Research Article

Nguyen Hong Giang, YuRen Wang*, Tran Dinh Hieu*, Quan Thanh Tho, Le Anh Phuong, and Hoang Ngo Tu Do

# Toward rainfall prediction by machine learning in Perfume River Basin, Thua Thien Hue Province, Vietnam

https://doi.org/10.1515/geo-2020-0276
received March 09, 2021; accepted July 13, 2021

**Abstract:** This study examines rainfall forecasting for the Perfume (Huong) River basin using the machine learning method. To be precise, statistical measurement indicators are deployed to evaluate the reliability of the actual accumulated data. At the same time, this study applied and compared two popular models of multi-layer perceptron and the *k*-nearest neighbors (*k*-NN) with different configurations. The calculated rainfall data are obtained from the Hue, Aluoi, and Namdong hydrological stations, where the rainfall demonstrated a giant impact on the downstream from 1980 to 2018. This study result shows that both models, once fine-tuned properly, enjoyed the performance with standard metrics of R_squared, mean absolute error, Nash–Sutcliffe efficiency, and root-mean-square error. In particular, once Adam stochastic is deployed, the implementation of the MLP model is significantly improving. The promising forecast results encourage us to consider applying these models with future data to help natural disaster non-stop mitigation in the Perfume River basin.

**Keywords:** distance metric, *k*-NN, rainfall prediction, multi-layer perceptron, stochastic optimization

## 1 Introduction

Global climate change has extreme effects on the annual volume and pattern of rainfall. It is also the main cause of several droughts and floods worldwide. This situation has negatively impacted people, such as farmers, peasants, and agriculturists, whose livelihood depends on regular rainfall [1–3]. These points indicate that the desirability of highly accurate rainfall forecasting is now an urgent situation. Therefore, several studies have proposed several prediction methods of hydrological processes for forecasting soil temperature with neural networks and machine learning methods for rain run-off prediction, forecasting water flow, semi-arid precipitation forecast, and drought prediction [4–9]. In addition, recently several studies have applied machine learning methods to predict the quality of dykes, water quality in rivers, and the amount of sludge in wastewater treatment plants [10–12].

The MLP and *k*-nearest neighbors (*k*-NN) that the models conduct for supervised learning techniques in classification math are mentioned [13,14]. Classification math is divided into three processes: collecting the input training data set, using the test data set to check the classification accuracy, and deploying the classifier to categorize the new data [15]. Its abilities identify the relationships of the high complexity of input and output variables without realizing the natural physical processes [16–22]. Specific functions of the *k*-NN model's salient features are the non-parametric approach and the most straightforward in both regression and classification functions [23–25]. In addition, the main advantages of

*** Corresponding author: YuRen Wang,** Civil Engineering Faculty, National Kaohsiung University of Science and Technology, Kaohsiung 80778, Taiwan, e-mail: yrwang@nkust.edu.tw
*** Corresponding author: Tran Dinh Hieu,** Faculty of Architecture, ThuDauMot University, ThuDauMot 820000, Vietnam, e-mail: hieutd@tdmu.edu.vn, tel: +84-90-515-3333
**Nguyen Hong Giang:** Faculty of Architecture, ThuDauMot University, ThuDauMot 820000, Vietnam; Civil Engineering Faculty, National Kaohsiung University of Science and Technology, Kaohsiung 80778, Taiwan, e-mail: giangnh@tdmu.edu.vn, giangh@hueuni.edu.vn
**Quan Thanh Tho:** Faculty of Computer Engineering, HoChiMinh University of Technology, HoChiMinh 27169, Vietnam, e-mail: qttho@hcmut.edu.vn
**Le Anh Phuong:** Department of Computer Science, Hue University of Education, Hue University, Hue 49118, Vietnam, e-mail: laphuong@hueuni.edu.vn, leanhphuong@dhsphue.edu.vn
**Hoang Ngo Tu Do:** Faculty of Geology and Geography of Sciences University, Hue University, Hue 49100, Vietnam, e-mail: hoangngotudo@hueuni.edu.vn

*k*-NN can be listed as fast calculation time, a simple algorithm, easy to interpret, useful for regression and classification, high accuracy, no assumptions about data, no need to make additional assumptions and adjust some parameters or build a model [26–28]. Meanwhile, the MLP provides reliable regression and classification for the neural networks, which involves data entry from the input units and passes through the network to output units. Its hierarchy includes an input layer and one or more invisible layers of computational nodes and an output layer of computational nodes [29–31]. The MLP model integrates with the backpropagation algorithm [32].

Therefore, several studies on rainfall forecasts had been published using the models. Dash et al. [33] applied the *k*-NN model to predict the rainfall season of the summer monsoon (June–September) and post-monsoon (from October to December) for 4 years (from 2011 to 2016) in Kerala state of Indian Peninsula. The study concluded that *k*-NN has been carried out reasonably well. Wu et al. [34] used the *k*-NN model to forecast rain from February to April every year at 18 major hydrological stations in the Southeastern Mediterranean region. The results indicated that *k*-NN model well narrowed the gap between the global and the coarse forecasts models for the Southeastern Mediterranean region. Vallam and Qin [35] developed a *k*-NN model to test predicted long-term rainfall simulation in Singapore over 30 years. The results showed that the *k*-NN model is satisfactory when forecasts were conducted in the wet seasons. Moreover, the model could repeat the values closely of extreme rainfall. Zhang et al. [36] used the MLP model to predict the annual and non-monsoon rainfall prediction in Odisha, India. The results indicated that MLP was more accurate when using the model for the rest of the eight non-monsoon months in future rainfall prediction. Zahmatkesh and Goharian [37] used the MLP model to predict long lead monthly rainfall forecast from 1925 to 2016 in Vancouver, British Columbia, Canada. The research pointed out that the model with the best forecasting performance is selected to forecast rainfall 1 month ahead of time.

The perfume River basin in Thua Thien Hue Province is a vulnerable place, sensitive to natural disasters and the impact of climate change. Therefore, this area needs many types of forecasting related to natural disasters. Toward rainfall prediction for the Perfume River basin will be deployed by Machine Learning based on the Python platform. Even though the first-time study methodology is applied, this study result may contribute to making more accurate predictions and supplying a new method for rainfall forecast in this basin.

This study proposes two MLP and *k*-NN models with four configurations: Adam, L-BFGS methods, Euclidean, and Minkowski distance metrics predict rainfall in the Perfume River basin, respectively. These models are also deployed to compare each other to find the most optimal model. Several accurate measurement parameters such as R_squared, Nash–Sutcliffe efficiency (NSE), root-mean-square error (RMSE), and mean absolute error (MAE) are used to evaluate the accuracy levels of the proposed models. In addition, statistical measurement indicators (the percentage, the average, minimum and maximum values, standard deviation (St Dev), coefficient of variation (Cv) are applied to evaluate the reliability of the actual accumulated data.

The rest of the paper is structured as follows: Section 2 describes the methodology and study area, Section 3 evaluates the study data and analyzes the study results, Section 4 discusses the study approaches and limitations, and Section 5 presents the conclusion.

# 2 Methodology, study area, and data collection

## 2.1 Methodology

### 2.1.1 Multi-layer perceptron

The MLP model is considered a typical representative. It includes an input layer, an output layer, and many hidden layers in between; all the nodes in the hidden layers and the output layer are named as neurons. The strength of the signal transmitting from one node to the others depends on the connection weight of the interconnections. Hidden layers improve the network's ability to complex functions of the model [38,39], appurtenant to a lot of the training process. The training principle for the MLP model is using a variety of backpropagation algorithms. Training is a process of adjusting the weights and bias connections and calculating the errors caused by the network. In the training process, the differences between the desired with actual responses that the output layer of the training process fit the best-desired output [40]. During training of neurons, the activation function is applied to this training process and the rectified linear unit (ReLU) is used for the activation function. ReLU does training for machine learning networks [41,42]. Due to ReLU convergence and gradient calculation almost instantly, ReLU solves explosion and the disappearance of gradients,

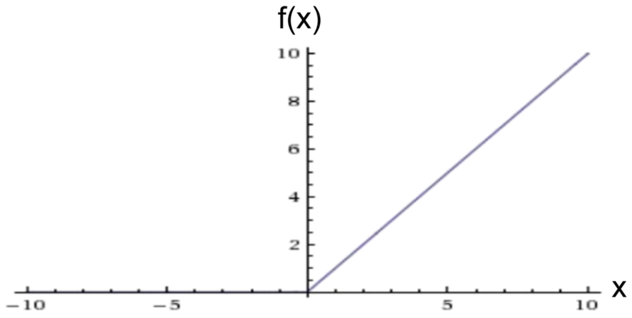

**Figure 1:** ReLU function graph.

maintaining a steady-state convergence rate as well [43]. In addition, the ReLU function is simple and effective for rainfall prediction [44]. For popular forecasts, the Adam or Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) method applies for a stochastic optimizer [45,46].

The specific characteristics of ReLU function, Adam, and L-BFGS methods are explained in detail.

The ReLU function is illustrated in Figure 1, and ReLU is described as follows:

$$f(x) = \max(0, x). \tag{1}$$

Equation (1) indicates that $f'(x) = 0$ when $x < 0$ and $f'(x) = 1$ when $x \geq 0$.

L-BFGS is an algorithm of optimization of the quasi-Newton methods. It applies to the estimation of parameters in Machine Learning [47,48]. L-BFGS was performed as an estimate of the Hessian matrix of inversion; the purpose of steer is to search through variable space. Due to its requiring linear memory, the L-BFGS method is particularly suitable for optimization problems with multiple variables [49,50].

Adam is derived from the estimation of the adaptive moment. The Adam method is applied for efficient stochastic optimization. It only requires a small memory for the first-order gradience; it calculates the learning rates for different parameters from approximate for the first and second moments of the gradients. The method has several advantages of deep neutral networks as follows. The parameter amplitude updates do not change the gradient scale, do not need the stationary objective, and the step-sizes approximate bounded by the step-sizes of hyper-parameter. At the same time, it carries out with sparse gradients and naturally works in the form of step-size annealing.

In this study, Figure 2 describes the structural MLP that input layer has 12 input nodes from $a_1$ to $a_{12}$ (which are also 12 months of the year), one neuron of the output layer has represented the values of rainfall. There are three hidden layers: the first hidden layer contains neurons from $H_{11}$ to $H_{112}$, the second one is from $H_{21}$ to $H_{212}$, and the last one is from $H_{31}$ to $H_{312}$. Each neuron of the hidden layer and the output layer has a corresponding weight and bias, as $W_{11}{}^{(2)}$, $B_1{}^{(1)}$ and $W_{12}{}^{(2)}$, $B_2{}^{(2)}$ are the weight and bias to correspond for neuron $H_{11}$ and neuron $H_{12}$, respectively, so on. Each neuron of the hidden layers takes the output from all neurons of the previous layers and converts these values with a weighted linear sum into the output layer, where $n$ is the number of neurons of class and corresponds to the component of the vector weights. The output class gets the values from the last hidden layer. The ReLU function is the activation function for three hidden layers. Adam and L-BFGS methods are two stochastic optimizations to the solver of weight optimization, and using these two methods, rainfall prediction of three station areas is compared. The training method for MLP is regression.

### 2.1.2 *k*-NN

The *k*-NN is the layer model for objects and locates on the nearest distance between the objects (query point) layer

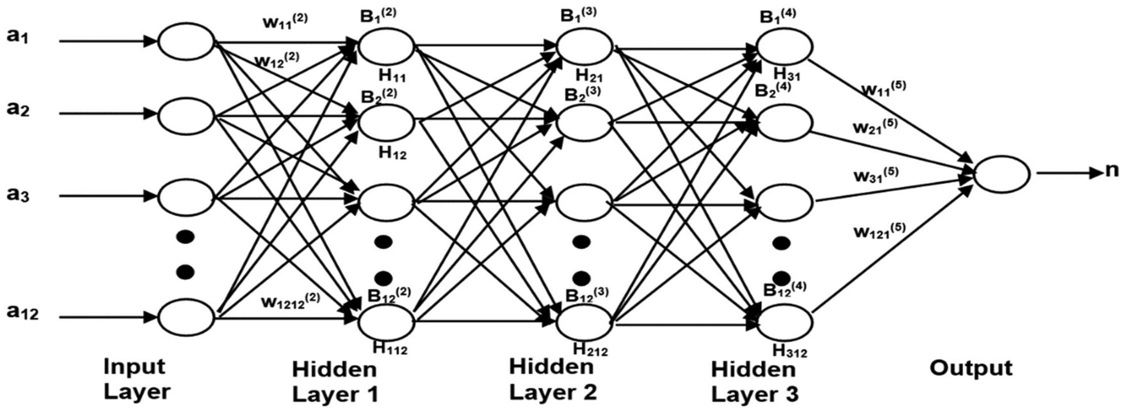

**Figure 2:** Structure of MLP network for rainfall prediction.

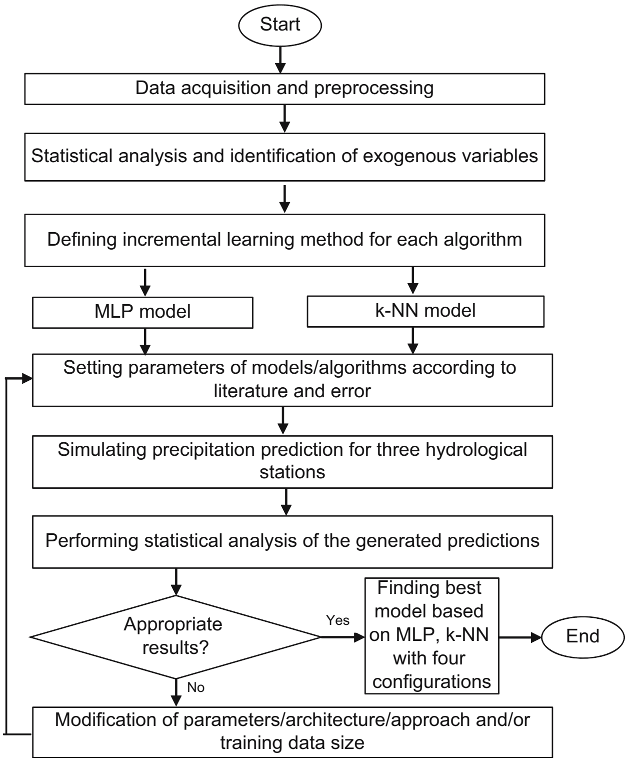
Start
Data acquisition and preprocessing
Statistical analysis and identification of exogenous variables
Defining incremental learning method for each algorithm
MLP model
k-NN model
Setting parameters of models/algorithms according to literature and error
Simulating precipitation prediction for three hydrological stations
Performing statistical analysis of the generated predictions
Appropriate results?
Yes
Finding best model based on MLP, k-NN with four configurations
End
No
Modification of parameters/architecture/approach and/or training data size

**Figure 3:** Flowchart of the experimental steps conducted in this study.

and remained objects in the training data. The *k*-NN algorithm is considered an easy learning algorithm and a simple implementation [49]. The response values are calculated as a weighted sum of the whole *k* neighbors when the *k*-NN model carries out the regression method. The weight is inversely proportional to the distance from the input record. This distance is called the Minkowski distance. Wilson and Martinez [51] defined the Minkowski distance of order *p* (*p* is an integer) between two vectors *X*, *Y* as follows:

$$M\ (X,\ Y\ )=\ \sqrt{\sum_{i}^{n}|x_i -\ y_i|^p}\ , \qquad (2)$$

where $x_i$ is the $i$th value in the vector $X = \sum_i^n x_I$ and $y_i$ is the $i$th value in the vector $Y = \sum_i^n y_i$; there are numeric input variables, while $n$ is the number of input variables. In this study, the $p$ values in equation (2) will be $p = 2$ and $p = \infty$, which are Euclidean and Minkowski distance metrics, respectively. Our purpose is to find the best model for rainfall prediction at three station areas; the results of the study compares *k*-NN classification error rates by using Euclidean distance versus Minkowski distance. To break the relation between different classes that we continuously decrease the neighbor size, ultimately classifying by just the $k = 3$ nearest neighbors.

After selecting the value of $k$, a prediction is an average over the outcomes for *k*-NN, and equation (3) is as follows [28]:

$$\theta\ =\ \frac{1}{k}\sum_{k}^{i=1} o_i, \qquad (3)$$

where $\mathrm{o}_i$ is the $i$ thvalue in the vector $\theta$ and $o$ is the number of output variables.

### 2.1.3 Accuracy measurements

Forecasting data will be calculated and compared with actual data to accurately evaluate the forecasted values.

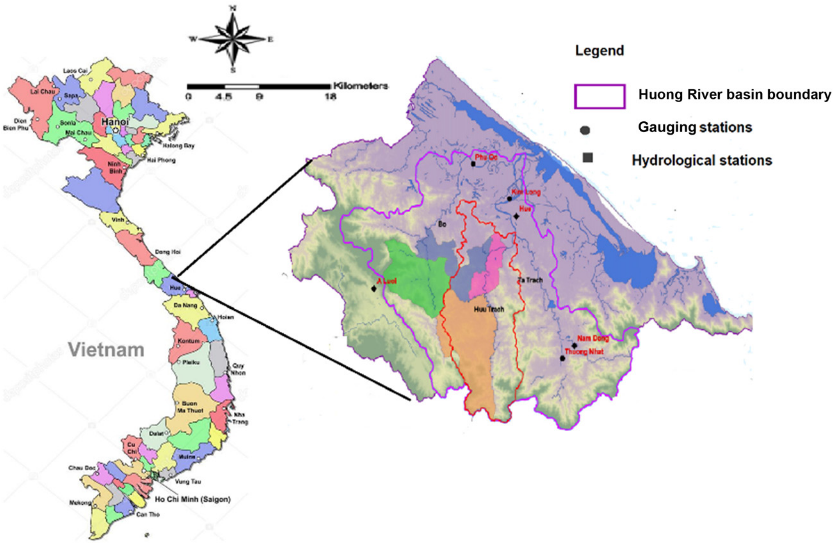

**Figure 4:** The position of the meteorological stations.

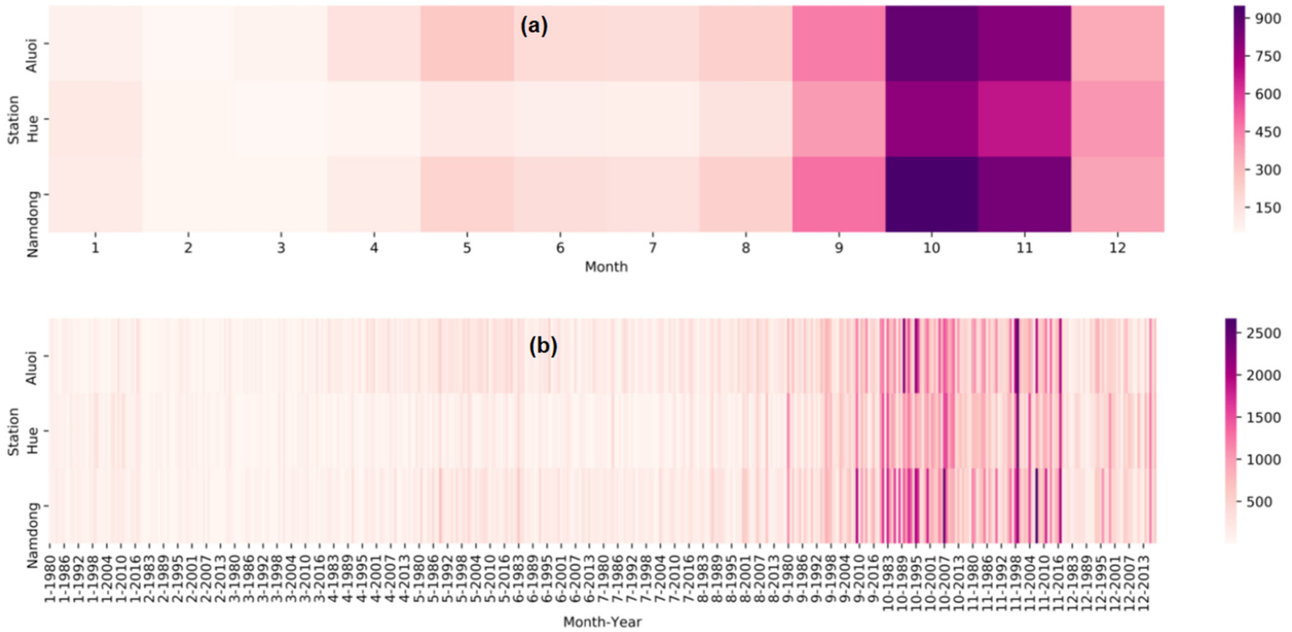

**Figure 5:** Monthly rainfall from 1980 to 2018 at the three hydrological stations.

The metrics that calculate the forecast accuracy include the MAE, the RMSE, and the R_squared. The error metrics are as follows:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|x_{f,t} - x_{a,t}|, \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(x_{f,t} - x_{a,t})^2}{n}}, \tag{5}$$

$$\text{R\_squared} = 1 - \frac{\sum_{t=1}^{n}(x_{a,t} - x_{f,t})^2}{\sum_{t=1}^{n}\left(x_{a,t} - \frac{1}{n}\sum_{t=1}^{n}x_{a,t}\right)^2}, \tag{6}$$

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{n}(x_{a,t} - x_{f,t})^2}{\sum_{t=1}^{n}(x_{a,t} - \bar{x})^2}, \tag{7}$$

where $x_{f,t}$ and $x_{a,t}$ are the forecast value and actual value in the period time $t$, respectively, $\bar{x}$ is mean of the observed value, and $n$ is the number of the observed values in the testing data. The forecasting study accuracy is conducted by the R_squared, NSE, and the error indicators (MAE, RMSE). The R_squared and NSE should be approaching up to 1 to indicate strong model performance, and the error indicators should be as close to zero as possible. Based on these error indicators, the best prediction model for each station area is chosen.

Thus, the methodology of this paper is summarized in Figure 3. The data in Figure 3 describe a flowchart illustrating the experiment steps for this study.

## 2.2 Study area

### 2.2.1 Brief of geography

Thua Thien Hue Province belongs to the North Central Coast Region of Vietnam. The province containing the largest basin is the Perfume River basin (see Figure 4), which is located between the North of Bach Ma mountain and the East of Truong Son range, its area is about 2,830 km$^2$, the altitude ranging from 200 m to 1,708 m, and the average slope ranging from 15 to 35°. Its main

**Table 1:** Hydrological location, record period, and years considered

| Station | Location | Earliest record year | Latest record year | Numbers of month |
|---|---|---|---|---|
| Hue | Hue city | 1980 | 2018 | 468 |
| Aluoi | Aluoi district | 1980 | 2018 | 468 |
| Namdong | Namdong district | 1980 | 2018 | 468 |

**Table 2:** Statistical characteristics of monthly precipitation data

| | Percentage | | Average (mm) | | St Dev | | Cv (%) | | Min (mm) | | Max (mm) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Station** | **Min** | **Max** | **Min** | **Max** | **Min** | **Max** | **Min** | **Max** | **Min** | **Max** | **Min** | **Max** |
| Hue | 20 | 314 | 50.6 | 788.3 | 46.7 | 451 | 48 | 106 | 3.2 | 35 | 353.7 | 2,452.3 |
| Aluoi | 21 | 275 | 68.2 | 912.4 | 68.2 | 912.4 | 31 | 89 | 4.5 | 132.7 | 499.0 | 2590.0 |
| Namdong | 20 | 293 | 66.2 | 974.0 | 50 | 681.1 | 43 | 76 | 1.6 | 123.5 | 412.4 | 2,672.3 |

branches originate from the high areas of Bach Ma mountain, flow from South to North about 104 km. At the same time, the basin has three relatively sub-drainage basins: Huu Trach branch (a catchment range of 691 $km^2$ with 70 km long), Ta Trach branch (a catchment range of 729 $km^2$ with 51 km long), and Bo River (a drainage basin of 938 $km^2$ with 94 km long). Perfume River basin has the highest rainfall in Vietnam. Annually, the dry season runs from March to August in this basin, and the rough often from the end of July to the end of August. Especially, hurricane season starts in September and finishes in December. The average precipitation in Hue, ALuoi, and Namdong areas is about 2,850, 3,500, and 3,200 mm, respectively (see Figure 5(a)–(b)). The basin topography has not transitional areas from the upstream of the mountain down to the plain and the lagoon system. Hence, this morphology mainly causes high runoff upstream and large floods downstream during the rainy season.

In addition, the black square dots in Figure 4 point out the Hue, Aluoi, Namdong hydrological stations. The areas signify various climatic characteristics. The precipitation of three hydrological stations is a key to flood or drought seasons in the downstream. Therefore, the obtained rainfall data are crucially important in this study.

## 2.3 Data collection

The annual statistical report by Thua Thien Hue Centre for Hydro-Meteorological Forecasting provided the monthly rainfall data of Hue, Aluoi, and Namdong hydrological stations. The data are also checked with the annual statistical report of Thua Thien Hue Province. This preliminary data evaluation process is crucial for the study input. Table 1 shows the features of the data deployed in this study.

Statistical features calculating from the monthly rainfall time series of each hydrological station are listed in Table 2. For comparative implementation, monthly rainfall data were measured with millimeters (mm). The range of the following characteristics was computed from the time series of the observed monthly rainfall: the percentage, average, minimum and maximum values, St Dev, and Cv.

Dataset included 468 rainfall months from January 1980 to December 2018. In this study, the dataset from January 1980 to December 2003 of the hydrological stations is used for the training phase, and the dataset from January 2004 to December 2018 is applied for the test phase.

# 3 Results

## 3.1 The rainfall forecasting of the MLP model

After many experiments to find the optimal MLP model with two methods of Adam and L-BFGS, the study found the optimal model with the values of the core parameters that are listed in Table 3.

The results of the simulation by the MLP model with Adam and L-BFGS stochastic optimizations is shown in Figure 6. The line charts in Figure 6(a)–(c) are relatively good fitness between trained data and tested data for

**Table 3:** MLP basic component

| Item | Configuration |
|---|---|
| Number of inputs | 12 |
| Number of hidden layers | 3 |
| Hidden layer sizes | 12/12/12 |
| Number of outputs | 1 |
| Learning rate init | 0.001 |
| Iter no change | 10 |
| Beta 1 | 0.9 |
| Validation_fraction | 0.1 |
| Alpha | 0.0001 |
| Max iter | 10000 |
| Power_t | 0.5 |
| Beta 2 | 0.999 |
| Solver | Adam, L-BFGS |

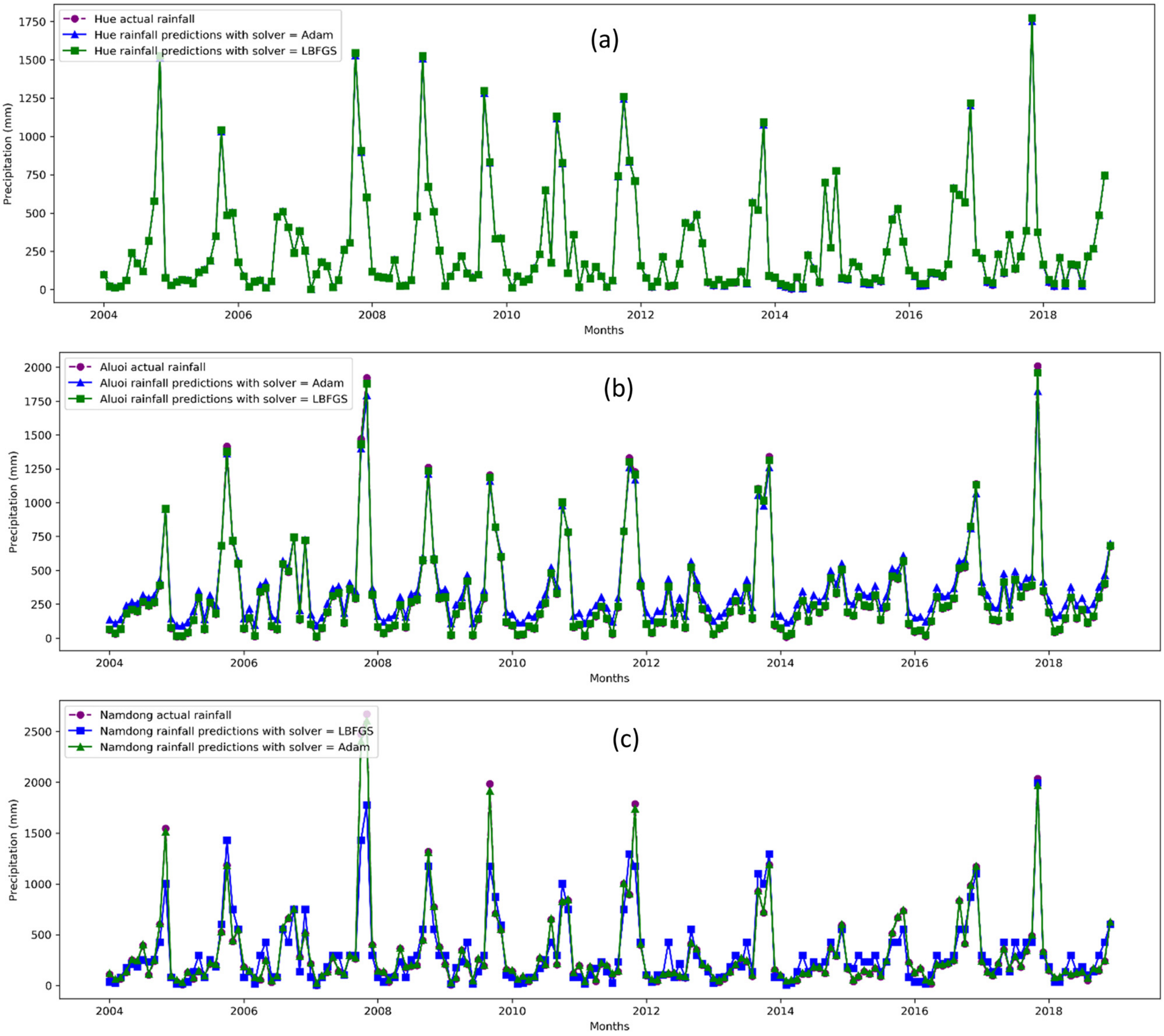

**Figure 6:** The actual and predicted rainfall forecasting based on the MLP model with Adam and L-BFGS stochastic optimizations at (a) Hue, (b) Aluoi, and (c) Namdong stations.

MLP models with Adam and L-BFGS stochastic optimizations. The difference between the two stochastic optimizations of the three hydrological stations is hardly distinguished by the figures. Hence, the accuracy parameters are provided in higher detail in the data in Table 4.

The data in Table 4 compares the two methods of Adam and L-BFGS stochastic optimizations. Results from the statistics show that these three hydrological stations have more accurate values when using the Adam method. The results show that the best model is Hue with R-squared = 0.999, NSE = 0.999, MAE = 2.97, and RMSE = 5.38, the second-best model is Aluoi with R-squared = 0.991, NSE = 0.998, MAE = 7.81, and RMSE = 9.85, and the third-best model is Namdong hydrological station with R-squared = 0.986, NSE = 0.996, MAE = 14.37, and RMSE = 17.18.

## 3.2 The rainfall forecasting of the *k*-NN model

The parameters in Table 5 give optimal values for the *k*-NN model with distance metrics $p = \{2, \infty\}$. These values are obtained after many experiments to get the optimal model.

**Table 4:** Accuracy parameters for rainfall prediction used MLP models at the three hydrological stations

| Parameter | Hue rainfall prediction used Adam | Hue rainfall prediction used L-BFGS | Average | Namdong rainfall prediction used Adam | Namdong rainfall prediction used L-BFGS | Average | Aluoi rainfall prediction used Adam | Aluoi rainfall prediction used L-BFGS | Average |
|---|---|---|---|---|---|---|---|---|---|
| R_squared | 0.999 | 0.997 | 0.998 | 0.986 | 0.984 | 0.985 | 0.991 | 0.988 | 0.990 |
| NSE | 0.999 | 0.998 | 0.999 | 0.996 | 0.995 | 0.996 | 0.998 | 0.997 | 0.998 |
| MAE | 2.97 | 5.12 | 4.045 | 14.37 | 16.36 | 15.37 | 7.81 | 9.59 | 8.70 |
| RMSE | 4.38 | 6.24 | 5.31 | 17.18 | 20.21 | 18.70 | 9.85 | 13.11 | 11.48 |

**Table 5:** *k*-NN basic components

| | | | |
|---|---|---|---|
| Algorithm | auto | Leaf_size | 30 |
| Metric: | Minkowski | P | {2, ∞} |
| N_neighbors: | 3 | Weights: | uniform |

The data in Figure 7 show the *k*-NN for rainfall forecasting to apply distance metric with $p = 2$ and $p = \infty$ in Hue, Namdong, and Aluoi hydrological stations. Figure 7(b) and (c) indicates that the rainfall prediction and actual rainfall are a very close relationship; moreover, there are no significant differences. Because the two graphs above are difficult to distinguish the best optimal distance metric, the data in Table 6 is provided to evaluate the best method. Figure 7(a) shows that the *k*-NN with $p = \infty$ is the best forecast for rainfall at Hue hydrological station; moreover, the prediction and actual data are very rigid. Meanwhile, the relationship between expected and actual rainfall of the values of $p = 2$ is loose-fitting.

The data in Table 6 show the value of prediction errors of the R_squared, NSE, MAE, and RMSE. These data were collected from the analysis of the rainfall prediction of the three hydrological stations using the *k*-NN model with distance metrics of $p = 2$ and $p = \infty$. At the same time, the result of the analysis indicated that the value of the forecast errors at the Hue station with $p = \infty$ is the lowest, and the second-lowest is the Aluoi station with $p = 2$. On the other hand, the value of forecast errors for the Namdong station with $p = 2$ is the highest. R_squared, NSE, MAE, and RMSE of the best model for the Hue, Aluoi, and Namdong hydrological station are 0.993, 0.998, 16.39, and 27.70; 0.987, 0.996, 19.05, 29.36; and 0.983, 0.992, 21.67, 61.88, respectively.

## 3.3 Comparison and analysis of simulation results between the MLP model and the *k*-NN model

The models of MLP and *k*-NN are carried out to assess rainfall during the 1980 to 2018 period in Thua Thien Hue Province. The line chart of Figure 8 summarizes the best rainfall projections at Hue, Aluoi, and Namdong hydrological stations after using the methods of distance metric and stochastic optimization for both the *k*-NN and MPL models.

Figure 9 and Table 7 show that the average R_squared and NSE indicators of the two models are from 0.987 to 0.997, which proves that simulation results in a highly

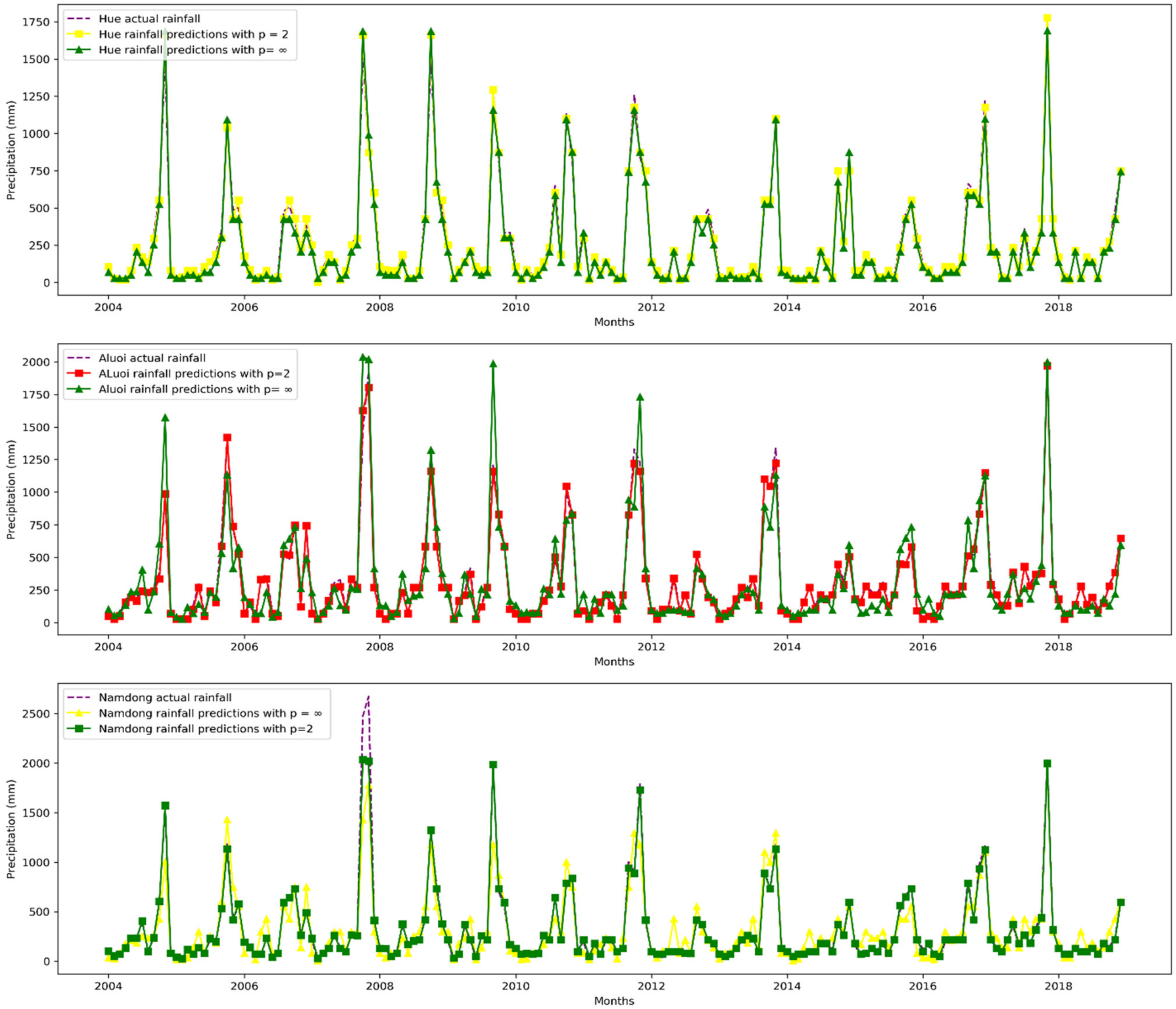

**Figure 7:** The actual and predicted rainfall forecasting based on the *k*-NN model with $p = 2$ and $p = \infty$ at (a) Hue hydrological station, (b) Aluoi hydrological station, and (c) Namdong hydrological station.

accurate forecast when compared with true data together. At the same time, the average error indicators of the MLP and *k*-NN models fluctuate from 8.38 to 39.65, in which the average values of MAE, RMSE parameters of the *k*-NN, and MPL models are 19.04, 8.38 and 39.45, 10.47, respectively, which mean that the indicators are fitness values for both the rainfall training data and the rainfall forecasting data.

In addition, Figure 10 shows a comparison between the predicted values of precipitation rainfall and the actual values of precipitation in the training and testing

**Table 6:** Accuracy parameters for rainfall prediction used *k*-NN models at the three hydrological stations

| Parameter | Hue $p = 2$ | Hue $p = \infty$ | Average | Namdong $p = 2$ | Namdong $p = \infty$ | Average | Aluoi $p = 2$ | Aluoi $p = \infty$ | Average |
|---|---|---|---|---|---|---|---|---|---|
| R_squared | 0.982 | 0.993 | 0.985 | 0.983 | 0.981 | 0.982 | 0.987 | 0.979 | 0.982 |
| NSE | 0.996 | 0.998 | 0.997 | 0.992 | 0.991 | 0.992 | 0.996 | 0.994 | 0.995 |
| MAE | 32.83 | 16.39 | 24.61 | 21.67 | 28.65 | 25.16 | 19.05 | 31.46 | 25.255 |
| RMSE | 43.62 | 27.70 | 35.66 | 61.88 | 76.21 | 69.045 | 29.36 | 44.78 | 37.07 |

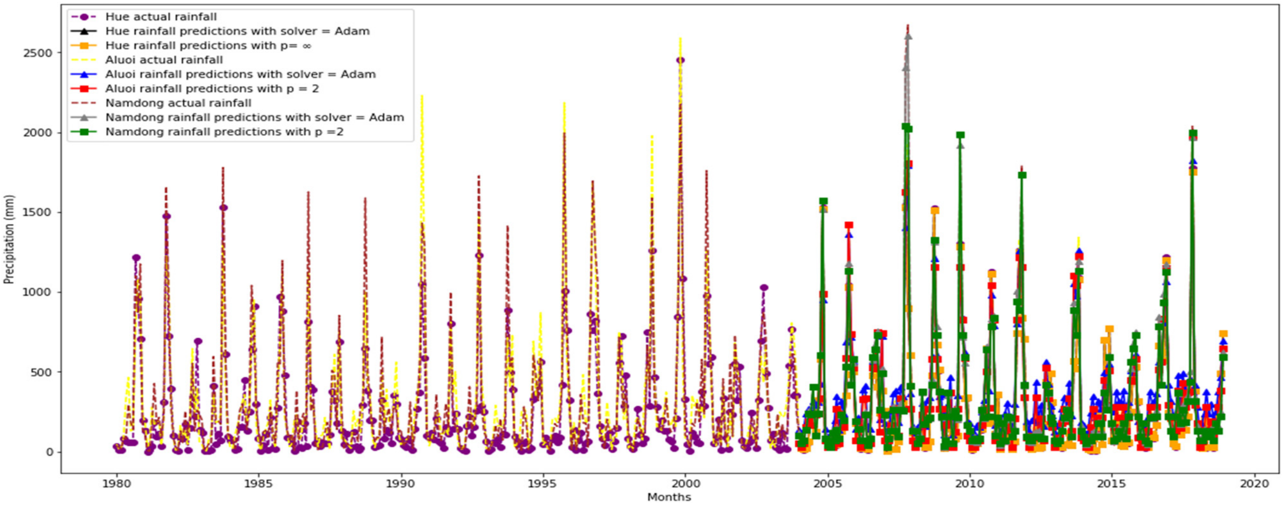

**Figure 8:** The best forecast of precipitation when using *k*-NN and MLP models.

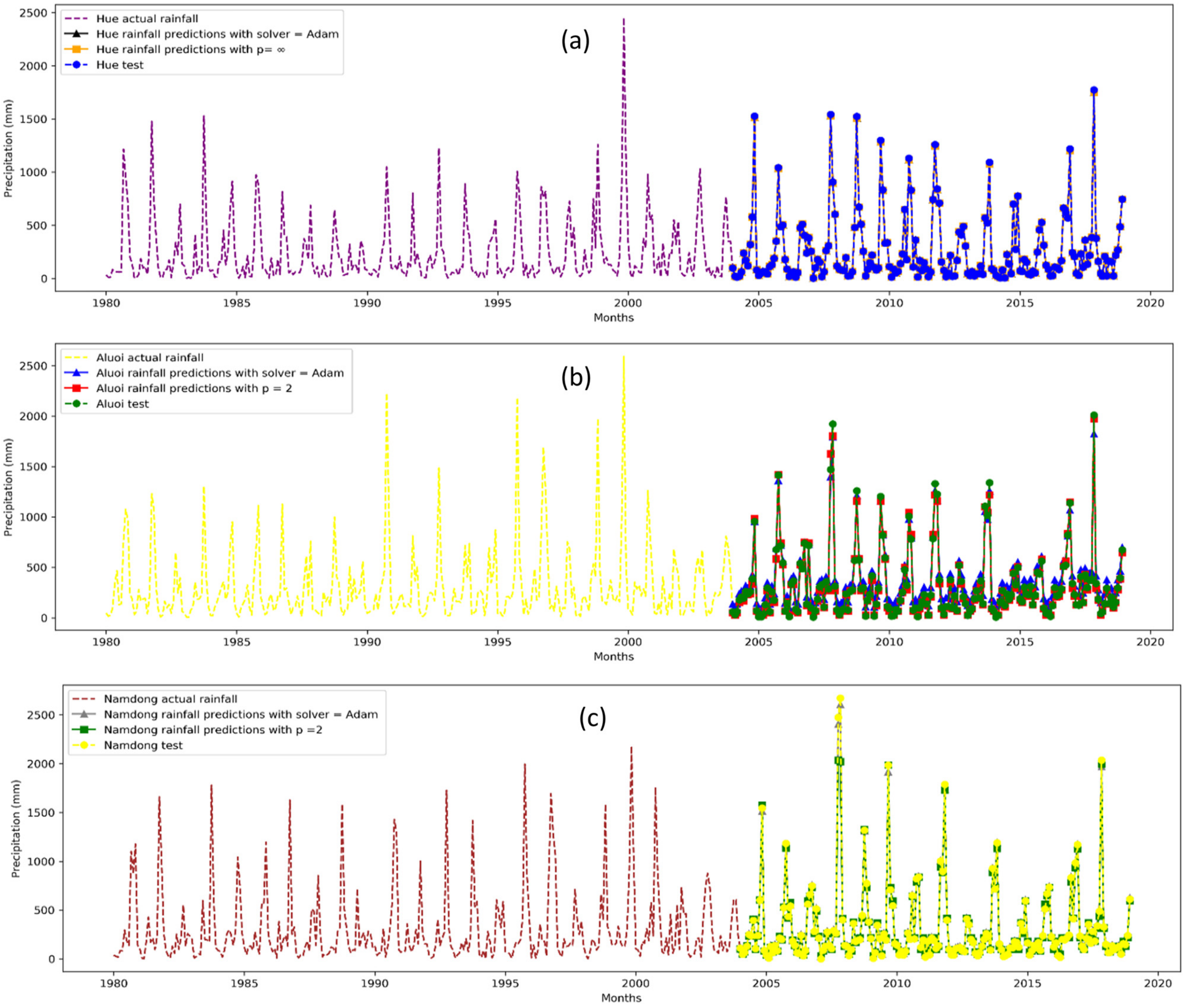

**Figure 9:** The best forecast of precipitation when using *k*-NN and MLP models at the (a) Hue, (b) Aluoi, and (c) Namdong hydrological stations.

**Table 7:** The best accuracy parameters for rainfall prediction used *k*-NN and MLP models at the three hydrological stations

| **Parameter** | ***k*-NN** | | | | **MLP** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Hue $p = \infty$** | **Namdong $p = 2$** | **Aluoi $p = 2$** | **Average** | **Hue Adam** | **Namdong Adam** | **Aluoi Adam** | **Average** |
| R_squared | 0.993 | 0.983 | 0.987 | 0.988 | 0.999 | 0.986 | 0.991 | 0.992 |
| NSE | 0.998 | 0.992 | 0.996 | 9.995 | 0.999 | 0.995 | 0.998 | 0.997 |
| MAE | 16.39 | 21.67 | 19.05 | 19.04 | 2.97 | 14.37 | 7.81 | 8.38 |
| RMSE | 27.7 | 61.88 | 29.36 | 39.65 | 4.38 | 17.18 | 9.85 | 10.47 |

periods and the correlation coefficient for the best *k*-NN model and MLP prediction model. The MLP method is more exact in the provision of the correlation coefficient.

## 4 Discussion

The result of simulating rainfall by MLP and *k*-NN models using four different configurations showed the following findings. Two models obtained the best performance and reliability for rainfall prediction; moreover, the forecast values compared to the actual parameters achieved high accuracy, where the R_squared and NSE values were higher than 0.979. At the same time, the RMSE values were lower than 76.21. The MLP model with the Adam optimization method gave the best accuracy for rainfall prediction to compare with the rest methods.

The study is conducted to predict a time series of annual rainfall from 1980 to 2018 in three hydrological stations: Hue station is located downstream and Aluoi and Namdong are located upstream.

However, several recent rainfall studies have incorporated rainfall and some effects on precipitation. The research of Choubin et al. [52] evaluated factors that may influence fall rain forecast in Kerman Province, Iran, which consisted of large-scale oceanic and atmospheric information. Hence, the combination between these factors and accumulated rainfall data has given high accuracy for the forecast of autumn rainfall. Rainfall data have non-linear variation. Therefore, Choubin et al. [53] deployed the data normalization method for the rainfall study at the Maharlu-Bakhtegan basin, Iran. And the results indicated that the MLP model using data after normalization have resulted in a lower RMSE than the RMSE of this study. In addition, the studies by

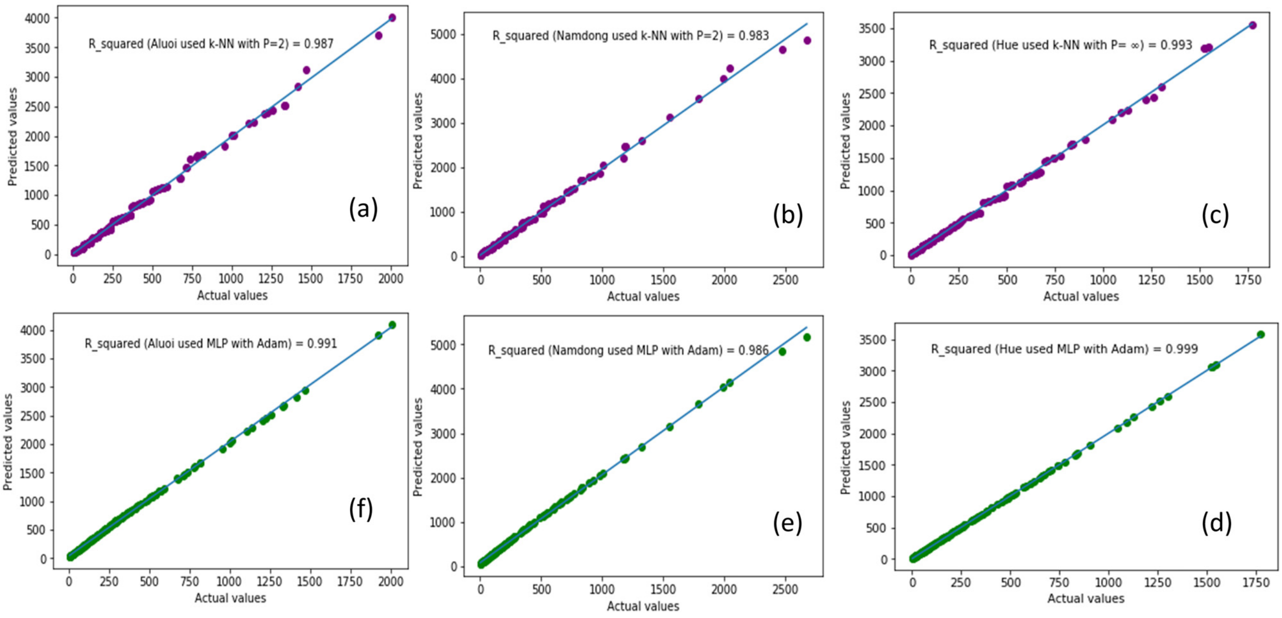

**Figure 10:** The best performance R_squared of MLP and *k*-NN terms of the correlation coefficient for Thua Thien Hue Province in (c), (d) Hue hydrological station, (b), (e) Namdong hydrological station, and (a), (f) Aluoi hydrological station.

Najafzadeh et al. [54,55] have used some models such as neuro-fuzzy group method of data handling (NF-GMDH) based on self-organized models and group method of data handling gene-expression programming (GMDH-GEP) model to forecast bridge pier scour depth under debris flow effects and free span expansion rates below pipelines under waves, respectively. Research results have shown that the RMSE index of these two models is also smaller than the RMSE value of this study.

Even though the precipitation at three hydrological stations has a seasonal variation with different complexity, applying these two models with four configurations has achieved high-reliability results. Hence, it can be used for rainfall forecasting for other regions in Vietnam. In addition, the study results are also a utility reference channel for the province authority to develop short-term plans for natural disaster mitigation.

## 5 Conclusion

This study performs the predicted precipitation of the Perfume River basin. This study also indicated that the MLP model is more accurate than the *k*-NN model. The measured rainfall was collected from three hydrological stations at the Hue, Namdong, and ALuoi areas of the province from 1980 to 2018. The dataset is separated using time-based criteria: training data (1980–2003) and test data (2004–2018). The results demonstrate that the effectiveness of the models for the core parameters has been mentioned earlier. In addition, this study result may help the Thua Thien Hue government formulate short-term plans of natural disasters to mitigate for the basin.

**Acknowledgments:** We would like to thank the school of Civil Engineering of National Kaohsiung University of Science and Technology of Taiwan, and the Thu Dau Mot University of Vietnam for implementing and financial study. We also thank the master class of MSE#07HCM at FSB School Of Business and Technology and Dr. Hector Tibo, Dr. June Raymon (who are currently Ph.D. students at the National Kaohsiung University of Science and Technology of Taiwan) for their help in correcting the academic writing of the paper.

**Funding information:** This research is funded by Thu Dau Mot University, Vietnam. The APC was funded by Thu Dau Mot University.

**Author contributions:** Conceptualization, discussion, and conclusions, material and methods: Nguyen Hong Giang; writing, original draft preparation: Tran Dinh Hieu, Hoang Ngo Tu Do; writing, review, and editing: Yu Ren Wang, Quan Thanh Tho, Le Anh Phuong; funding acquisition: Tran Dinh Hieu and Nguyen Hong Giang. All authors have read and agreed to the published version of the manuscript.

**Conflict of interest:** The authors declare no conflict of interest.

**Data availability statements:** The datasets analyzed during the study are available from the corresponding author on request.

## References

[1] Wang B, Xiang B, Li J, Webster PJ, Rajeevan MN, Liu J, et al. Rethinking Indian monsoon rainfall prediction in the context of recent global warming. Nat Commun. 2015;6(1):1–9.

[2] Cramer S, Kampouridis M, Freitas AA, Alexandridis AK. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. Expert Syst Appl. 2017;85:169–81.

[3] Kusiak A, Wei X, Verma AP, Roz E. Modeling and prediction of rainfall using radar reflectivity data: a data-mining approach. IEEE Trans Geosci Remote Sens. 2012;51(4):2337–42.

[4] Bui DT, Pradhan B, Lofman O, Revhaug I, Dick ØB. Regional prediction of landslide hazard using probability analysis of intense rainfall in the Hoa Binh province, Vietnam. Nat Hazards. 2013;66(2):707–30.

[5] Bonakdari H, Moeeni H, Ebtehaj I, Zeynoddin M, Mahoammadian A, Gharabaghi B. New insights into soil temperature time series modeling: linear or nonlinear? Theor Appl Climatol. 2019;135(3):1157–77.

[6] Labat D, Ababou R, Mangin A. Linear and nonlinear input/output models for karstic springflow and flood prediction at different time scales. Stoch Environ Res risk Assess. 1999;13(5):337–64.

[7] Adamowski J, Sun K. Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. J Hydrol. 2010;390(1–2):85–91.

[8] Choubin B, Khalighi-Sigaroodi S, Malekian A, Ahmad S, Attarod P. Drought forecasting in a semi-arid watershed using climate signals: a neuro-fuzzy modeling approach. J Mt Sci. 2014;11(6):1593–605.

[9] Choubin B, Malekian A, Samadi S, Khalighi-Sigaroodi S, Sajedi-Hosseini F. An ensemble forecast of semi-arid rainfall using large-scale climate predictors. Meteorol Appl. 2017;24(3):376–86.

[10] Zeinolabedini M, Najafzadeh M. Comparative study of different wavelet-based neural network models to predict sewage

sludge quantity in wastewater treatment plant. Environ Monit Assess. 2019;191(3):1–25.

[11] Najafzadeh M, Oliveto G. Riprap incipient motion for overtopping flows with machine learning models. J Hydroinf. 2020;22(4):749–67.

[12] Najafzadeh M, Ghaemi A. Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. Environ Monit Assess. 2019;191(6):1–21.

[13] Hosseini S, Azizi M. The hybrid technique for DDoS detection with supervised learning algorithms. Computer Netw. 2019;158:35–45.

[14] Govindarajan M, Chandrasekaran RM. Intrusion detection using neural based hybrid classification methods. Computer Netw. 2011;55(8):1662–71.

[15] Eslamloueyan R. Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the Tennessee–Eastman process. Appl Soft Comput. 2011;11(1):1407–15.

[16] Mahsin MD. Modeling rainfall in Dhaka division of Bangladesh using time series analysis. J Math Model Appl. 2011;1(5):67–73.

[17] Alizadeh Z, Yazdi J, Kim JH, Al-Shamiri AK. Assessment of machine learning techniques for monthly flow prediction. Water. 2018;10(11):1676.

[18] Ren J, Ren B, Zhang Q, Zheng X. A Novel hybrid extreme learning machine approach improved by *k*-nearest neighbor method and fireworks algorithm for flood forecasting in medium and small watershed of Loess region. Water. 2019;11(9):1848.

[19] Nkoana R. Artificial neural network modelling of flood prediction and early warning. Master Degree. Bloemfontein: University of the Free State; 2011. ufs.ac.za.

[20] Di Piazza A, Conti FL, Noto LV, Viola F, La Loggia G. Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. Int J Appl Earth Obs Geoinf. 2011;13(3):396–408.

[21] Chang TK, Talei A, Alaghmand S, Ooi MPL. Choice of rainfall inputs for event-based rainfall-runoff modeling in a catchment with multiple rainfall stations using data-driven techniques. J Hydrol. 2017;545:100–8.

[22] Martínez-Acosta L, Medrano-Barboza JP, López-Ramos Á, Remolina López JF, López-Lambraño ÁA. SARIMA approach to generating synthetic monthly rainfall in the Sinú River watershed in Colombia. Atmosphere. 2020;11(6):602.

[23] Loh WY. Classification and regression trees. Wiley Interdiscip Rev Data Min Knowl Discov. 2011;1(1):14–23.

[24] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J. Efficient water quality prediction using supervised machine learning. Water. 2019;11(11):2210.

[25] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.

[26] Gayathri K, Marimuthu A. Text document pre-processing with the KNN for classification using the SVM. 2013 7th International Conference on Intelligent Systems and Control (ISCO). IEEE; 2013. p. 453–7.

[27] Amra IAA, Maghari AY. Students performance prediction using KNN and Naïve Bayesian. 2017 8th International Conference on Information Technology (ICIT). IEEE; 2017 May. p. 909–13.

[28] Imandoust SB, Bolandraftar M. Application of *k*-nearest neighbor (knn) approach for predicting economic events: Theoretical background. Int J Eng Res Appl. 2013;3(5):605–10.

[29] Tfwala SS, Wang YM. Estimating sediment discharge using sediment rating curves and artificial neural networks in the Shiwen River, Taiwan. Water. 2016;8(2):53.

[30] Jozdani SE, Johnson BA, Chen D. Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification. Remote Sens. 2019;11(14):1713.

[31] Abdullah S, Ismail M, Ahmed AN, Abdullah AM. Forecasting particulate matter concentration using linear and non-linear approaches for air quality decision support. Atmosphere. 2019;10(11):667.

[32] Naganna SR, Deka PC, Ghorbani MA, Biazar SM, Al-Ansari N, Yaseen ZM. Dew point temperature estimation: application of artificial intelligence model integrated with nature-inspired optimization algorithms. Water. 2019;11(4):742.

[33] Dash Y, Mishra SK, Panigrahi BK. Rainfall prediction for the Kerala state of India using artificial intelligence approaches. Comput Electr Eng. 2018;70:66–73.

[34] Wu W, Liu Y, Ge M, Rostkier-Edelstein D, Descombes G, Kunin P, et al. Statistical downscaling of climate forecast system seasonal predictions for the Southeastern Mediterranean. Atmos Res. 2012;118:346–56.

[35] Vallam P, Qin XS. Multi-site rainfall simulation at tropical regions: a comparison of three types of generators. Meteorol Appl. 2016;23(3):425–37.

[36] Zhang X, Mohanty SN, Parida AK, Pani SK, Dong B, Cheng X. Annual and non-monsoon rainfall prediction modelling using SVR-MLP: an empirical study from Odisha. IEEE Access. 2020;8:30223–33.

[37] Zahmatkesh Z, Goharian E. Comparing machine learning and decision making approaches to forecast long lead monthly rainfall: The city of Vancouver, Canada. Hydrology. 2018;5(1):10.

[38] Cao W, Wang X, Ming Z, Gao J. A review on neural networks with random weights. Neurocomputing. 2018;275:278–87.

[39] Patra JC, Pal RN, Chatterji BN, Panda G. Identification of nonlinear dynamic systems using functional link artificial neural networks. IEEE Trans Syst Man Cyber Part B. 1999;29(2):254–62.

[40] Simpson PK. Artificial neural systems: foundations, paradigms, applications, and implementations. 1st ed. Elmsford, NY: Pergamon Press, Inc.; 1990. worldcat.org.

[41] Freire-Obregon D, Narducci F, Barra S, Castrillon-Santana M. Deep learning for source camera identification on mobile devices. Pattern Recognit Lett. 2019;126:86–91.

[42] Wang Y, Li Y, Song Y, Rong X. The influence of the activation function in a convolution neural network model of facial expression recognition. Appl Sci. 2020;10(5):1897.

[43] Huang X, Gao L, Crosbie RS, Zhang N, Fu G, Doble R. Groundwater recharge prediction using linear regression, multi-layer perception network, and deep learning. Water. 2019;11(9):1879.

[44] Xiang Z, Yan J, Demir I. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resour Res. 2020;56(1):e2019WR025326.

[45] Verma C, Stoffová V, Illés Z, Tanwar S, Kumar N. Machine learning-based student's native place identification for real-time. IEEE Access. 2020;8:130840–54.

[46] Basu M, Kumar S, Gupta P, Kumar Singh R. A quantitative analysis of machine learning based regressors for pressure reconstruction in particle image velocimetry applications. Fluids Engineering Division Summer Meeting. Vol. 83716, American Society of Mechanical Engineers; 2020 July. p. V001T02A016

[47] Malouf R. A comparison of algorithms for maximum entropy parameter estimation. In COLING-02. The 6th Conference on Natural Language Learning 2002 (CoNLL-2002); 2002.

[48] Andrew G, Gao J. Scalable training of l 1-regularized log-linear models. Proceedings of the 24th International Conference on Machine Learning; 2007 June. p. 33–40

[49] Morales JL, Nocedal J. Remark on "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". ACM Trans Math Softw. 2011;38(1):1–4. Researchgate.net.

[50] Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Trans Math Softw. 1997;23(4):550–60.

[51] Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. Mach Learn. 2000;38(3):257–86.

[52] Choubin B, Zehtabian G, Azareh A, Rafiei-Sardooi E, Sajedi-Hosseini F, Kişi Ö. Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches. Environ Earth Sci. 2018;77(8):1–13.

[53] Choubin B, Malekian A, Golshan M. Application of several data-driven techniques to predict a standardized precipitation index. Atmósfera. 2016;29(2):121–8.

[54] Najafzadeh M, Saberi-Movahed F. GMDH-GEP to predict free span expansion rates below pipelines under waves. Mar Georesour Geotechnol. 2019;37(3):375–92.

[55] Najafzadeh M, Saberi-Movahed F, Sarkamaryan S. NF-GMDH-Based self-organized systems to predict bridge pier scour depth under debris flow effects. Mar Georesour Geotechnol. 2018;36(5):589–602.