAE	Number of Features	NSE	Number of Features
ANN-Original Data	1.5630	8	1.1560	8	-0.1319	8
ANN-Opt1	1.2862	6	0.9524	6	0.2335	6
ANN-Opt2	1.3478	8	1.0288	8	0.1584	8
ANN-Opt3	1.1302	7	0.9181	7	0.4082	7
LSTM-Original Data	2.0334	8	1.4090	8	-0.9158	8
LSTM-Opt1	1.0447	4	0.7519	4	0.4943	4
LSTM-Opt2	1.2259	4	0.8253	3	0.3037	4
LSTM-Opt3	**0.9356**	**3**	**0.6742**	**3**	**0.5945**	**3**
ANFIS-Original Data	3.3771	8	1.7927	8	-0.0066	8
ANFIS-Opt1	1.3058	1	0.8688	1	0.1850	4
ANFIS-Opt2	1.2184	4	0.7769	4	-0.2919	5
ANFIS-Opt3	1.0046	3	0.6885	3	-0.1619	6

The performance measurement trend is presented in Fig. 3 for RMSE, Fig. 4 for MAE, and Fig. 5 for NSE. Fig. 6 presents the comparison of forecast data and observed data for the testing datasets of LSTM-Opt3 model that recorded the lowest of RMSE and MAE, and highest performance of NSE.

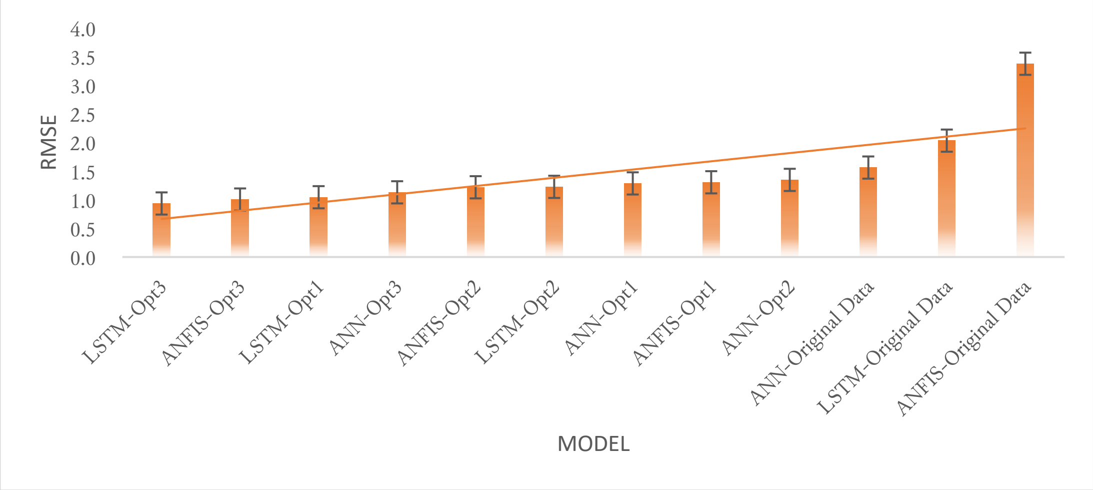

Fig. 3. Lowest to Highest RMSE Value for All Models

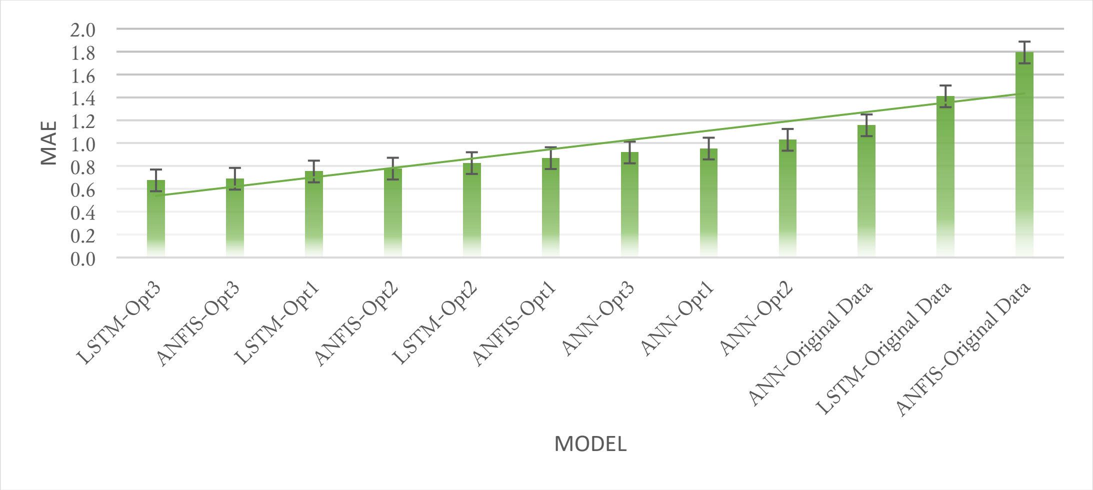

Fig. 4. Lowest to Highest MAE Value for All Models

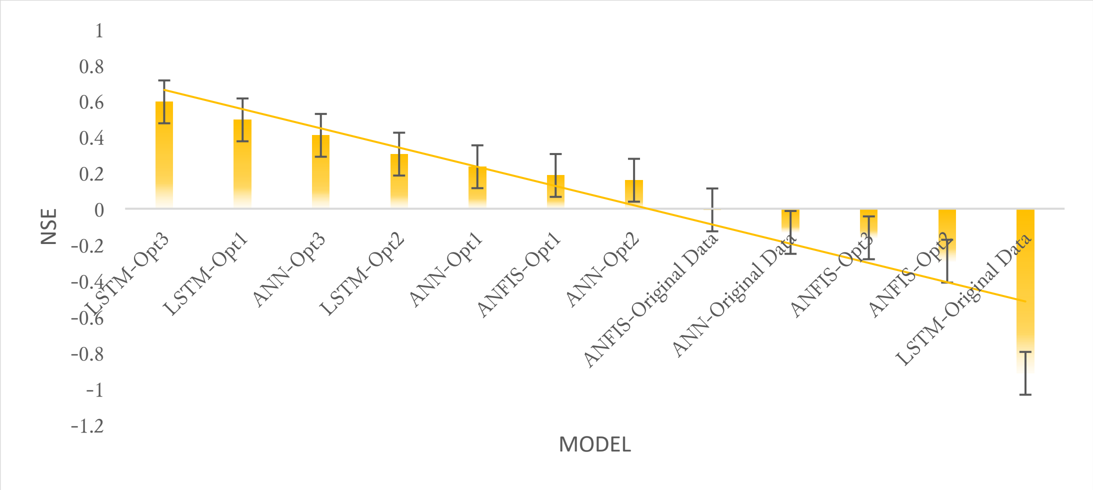

Fig. 5. Highest to Lowest NSE Value for All Models

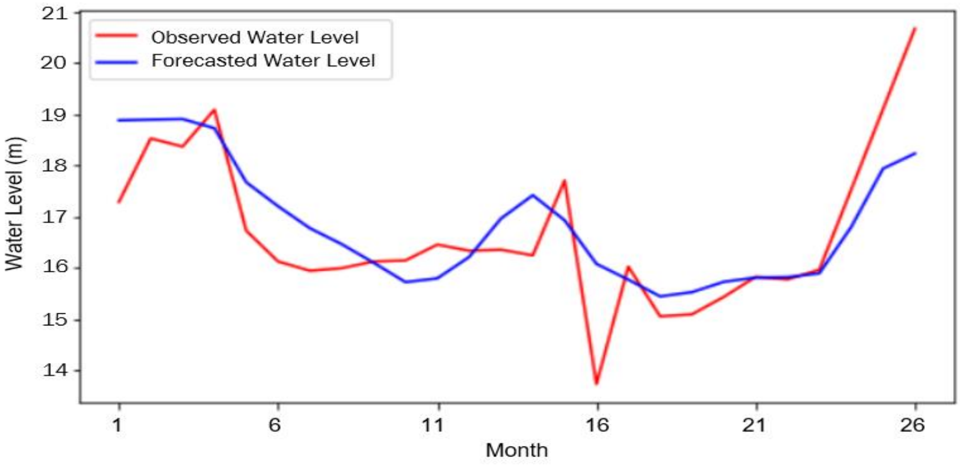

Fig. 6. Forecasted Water Level Compared with Observed Water Level for LSTM-Opt3 Model

The performance measurement shows that most of the hybrid model has achieved better performance in terms of RMSE, NSE, and MAE than the model with original data. The best performance is achieved by LSTM-Opt3 model when it recorded the lowest value in RMSE and MAE while the highest value in NSE. Lowest value of RMSE and MAE is an indication of improved forecast in which minimal error are produced compared to another model.

Although LSTM-Opt3 model has achieved a great result, however according to the forecast pattern in Fig. 6, at some points of the data where a major sudden spike occurred in original data, the forecast model seems not to behave following the trend. This may be due to the model might not have enough prior knowledge to correctly forecast the subsequent water levels whenever sudden changes happened in the input data, such as a major increase in rainfall intensity. Spearman coefficient correlation is conducted between the observed value and forecasted value for the LSTM-Opt3 model which resulted the value of 0.789 that indicates a positive and strong relationship of water level towards the rainfall trends. For flood forecast models to work, the connection between water level and rainfall must be positive. When rain falls heavily or lasts a long time, there is more water runoff, which raises the water levels in rivers and streams. Forecast model can anticipate flood events by examining past patterns of precipitation and the responses of the river levels. This knowledge is crucial for delivering prompt alerts, putting emergency action plans into action, and reducing possible flooding damage.

In the case of hybrid model using ANFIS, it is revealed that all hybrid model using ANFIS has achieved the lowest value in RMSE and MAE than the model with original data. In contrary, only ANFIS-Opt2 managed to perform better than the ANFIS model with original data in term of NSE. Although the model has less energy, it can still provide a better accuracy. All hybrid model based on ANN has outperformed the ANN model with original data. The improvement of the model's accuracy is varied among the models. Even so, the real and significant improvement can be found in ANFIS based model in which the value of RMSE and MAE are greatly reduced. Hybrid model of ANN, ANFIS and LSTM with optimization has proven to produce better modelling performance. This study also indicates that the best performance of RMSE, NSE, and MAE are achieved when around 38%-40% of features with the highest rank are used in the machine learning model. The model has utilized the advantage of reducing the dimension dataset produce by optimization method to produce a more accurate result.

4. Conclusion

In this study, hybrid models are developed by combining machine learning model with optimizations method. The optimizations method is based on integration of decomposition signals with entropy called Mutual Information. They are not only producing datasets that have the strongest relationship with the original data but also can reduces the dimension of features used in the models. The adaptation of the optimization methods in machine learning model has demonstrated a significant improvement in the forecast performance for all the developed hybrid models. Machine learning models LSTM-Opt3 has shown an impressive result of being the best performance model. It reduced the RMSE value by 54%, MAE value by 52% and increased the NSE value by more than 100% compared to the original data. The advantage of the proposed hybrid model is that by integrating the optimization methods with machine learning, it will provide the researcher or authorities a dominant input to the forecast model and improving its performance. It will also provide an early warning to the citizen and relevant agencies that can help them in triggering any action such evacuation plan to minimize the impact of the flood to lives, infrastructure, and crops. Although the proposed hybrid model has shown significant result in term of the forecast performance, but it can be difficult to guarantee that a hybrid model performs effectively across various datasets, contexts, or time periods. When used in new situations, the hybrid models may be sensitive to variations in data distribution and need to be modified or retrain. For future research, it is valuable to investigate how a hybrid approach might be used to improve flood forecasting in areas with a lack of data and observations or incomplete data that make it difficult to use conventional modeling techniques.

Acknowledgment

We would like to acknowledge the Department of Irrigation and Drainage Malaysia under the Ministry of Environment and Water Malaysia for providing us with the data for the purpose of this study.

Declarations

Author contribution. Conceptualization, Nazli Mohd Khairudin; Data curation, Nazli Mohd Khairudin; Funding ac-quisition, Nazli Mohd Khairudin; Investigation, Nazli Mohd Khairudin; Methodology, Nazli Mohd Khairudin; Resources, Teh Noranis Mohd Aris and Maslina Zolkepli; Software, Nazli Mohd Khairudin; Supervision, Norwati Mustapha; Visualization, Norwati Mustapha; Writing – original draft, Nazli Mohd Khairudin; Writing – review & editing, Teh Noranis Mohd Aris and Maslina Zolkepli
Funding statement. This research received no external funding.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

Data and Software Availability Statements

3rd Party Data. Restrictions apply to the availability of these data. Data was obtained from Department of Irrigation and Drainage Malaysia and are available with the permission of Department of Irrigation and Drainage Malaysia.

References

[1] N. Khairuddin, A. Z. Aris, A. Elshafie, T. Sheikhy Narany, M. Y. Ishak, and N. M. Isa, “Efficient forecasting model technique for river stream flow in tropical environment,” *Urban Water J.*, vol. 16, no. 3, pp. 183–192, 2019, doi: 10.1080/1573062x.2019.1637906.

[2] Y. Zhou, S. Guo, and F. J. Chang, “Explore an evolutionary recurrent ANFIS for modelling multi-step-ahead flood forecasts,” *J. Hydrol.*, vol. 570, no. December, pp. 343–355, 2019, doi: 10.1016/j.jhydrol.2018.12.040.

[3] Y. Yu, H. Zhang, and V. P. Singh, “Forward prediction of Runoffdata in data-scarce basins with an improved ensemble empirical mode decomposition (EEMD) model,” *Water (Switzerland)*, vol. 10, no. 4, pp. 1–15, 2018, doi: 10.3390/w10040388.

[4] Q. F. Tan *et al.*, “An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach,” *J. Hydrol.*, vol. 567, pp. 767–780, 2018, doi: 10.1016/j.jhydrol.2018.01.015.

[5] M. Tayyab, J. Zhou, X. Dong, I. Ahmad, and N. Sun, “Rainfall-runoff modeling at Jinsha River basin by integrated neural network with discrete wavelet transform,” *Meteorol. Atmos. Phys.*, vol. 131, no. 1, pp. 115–125, 2019, doi: 10.1007/s00703-017-0546-5.

[6] F. F. Li, Z. Y. Wang, and J. Qiu, "Long-term streamflow forecasting using artificial neural network based on preprocessing technique," *J. Forecast.*, vol. 38, no. 3, pp. 192–206, 2019, doi: 10.1002/for.2564.

[7] K. Roushangar, F. Alizadeh, and V. Nourani, "Improving capability of conceptual modeling of watershed rainfall–runoff using hybrid wavelet-extreme learning machine approach," *J. Hydroinformatics*, vol. 20, no. 1, pp. 100–116, 2018, doi: 10.2166/hydro.2017.011.

[8] C. P. Dautov and M. S. Ozerdem, "Wavelet transform and signal denoising using Wavelet method," *26th IEEE Signal Process. Commun. Appl. Conf. SIU 2018*, pp. 1–4, 2018, doi: 10.1109/SIU.2018.8404418.

[9] G. Zuo, J. Luo, N. Wang, Y. Lian, and X. He, "Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting," *J. Hydrol.*, vol. 585, no. December 2019, p. 124776, 2020, doi: 10.1016/j.jhydrol.2020.124776.

[10] V. Nourani, G. Andalib, and F. Sadikoglu, "Multi-station streamflow forecasting using wavelet denoising and artificial intelligence models," *Procedia Comput. Sci.*, vol. 120, pp. 617–624, 2017, doi: 10.1016/j.procs.2017.11.287.

[11] N. M. Khairudin, N. Mustapha, T. N. M. Aris, and M. Zolkepli, "in-Depth Review on Machine Learning Models for Long-Term Flood Forecasting," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 10, pp. 3360–3378, 2022. [Online]. Available at: https://www.jatit.org/volumes/Vol100No10/19Vol100No10.pdf.

[12] T. Xie, G. Zhang, J. Hou, J. Xie, M. Lv, and F. Liu, "Hybrid forecasting model for non-stationary daily runoff series: A case study in the Han River Basin, China," *J. Hydrol.*, vol. 577, no. April, p. 123915, 2019, doi: 10.1016/j.jhydrol.2019.123915.

[13] F. F. Li, Z. Y. Wang, X. Zhao, E. Xie, and J. Qiu, "Decomposition-ANN Methods for Long-Term Discharge Prediction Based on Fisher's Ordered Clustering with MESA," *Water Resour. Manag.*, vol. 33, no. 9, pp. 3095–3110, 2019, doi: 10.1007/s11269-019-02295-8.

[14] M. Rezaie-Balf, S. F. Nowbandegani, S. Z. Samadi, H. Fallah, and S. Alaghmand, "An ensemble decomposition-based artificial intelligence approach for daily streamflow prediction," *Water (Switzerland)*, vol. 11, no. 4, pp. 1–31, 2019, doi: 10.3390/w11040709.

[15] Z. M. Yaseen, S. M. Awadh, A. Sharafati, and S. Shahid, "Complementary data-intelligence model for river flow simulation," *J. Hydrol.*, vol. 567, no. October, pp. 180–190, 2018, doi: 10.1016/j.jhydrol.2018.10.020.

[16] M. M. M. Fuad, "A differential evolution optimization algorithm for reducing time series dimensionality," *2016 IEEE Congr. Evol. Comput. CEC 2016*, pp. 249–254, 2016, doi: 10.1109/CEC.2016.7743802.

[17] N. T. N. Anh, N. Q. Dat, N. T. Van, N. N. Doanh, and N. Le An, "Wavelet-Artificial Neural Network Model for Water Level Forecasting," *Proc. 2018 3rd IEEE Int. Conf. Res. Intell. Comput. Eng. RICE 2018*, pp. 1–6, 2018, doi: 10.1109/RICE.2018.8509064.

[18] W. jing Niu *et al.*, "Forecasting reservoir monthly runoff via ensemble empirical mode decomposition and extreme learning machine optimized by an improved gravitational search algorithm," *Appl. Soft Comput. J.*, vol. 82, p. 105589, 2019, doi: 10.1016/j.asoc.2019.105589.

[19] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–656, 1948, doi: 10.1002/j.1538-7305.1948.tb00917.x.

[20] L. Chen *et al.*, "Flood forecasting based on an improved extreme learning machine model combined with the backtracking search optimization algorithm," *Water (Switzerland)*, vol. 10, no. 10, pp. 1-17, 2018, doi: 10.3390/w10101362.

[21] D. Liu, W. Jiang, L. Mu, and S. Wang, "Streamflow Prediction Using Deep Learning Neural Network: Case Study of Yangtze River," *IEEE Access*, vol. 8, pp. 90069–90086, 2020, doi: 10.1109/ACCESS.2020.2993874.

[22] M. Hammad, M. Shoaib, H. Salahudin, M. A. I. Baig, M. M. Khan, and M. K. Ullah, "Rainfall forecasting in upper Indus basin using various artificial intelligence techniques," *Stoch. Environ. Res. Risk Assess.*, vol. 35, no. 11, pp. 2213–2235, 2021, doi: 10.1007/s00477-021-02013-0.

[23] M. Kumar and R. R. Sahay, "Wavelet-genetic programming conjunction model for flood forecasting in rivers," *Hydrol. Res.*, vol. 49, no. 6, pp. 1880–1889, 2018, doi: 10.2166/nh.2018.183.

[24] H. Tongal and M. J. Booij, "Simulation and forecasting of streamflows using machine learning models coupled with base flow separation," *J. Hydrol.*, vol. 564, pp. 266–282, 2018, doi: 10.1016/j.jhydrol.2018.07.004.

[25] A. Azizpour, M. A. Izadbakhsh, S. Shabanlou, F. Yosefvand, and A. Rajabi, "Estimation of water level fluctuations in groundwater through a hybrid learning machine," *Groundw. Sustain. Dev.*, vol. 15, p. 100687, Nov. 2021, doi: 10.1016/j.gsd.2021.100687.

[26] Q. Chen *et al.*, "Empirical mode decomposition based long short-term memory neural network forecasting model for the short-term metro passenger flow," *PLoS One*, vol. 14, no. 9, pp. 1–18, 2019, doi: 10.1371/journal.pone.0222365.

[27] M. Hejazi, S. A. R. Al-Haddad, Y. P. Singh, S. J. Hashim, and A. F. Abdul Aziz, "ECG biometric authentication based on non-fiducial approach using kernel methods," *Digit. Signal Process. A Rev. J.*, vol. 52, pp. 72–86, 2016, doi: 10.1016/j.dsp.2016.02.008.

[28] N. Lv *et al.*, "A long Short-Term memory cyclic model with mutual information for hydrology forecasting: A Case study in the xixian basin," *Adv. Water Resour.*, vol. 141, no. May, p. 103622, 2020, doi: 10.1016/j.advwatres.2020.103622.

[29] A. Mosavi, P. Ozturk, and K. W. Chau, "Flood prediction using machine learning models: Literature review," *Water (Switzerland)*, vol. 10, no. 11, pp. 1–40, 2018, doi: 10.3390/w10111536.

[30] I. R. Widiasari, L. E. Nugoho, Widyawan, and R. Efendi, "Context-based Hydrology Time Series Data for A Flood Prediction Model Using LSTM," *Proc. - 2018 5th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2018*, pp. 385–390, 2018, doi: 10.1109/ICITACEE.2018.8576900.

[31] N. B. M. Khairudin, N. B. Mustapha, T. N. B. M. Aris, and M. B. Zolkepli, "Comparison of Machine Learning Models for Rainfall Forecasting," *2020 Int. Conf. Comput. Sci. Its Appl. Agric. ICOSICA 2020*, p. 5, 2020, doi: 10.1109/ICOSICA49951.2020.9243275.

[32] Y. Da Jhong, C. S. Chen, H. P. Lin, and S. T. Chen, "Physical hybrid neural network model to forecast typhoon floods," *Water (Switzerland)*, vol. 10, no. 5, pp. 1–17, 2018, doi: 10.3390/w10050632.

[33] S. K. Jain *et al.*, "A Brief review of flood forecasting techniques and their applications," *Int. J. River Basin Manag.*, vol. 16, no. 3, pp. 329–344, 2018, doi: 10.1080/15715124.2017.1411920.

[34] E. S. K. Tiu, Y. F. Huang, J. L. Ng, N. AlDahoul, A. N. Ahmed, and A. Elshafie, *An evaluation of various data pre-processing techniques with machine learning models for water level prediction*, vol. 110, no. 1. pp. 121–153, 2022, doi: 10.1007/s11069-021-04939-8.