USING MACHINE LEARNING METHOD TO FORECAST RIVER WATER LEVELS IN THE BAC HUNG HAI IRRIGATION SYSTEM IN VIETNAM

Hung Viet Ho[1]

Abstract: *In recent years, the application of the Machine Learning (ML) method in analyzing and studying hydrological problems is increasingly becoming common. The numerical models based on ML algorithms have been widely used for predicting river water levels or flowrate. This paper proposes a new approach using one of the applications of deep learning models to predict river water levels in irrigation systems. A predictive model has been developed based on the Long Short-Term Memory (LSTM) neural networks to forecast the water levels upstream of Tranh Culvert in the Bac Hung Hai irrigation system in Vietnam. The findings of this study indicate that although only a modest amount of data is required, the proposed model produced superior results. These results can be used to construct an operating regime for irrigation sluice gates in the Bac Hung Hai system.*

Keywords: Long Short-Term Memory (LSTM), machine learning, Bac Hung Hai, water level.

1. INTRODUCTION

Nowadays, in Vietnam and the world, Machine Learning (ML) algorithms and Artificial Neural Network (ANN) models are increasingly being applied in different fields, including Hydrology and Hydraulics. The ANN models are used to forecast water levels or flowrate in rivers to warn of floods. Moreover, for irrigation systems, the water level forecasting results are essential requirements for building a real-time sluice gate operation regime to serve the water demands for agriculture. Therefore, accurate forecasts contribute to ensuring the safety of people's lives and socio-economic development.

In recent years, there have been two methods for predicting river flow. The first method includes mathematical models based on hydraulic concepts to simulate the hydrodynamic process in an open channel. However, these models generally require a large amount of input data (i.e., bathymetry data, meteorological parameters, and geometric data including river networks, cross-sections) which could be challenging to obtain or not always available (Viet-Hung Truong et al., 2021). Furthermore, the first method has limitations on flood warnings due to the long-running time of the mathematical models (Thirumalaiah and C. Deo, 2000).

The second method is based on the statistical relationship between input and output data for predictions. The ANN model is one of the standard models of the second method. Currently, with the development of computational techniques, the ANN models have been widely used in various aspects of science and engineering because of the simplicity of their model structure. The latter approach may provide a promising alternative to existing computational techniques for river flow forecasting (Kişi, 2011; Hidayat et al., 2014). Several works have focused on forecasting water levels in Vietnam and the world using ANN models, Long Short-Term Memory (LSTM) neural network models, or in combination with ML algorithms (Sung et al., 2017).

LSTM neural network, one of the deep neural network applications (DNN), has been successfully applied in various fields, especially for time sequence problems (Le et al., 2019a).

[1]*Division of Hydraulics, Thuyloi University, Vietnam.*

Regarding water resources, LSTM based models for predicting water levels in agricultural areas were applied and evaluated in Korea and Vietnam. Several studies on ML applications in flood forecasting have also been performed on the river basins in Vietnam (Le et al., 2019a; Le et al., 2019b). An LSTM model also has been applied to predict river water levels downstream of Cau Cat Culvert in the Bac Hung Hai irrigation system in Vietnam (Ho and Ho, 2019).

In this paper, the author has developed LSTM models to forecast the water levels upstream of Tranh Culvert from one to four time-steps of lead time. The Tranh Culvert is one of 11 culverts in the Bac Hung Hai irrigation system, one of the largest irrigation systems in Vietnam constructed in 1958. The forecasted water levels will be used to operate Tranh sluice gates.

2. METHODOLOGY AND INPUT DATA

2.1. LSTM Neural Network

The developed models are based on the LSTM neural networks that Hochreiter and Schmidhuber (1997) introduced to solve multi-step time series forecasting problems. LSTM neural networks can learn long-term dependencies and remember information for long periods. According to Olah (2015), each LSTM neural network has a chain structure with repeating modules. The LSTM module contains four layers that interact in a very distinctive way. The structure of an LSTM module is shown in Figure 1.

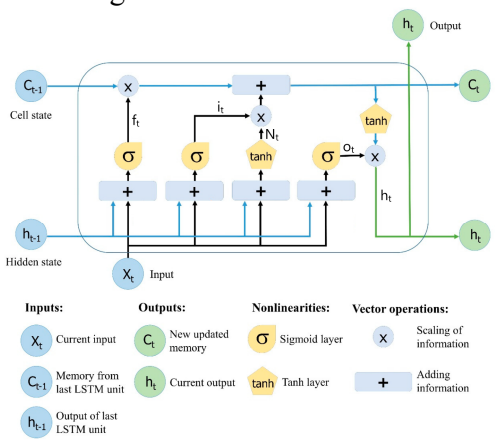

Figure 1. The structure of the LSTM module, Adapted from Le et al., (2019b)

A standard LSTM neural network consists of several modules of memory. There are two states in an LSTM module: the cell state (C_t) is long-term memory, the hidden state (h_t) is short-term memory. The cell state is being transferred from the previous module to the next module. This state is a conveyor belt of data flow running through the network modules, with only minor linear interactions. It allows data to be transmitted along with the network almost unchanged. The input data can be added to the cell state or removed by gates, which are a way to let information go through. An LSTM module has three gates: Forget, Input, Output gates to protect and control the cell state.

The first step in an LSTM network is to decide what information should be removed from the cell state. A sigmoid layer called the "forget gate layer" fixes this. The output value from the module at time t-1 (h_{t-1}) and the input data at time t (X_t) are the input in the module at time t. The output (f_t) is a number from 0 to 1 for each number in the cell state C_{t-1}.

$$f_t = \sigma(W_f.[h_{t-1}, X_t] + b_f) \quad (1)$$

Herein σ is the sigmoid activation function; W_f and b_f are the forget gate's weight matrices and bias.

The second step is deciding what new information from the input (X_t) will be stored in the cell state. This step contains the sigmoid layer called the "Input gate layer" and the tanh layer. The sigmoid layer decides which values should be updated or ignored; the tanh layer generates a vector of new candidate values, N_t, which could be added to the state. In the next step, the two values are multiplied to create an update to the cell state. In that way, the old cell state, C_{t-1}, is updated into the next new cell state C_t (See Equation 4).

$$i_t = \sigma(W_i.[h_{t-1}, X_t] + b_i) \quad (2)$$

$$N_t = tanh(W_n.[h_{t-1}, X_t] + b_n) \quad (3)$$

$$C_t = f_t \times C_{t-1} + i_t \times N_t \quad (4)$$

Here, C_{t-1} and C_t are the cell states at time t-1 and t, while W and b are the cell state's weight matrices and bias, respectively; *tanh* is the Hyperbolic tangent function.

We receive the output values (h_t) based on the cell state (O_t) but are filtered in the final step. First, a sigmoid layer runs to decide which parts of the cell state go to the output. Then, the sigmoid gate output (O_t) is multiplied by the numbers generated by putting the cell state through the tanh layer, pushing the values to numbers from -1 to +1 (See Equation 6).

$$O_t = \sigma(W_o.[h_{t-1}, X_t] + b_o) \quad (5)$$

$$h_t = O_t \times tanh(C_t) \quad (6)$$

Where: W_o and b_o are the weight matrices and bias, respectively, of the output gate.

In this study, the author used some open-source software libraries such as Keras, Sklearn, Numpy, Pandas, and Python 3.6 programming language to develop an LSTM model.

2.2. Input data

The data collected in the study area included data on Bac Hung Hai irrigation system and water levels downstream of Xuan Quan Culvert, upstream of Cau Cat Culvert, downstream and upstream of Tranh Culvert, downstream and upstream of Ba Thuy Culvert over 21 years, from January 1, 2000, to January 1, 2021 (30596 observed values for each gauge station). The time steps of the data series are 6 hours which means that the water level is measured every 6 hours at 1, 7, 13, and 19 h. The locations of the study area and the water level gauge stations of 4 culverts are shown in Figure 2. We measured the water level values and the forecasted results in meters (m) units in this paper. Correlation coefficients of measured values at gauge stations are calculated and presented in Table 1. We will use these coefficients to select the data series and design a predictive model. Moreover, the values in Table 1 show that the data series has a reasonable correlation.

For training, validating, and testing LSTM models, the available data set is divided into three independent parts for different purposes. For model training purposes, the first data section is the observed data series for 19 years from 2000 to 2018. The second data section used for validating the models is the observed water level data measured in 2 years: 2019 and 2020. The third data section is a series of 60-day data in January and February 2021, from January 1 to February 28, used to predict results, evaluate forecasting errors, and test the model performance.

Table 1. Correlation coefficients of observed values at gauge stations

Gauge station	Tranh (US)	Tranh (DS)	Ba Thuy (US)	Cau Cat (US)	Ba Thuy (DS)	Xuan Quan (DS)
Tranh (Upstream-US)	1					
Tranh (Downstream-DS)	0.972	1				
Ba Thuy (Upstream-US)	0.964	0.939	1			
Cau Cat (Upstream-US)	0.958	0.933	0.977	1		
Ba Thuy (Downstream-DS)	0.596	0.618	0.550	0.564	1	
Xuan Quan (Downstream-DS)	0.595	0.586	0.561	0.548	0.368	1

Figure 2. The locations of the study area and the water level gauge stations

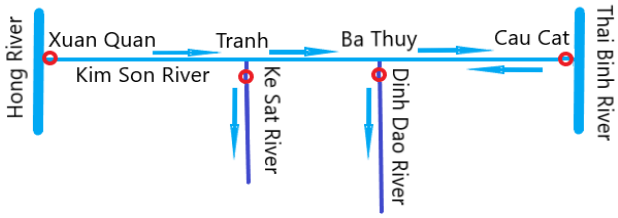

Figure 3. Diagram illustrating the rivers and four culverts in the system

2.3. Evaluation Criteria

To evaluate the model performance in the hydraulic field, we frequently use statistical criteria such as the Nash–Sutcliffe model efficiency coefficient (NSE), the root mean square error (RMSE), and the mean absolute error (MAE). The LSTM model produces good predictive results when the RMSE and MAE values are minimal, and the NSE values are approximately one.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O_i})^2} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \quad (8)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \quad (9)$$

Where: O_i and P_i are observed and predicted values at time t; $\overline{O_i}$ is the mean of observed values; n is the total number of observations.

3. MODEL DESIGN

Each LSTM model has been trained and validated to forecast the water level at one specific lead time. In this paper, the author has developed four LSTM models to predict four cases (from one to four time-steps in advance): the first is the forecasting model for the next one time-step, then the forecasting models for the next two, three, and four time-steps. Several different model parameters have been applied to ensure the best forecasting results. The number of hidden layers is one. The developed models are set to training and validation with a maximum epoch number of 5000. The calculation process will be stopped when the MAE value is unchanged (Early stopping technic). The Adam algorithm with the default learning rate of 0.001 is selected for the model training process to optimize the error. The number of units per module is 35, the optimizer and the learning rate remain constant for all cases. There are four specific cases listed in Table 2.

Table 2. Forecasting cases with a different lead time

Cases	Dependent variable (Forecast target)	Independent variables (Input variables)
1	Water level upstream of the Tranh Culvert for: 6 hours of lead time, or at the time (t + 1)	*Observed water levels in 8 last time-steps, from (t-7), (t-6), (t-5) … to (t) at 6 stations: downstream of Xuan Quan Culvert, upstream of Cau Cat Culvert, downstream*
2	Water level upstream of the Tranh Culvert for: 12 hours of lead time, or at the time (t + 2)	

Cases	Dependent variable (Forecast target)	Independent variables (Input variables)
3	Water level upstream of the Tranh Culvert for: 18 hours of lead time, or at the time (t + 3)	*and upstream of Tranh Culvert, downstream and upstream of Ba Thuy Culvert*
4	Water level upstream of the Tranh Culvert for: 24 hours of lead time, or at the time (t + 4)	

4. RESULTS AND DISCUSSION

4.1. Results for Validating Phase

The proposed model was trained on the first data section (19 years) and validated on the second (2 years) for all suggested cases using the recommended model parameters above. The NSE, RMSE, and MAE values were used to evaluate the model performance by comparing observed and forecasted values. After finishing the training and validating process, the best results of each predictive model have been reported corresponding to the four forecast cases. The results of the validating phase are summarized in Table 3 and presented in Figures 4 to 7, which illustrate the two-year observed data and the predicted values of water levels for all cases.

Table 3. The results of the validating process

Cases	Forecast time (hours)	Sequence length	Number of units	Number of epochs	RMSE (m)	MAE (m)	NSE
1	6	8	35	494	0.056	0.034	0.97
2	12	8	35	303	0.094	0.059	0.92
3	18	8	35	281	0.127	0.081	0.85
4	24	8	35	716	0.141	0.084	0.80

The figures in Table 3 demonstrate that the forecasting models provide impressive results, and forecasting accuracy declines with more extended forecast periods. The values of MAE are stable and less than 0.09 m for all cases. Figures 4 and 5 show that the time of occurrence of the forecasted maximum water level coincides with the actual time measured at the field. For the 18-hour prediction, the predicted top water level is lower than the observed maximum water level (Figure 6), but for the 24-hour prediction, the forecasted maximum water level is slightly higher than the observed value (Figure 7). However, the predicted minimum water level is slightly smaller than the measured water level in both cases.

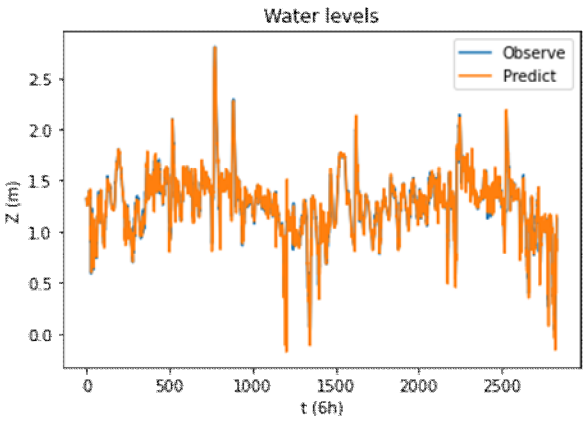

Figure 4. Forecast results in the next 6 hours

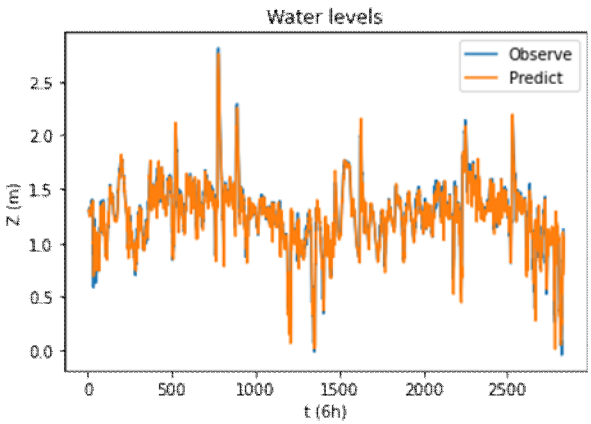

Figure 5. Forecast results in the next 12 hours

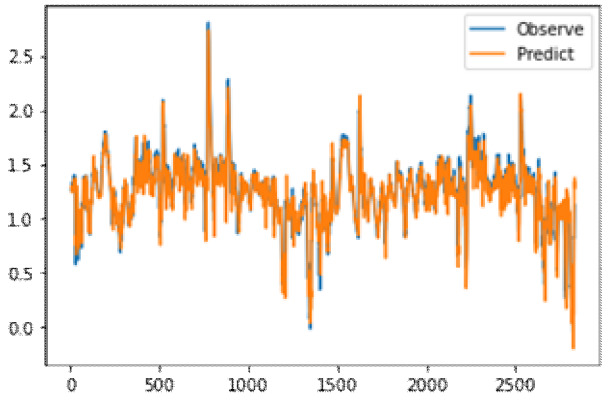

Figure 6. Forecast results in the next 18 hours

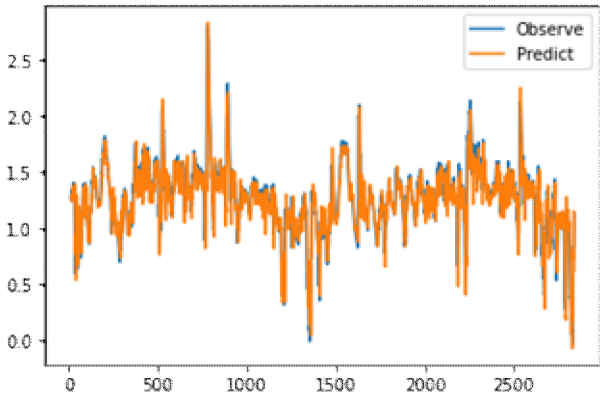

Figure 7. Forecast results in the next 24 hours

4.2. Results for Testing Phase

The LSTM models were tested using the third data section to objectively verify the performance of the predictive models and their accuracy. The input data for February 2021, from the 14th to the 16th, is used for this purpose. Especially, observed data on the 14th and the 15th day of February is a data series consisting of eight time-steps, which is used as input data to predict water levels for February 16. The testing process results for four cases are summarized in Table 4.

Table 4. The results of the testing process for all cases

Cases	Forecast Time (hours)	Predicted Time	Water Level (m)		Absolute error (m)
			Observe	Forecast	
1	6	16/02/2021, 01:00	1.72	1.72	0
2	12	16/02/2021, 07:00	1.73	1.73	0
3	18	16/02/2021, 13:00	1.73	1.72	0.01
4	24	16/02/2021, 19:00	1.74	1.69	0.05

The predictive values reported in Table 4 show that the proposed forecasting models still produce superior results. The forecast results in the next 6 hours and 12 hours are correct. In the case of forecasting 18 hours and 24 hours in advance, the absolute errors of the predictions are not greater than 5 cm. The testing results are better than the validating results, which certify that the developed models produce highly accurate results in both phases.

The above results show that the LSTM Neural Network Model has apparent advantages over other ML-based models when solving time series problems. According to Viet-Hung Truong et al. (2021), calculation methods based on ML algorithms give results with different accuracy. The NSE value ranges from 0.975 to 0.950 when using nine ML-based methods such as GTB, XGBoost, SVM, DT, RF, Adaboost, Hist, LightGBM, DL to forecast the water levels upstream of Tranh Culvert for one time-step.

We can see that the data of Cau Cat Culvert has more influence on the forecast results than the data of Xuan Quan and Ba Thuy Culverts due to the distance between the culverts and their operating modes. The data series with a measurement time of every 6 hours is the most effective for forecasting because it ensures accuracy and gives a long enough forecasting period.

5. CONCLUSIONS

The author of this study developed four predictive models based on the LSTM Neural

Network to forecast the river water levels upstream of Tranh Culvert. The proposed models produce accurate forecast results validated and tested by RMSE, MAE, and NSE values. The predictive results show that the prediction accuracy tends to decrease when forecasting multiple time-steps in advance. As a result of this study, the proposed models can be applied to predict water levels to construct the operating regime of irrigation sluice gates in the Bac Hung Hai system.

Contrary to hydraulic mathematical models requiring a variety of input data such as a river network and its cross-sections, the developed LSTM model uses only the measured water levels available at the target station and the hydrological stations upstream and downstream from the target station to forecast the water levels at the target station for multi-step output. The findings of this study demonstrate the superior performance of the ML-based method in forecasting river water levels. Furthermore, LSTM models have great potential for solving problems of hydrology and hydraulics.

However, a thorough analysis and evaluation of the influence of factors such as space, time, and calculation methods on the forecast results need to be studied further.

REFERENCES

Hidayat, H., Hoitink, A. J. F., Sassi, M. G., & Torfs, P. J. J. F. (2014). *Prediction of discharge in a tidal river using artificial neural networks.* Journal of Hydrologic Engineering, 19(8), p 04014006. doi: 10.1061/(ASCE)HE.1943-5584.0000970.

Ho, V. T., & Ho, V. H. (2019). *Using a recurrent neural network to forecast river water levels affected by tides.* Vietnamese Journal of Water Resources Science and Technology, No 52 (01/2019), pp. 108-116. [in Vietnamese]

Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory.* Neural Computation, 9(8), pp. 1735-1780. doi: 10.1162/neco.1997.9.8.1735.

Kişi, Ö. (2011). *A combined generalized regression neural network wavelet model for monthly streamflow prediction.* KSCE Journal of Civil Engineering, 15(8), pp. 1469-1479. doi: 10.1007/s12205-011-1004-4.

Le, X. H., Ho, H. V., & Lee, G. (2019a). *River streamflow prediction using a deep neural network: a case study on the Red River, Vietnam*. Korean Journal of Agricultural Science, 46(4), pp. 843-856. doi: 10.7744/kjoas.20190068.

Le, X. H., Ho, H. V., Lee, G., & Jung, S. (2019b). *Application of long short-term memory (LSTM) neural network for flood forecasting*. Water, 11(7), p 1387. doi: 10.3390/w11071387.

Olah, C. (2015, August 27). *Understanding LSTM networks.* Retrieved from http://colah. github.io/posts/2015-08-Understanding-LSTMs/

Sung, J., Lee, J., Chung, I.-M., & Heo, J.-H. (2017). *Hourly water level forecasting at tributary affected by main river condition.* Water, 9(9), p 644. doi: 10.3390/w9090644.

Thirumalaiah, K., & Deo, M. C. (2000). *Hydrological forecasting using neural networks.* Journal of Hydrologic Engineering, 5(2), pp. 180-189. doi: 10.1061/(ASCE)1084-0699(2000)5:2(180).etworks with the cuckoo search algorithm. Information, 5(4), p 570. doi: 10.3390/info5040570.

Viet-Hung Truong, Quang Viet Ly, Van-Chin Le, Trong-Bang Vu, Thi-Thanh-Thuy Le, Tuan-Thach Tran, Peter Goethals (2021). *Machine learning-based method for forecasting water levels in irrigation and drainage systems.* Environmental Technology & Innovation 23, 101762.

Tóm tắt:

ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY DỰ BÁO MỰC NƯỚC SÔNG TRONG HỆ THỐNG THỦY LỢI BẮC HƯNG HẢI Ở VIỆT NAM

Trong những năm gần đây, việc áp dụng phương pháp Học máy (ML) trong phân tích và nghiên cứu các vấn đề thủy văn, thủy lực ngày càng trở nên phổ biến. Các mô hình số dựa trên thuật toán ML đã được sử dụng rộng rãi để dự báo mực nước sông hoặc lưu lượng dòng chảy. Bài báo này đề xuất một cách tiếp cận mới bằng việc sử dụng một trong những ứng dụng của mô hình học sâu để dự báo mực nước sông, kênh trong các hệ thống thủy lợi. Một mô hình dự báo đã được phát triển dựa trên mạng nơ ron Bộ nhớ gần xa (LSTM) để dự báo mực nước ở thượng lưu Cống Tranh trong hệ thống thủy lợi Bắc Hưng Hải ở Việt Nam. Nghiên cứu này cho thấy, với một lượng dữ liệu khiêm tốn, mô hình mà tác giả đề xuất đã tạo ra kết quả vượt trội. Có thể sử dụng các kết quả này để xây dựng chế độ vận hành cho các cống tưới - tiêu trong hệ thống Bắc Hưng Hải.

Từ khóa: LSTM, học máy, Bắc Hưng Hải, mực nước.

Ngày nhận bài: 13/12/2021

Ngày chấp nhận đăng: 31/12/2021