DEVELOPING A FLOOD FORECASTING SYSTEM WITH MACHINE LEARNING AND APPLYING TO GEOGRAPHIC INFORMATION SYSTEM

Jirayu PUNGCHING[1] and Sitang PILAILAR[1]

DOI: 10.21163/GT_2023.181.01

ABSTRACT:

Floods are natural disasters that can damage lives, property, and the economy. Therefore, it is necessary to have a reliable and accurate flood forecasting system to provide early warning in time. Although several Mathematical models have been developed and used to forecast floods continuously for decades, most require up-to-date and specific physical data, including a high experience user, to provide and interpret the result. It is an obstacle for use in remote areas with incomplete information and a lack of specialists. This study, therefore, developed a real-time flood forecasting system with Machine Learning by applying a 2-variable sliding window technique to restructure the data, which can solve the problem of data limitation. Thung Song District Nakhon Si Thammarat Province was selected to test this newly developed model. By importing the water level data of two water level observed stations, SWR025 at the upstream and NKO001 at Thung Song Municipality, into five machine learning algorithms (Linear Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Random Forest) for forecasting the water level every 30 minutes for the next 5 hours. Their performance was compared by the MSE, MAE, and $R^{2,}$ which ranged from 0.006-0.013, 0.044-0.063, and 0.518-0.750, respectively. The Random Forest was the most efficient algorithm for the 3-hour forecast with an efficiency value of MSE 0.006, MAE 0.044, and R^2 0.75. The developed ML flood forecasting model was validated by the flood data in November 2021 and showed good agreement. Then, the extent of the inundation area was evaluated by the mathematical model. Next, the water depth and surface elevation were transformed and applied to GIS. Finally, the flood risk areas on Google Maps under that specific rainfall are promptly notified to the people three hours before the flood occurs.

Key-words: *Flood Forecasting, Flood Maps, Machine Learning, Linear Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, Thung Song Municipality*

1. INTRODUCTION

Floods are threatening natural disasters that cause damage to life, property, and economic security at both local and national levels. There has been a tendency to occur more frequently, more violently, and become increasingly difficult to predict due to many factors such as changing rain behavior from the past, land-use change, encroachment on people's waterways, and shallowness of natural waterways, etc. Urbanization significantly changes hydrological behavior by changing the infiltration rate, baseflow, and lag time and influencing flow patterns (Rafiei et al., 2016). It emphasizes the need for a reliable flood forecasting system to provide timely early warning for citizens and government agencies to take preventative or evacuation measures to alleviate the damage. Additionally, a real-time flood inundation map that indicates the extent of flood risk areas corresponding to the amount of precipitation upstream helps deal with potential disasters.

Although flood prediction mathematical models have been constantly developed and applied for hazard assessment and extreme event management for decades, most physical models require accurate physical data of the basin for model setting up (Chang et al., 2020). The recorded events over a long period are also necessary. Obviously, the more extended and accurate data results in more precise forecasting closer to reality, but it requires substantial computational effort. Furthermore, changes in the physical data of the area inevitably affect the accuracy of forecasts over the expectations of system users. There are also quite specific data import restrictions. In addition, the area risk factors cannot be readily increased or decreased, and aspects of expertise and user experience can significantly

[1] *Department of Water Resource Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand; jirayu.pun@ku.th, fengstpl@ku.ac.th*

influence the forecasting system's accuracy. Likewise, several studies suggested a gap in the short-term prediction capability of physical models (Castabile and Macchione, 2015). For example, Van de Honert and McAneney, 2011 published the failure in flood prediction in Queensland, Australia. Shrestha et al., 2013 also reported the systematic error caused by the unreliable numerical weather prediction model. The inappropriate short-term prediction that needs significant improvement has encouraged the usage of advanced data-driven models, e.g., machine learning (ML). Machine Learning Neural Networks model how the brain performs a particular task or function of interest (Haykin, 2008). The use of a Neural Network is computation through the process of "learning" that is adaptive in the machine, known as "Machine Learning (ML)." The computing algorithms can improve automatically through experience and by using information for training (Learning) to find reasonable approximate solutions to complex problems. Researchers have moved from physical-based flood forecasting methods to ML techniques over the last two decades (Chang et al.,2018). Chang et al., 2014 conducted the prediction of flood volume using hybrid Som and dynamic neural networks. Lohani et al., 2014 and Badrzadeh et al., 2015 applied ML for runoff prediction in real-time flood forecasting. Granata et al., 2016 examined the runoff in an urban area and compared the performances between the support vector regression and the EPA's stormwater management model. The study confirmed the usefulness of SVR for urban flood prediction.

After estimating the runoff and flood volume according to various return periods of rainfall, the predicted inundation map has been determined. In Kemaman river, West Malaysia, Chang et al., 2018 developed the flood inundation forecast model by combining two artificial neural networks, Self Organizing Map (SOM) and Recurrent nonlinear autoregressive with exogenous inputs (RNARX). The forecast inundation maps were then generated ahead of the event. The accuracy of inundation maps RMSE was found to range from 0.08 to 0.68, and R^2 ranged from 0.94 to 1. A few seconds carried out three to twelve hours ahead of inundation maps. Chang et al., 2020 predicted the inundation maps to build a real-time warning system for people in the risk area. He developed the effective real-time pluvial flood forecasting AI platform in Taoyuan City, Taiwan, with 6,000 sets of color-classified rainfall hyetograph maps. Three hundred thousand simulated flooding maps were used for learning with Inception V3 Convolutional Neural Network (CNN). The method's accuracy is shown by comparing AI-generating map images and simulation model images in RMSE ranging from 0.02 to 0.44. This AI platform can predict pluvial floods one-hour ahead; the total execution time is less than 6 minutes. Kim, 2020 provided the expected inundation area in Gangnam, Korea, due to simultaneous rainfall. The Probabilistic Neural Network (PNN) was used to perform the return period for observed rainfall, and the Support Vector Machine (SVM) and Self-Organizing Mapping (SOM) were used to predict the flood volume and inundation maps.

Although the above studies successfully applied machine learning techniques for flood forecasting and inundation map-generating, the rainfall data, including the rainfall return period statistic, has been crucial. Unfortunately, flood forecasting has still been problematic because of insufficient rainfall data, especially in remote urban areas. Therefore, this study attempts to apply the sliding window techniques for a machine learning-based model to cope with the future flood situation under data scarcity from a limited number of stations. Five algorithms of ML were selected and applied with the proposed sliding window techniques. The most suitable algorithms that resulted in the best agreement of water level at the concerned station were then applied under the specific return period-rainfall. Finally, the example of an application with GIS was conducted. Thus, the inundation map of the study area, Thung Song Municipality, Nakhon Si Thammarat Province, was generated to support flood surveillance and disaster alleviation.

2. STUDY AREA AND METHODOLOGY

2.1 Study Area

Thung Song Municipality in Nakhon Si Thammarat Province locates in the southern part of Thailand. It covers an area of 802.977 sq. km. The geographical characteristic is the high land in the northeast and low land on the Middle side, where the elevation ranges from +19.00 MSL to +1,255.00

MSL. The main river that flows through Thung Song Municipality is Khlong Thalao, as shown in **Fig. 1**, with a steep upstream slope of approximately 0.002. Thus, the runoff from the upstream, namely Namtok Yong, Khlong Thaloan, and Khlong Purk, flow through the city quickly. Consequently, it causes flash floods almost yearly, such as the large February 2006.

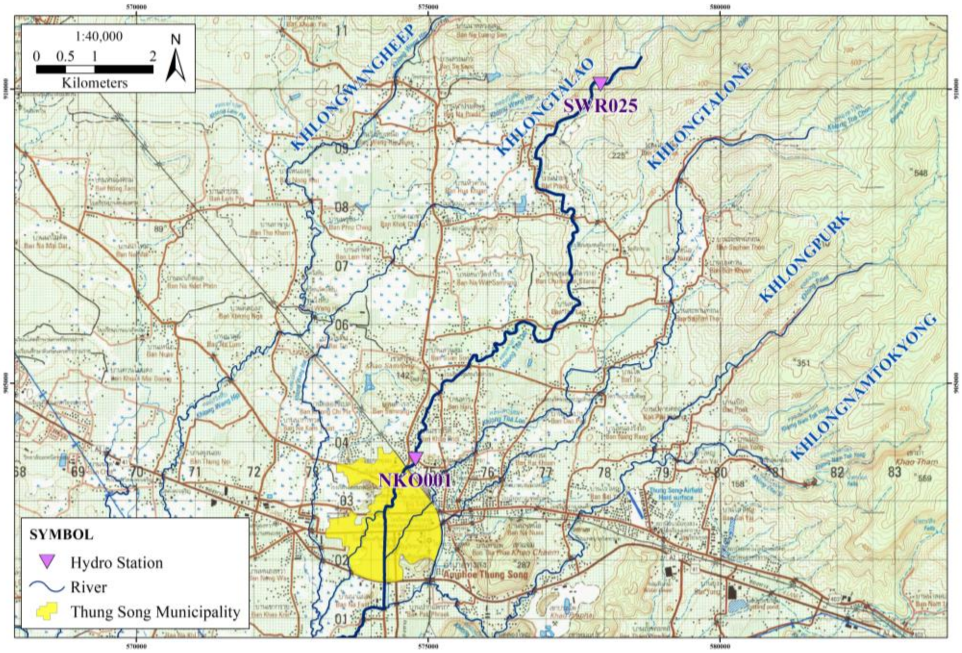


Fig. 1. Thung Song Municipality; Study Area and Route of Runoff Flow through the City.

On February 13-14, 2006, Thung Song accumulated 160 mm. of rainfall, estimated to have a volume of 3.80 million cum, causing the flow in Khlong Thalao of 151 cum/s. As a result, a flash flood occurred at 0200 a.m. February 14 without warning. The water level rose along the river to +51.42 MSL at Sapan Talad Kaset Bridge, +50.47 MSL at Asia Road Bridge, and +51.80 MSL at the Fire Station.

2.2 Flood Event for the Study

To generate the inundation map under a particular flood, this study reviewed data on flood events from 2020 to 2022. The severe flooding was recorded in December 2020, as they appear flood stains on the wall of Thung Song Municipality Office, measuring 1.20 meters, as shown in **Fig. 2**.

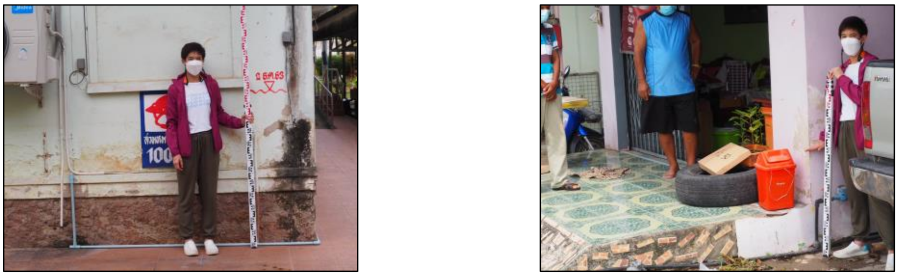


Fig.2. Flood Stains on the wall of Thung Song Municipality Office
due to Flood Event in December 2020.

The water level at the municipality, station NKO001, increased from +52.000 MSL at 0700 p.m. December 1 to the peak of +55.80 MSL at 0200 p.m. December 2, 2020, as shown in **Fig. 3.**

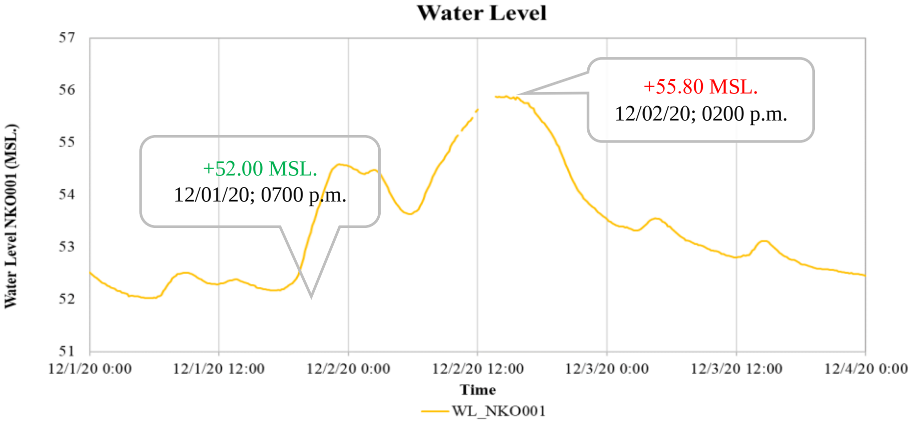


Fig. 3. Water Level of Flood Event in December 2020 at Thung Song Municipality, Station NKO001.

Another flood event occurred in November 2021, where the flood depth at Thung Song Municipality was about 30 cm, as shown in **Fig. 4.**

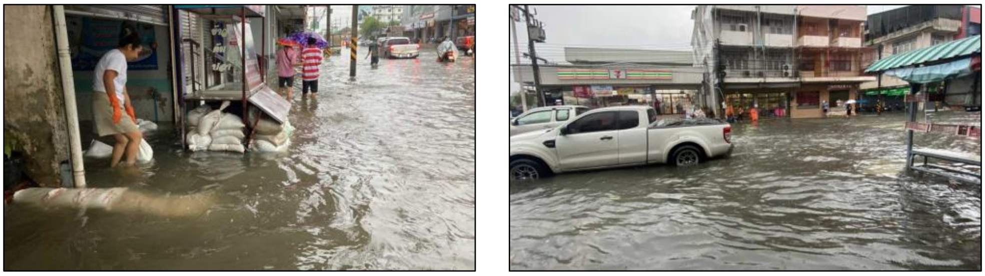


Fig. 4. Flood Event in November 2021 at Thung Song Municipality (Thairath, 2021).

The flood peak reached within five hours, from 51.33 MSL to 53.26 MSL, as shown in **Fig.5.** However, under the water situation, Thung Song Municipality has a surveillance and warning practice by monitoring the water level at the upstream station. As soon as the officials notice the water level at Station SWR025, as shown in **Fig.1**, rises to the watchful level, they will inform people to lift things high, arrange sandbags and avoid traveling in the path that may be flooded. Generally, the water travel time from Station SRW025 to Station NKO001 at the municipality, which is 10 km apart, is about 3 hours; thus, the people have enough time to deal with the flood situation.

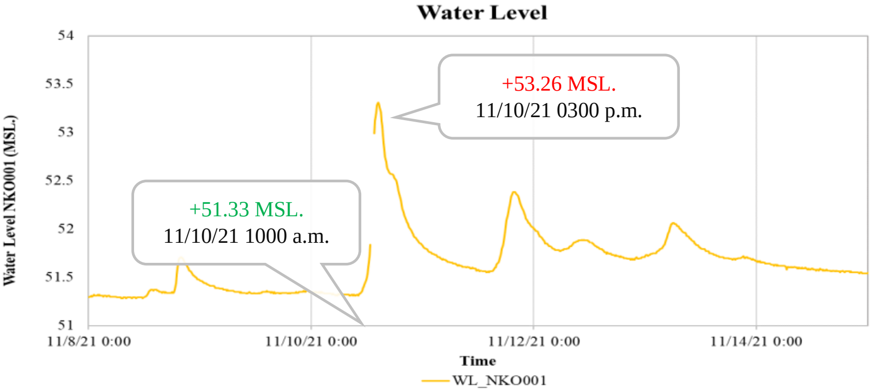


Fig. 5. Water Level of Flood Event in November 2021 at Thung Song Municipality, Station (NKO001).

2.3 Methodology

This study attempts to develop the ML flood forecasting model with the best algorithm. Then, the inundation map, which is practically helpful in flood warning and protection, can be generated. Thus, the works have been divided into six steps, as shown in **Fig. 6.**

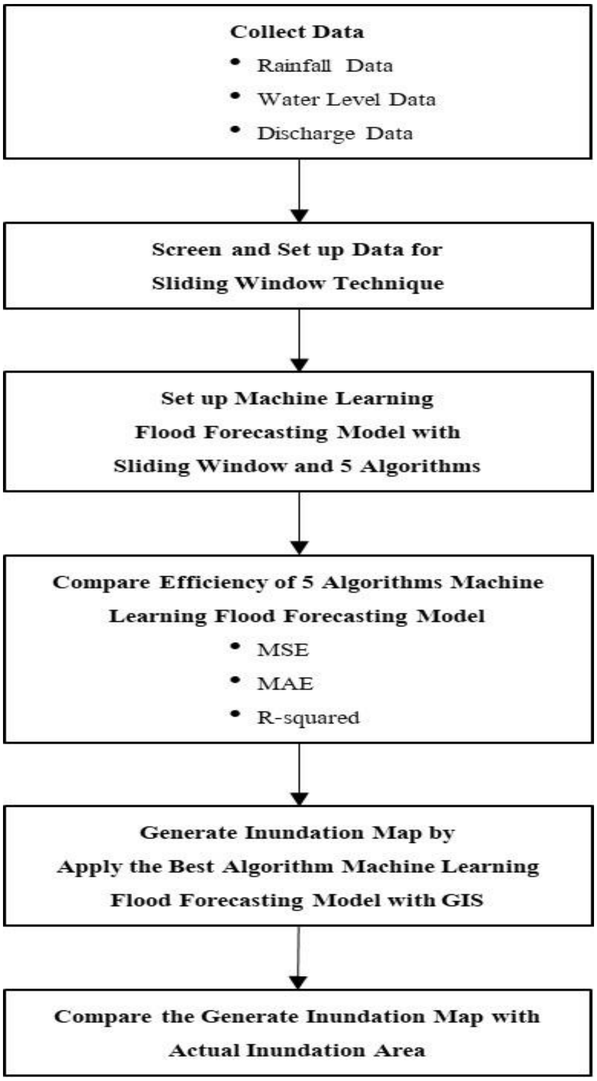

Collect Data
Rainfall Data
Water Level Data
Discharge Data
Screen and Set up Data for
Sliding Window Technique
Set up Machine Learning
Flood Forecasting Model with
Sliding Window and 5 Algorithms
Compare Efficiency of 5 Algorithms Machine
Learning Flood Forecasting Model
MSE
MAE
R-squared
Generate Inundation Map by
Apply the Best Algorithm Machine Learning
Flood Forecasting Model with GIS
Compare the Generate Inundation Map with
Actual Inundation Area

Fig. 6. Steps of Work.

2.3.1. Data Collection and Handling with Sliding Window Techniques

To forecast the water level downstream where the city is located, the developed model has to learn the flow pattern and relationship between the upstream and downstream water levels. Thus, supervised learning for Machine Learning is considered, and the available data has been explored.

Along Khlong Thalao, which flows directly through Thung Song Municipality, two water level stations belong to Hydro-Informatics Institute (HII). The data throughout 2021 have been selected for ML algorithms testing and learning. Nevertheless, for every 10 minutes of data, 61,566 timesteps, it was found that 3,152 data at SWR025 and 1,045 data at NKO001 were missing. The information in that period was, therefore, cut off.

Considering the consistency of water level data at the two stations 10 km apart, they have a good agreement, as shown in **Fig. 7** and **Fig. 8**. The travel time is about three hours; thus, the time lag could be observed.

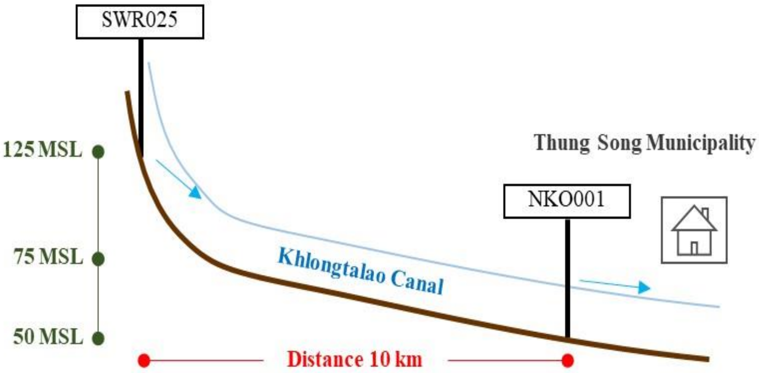


Fig.7 Longitudinal Profile of SWR025 and NKO001 Water Level Stations.

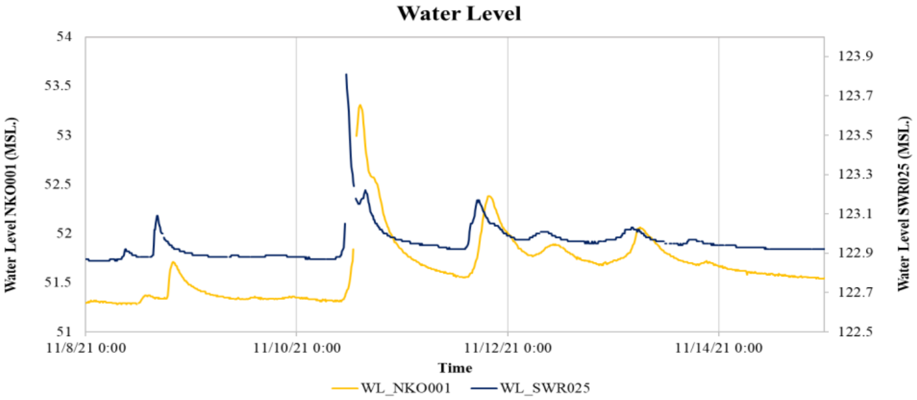
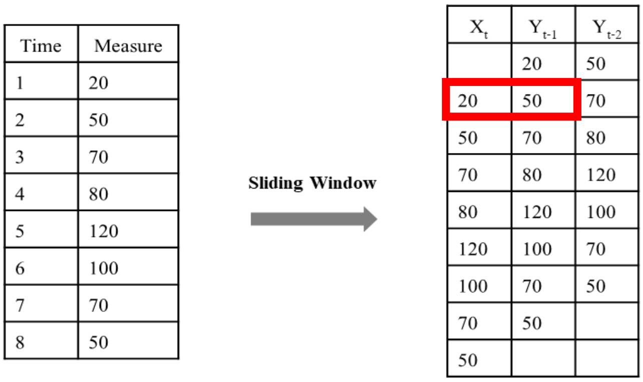


Fig. 8. Water Levels at SRW025 and NKO001 stations during the Flood Event in November 2021.

Time series forecasting can be framed as a supervised problem. The use of previous time steps to predict the next time step is called the sliding window method. Brownlee, 2020 introduced the sliding window technique in reframing data for predicting future values with regression problems. This technique has advantages in predicting multiple time-step ahead of one output variable. The function of regression in the algorithm learns the mapping function from the input variables (X) to predict the output variable (y) as $y = f(X)$. Sequential data analysis is used to restructure the data for supervision for predicting the output of variable (y) in the present time of variable input (X). The input and output sequence data can be used to directly supervise the machine learning algorithm without restruct data ($y_t = f(X_t)$), where t is time. To predict variable (y) multiple steps ahead, the sequence data have to be restructed for the learning algorithm by using previous timesteps for supervised data ($y_{t-n} = f(X_t)$), where n is the number of timesteps (Dietterich, 2002). Variable (y) is shifted to many steps, such as $y_t, y_{t-0.5}, y_{t-1}, \dots, y_{t-5}$ in the data frame, as shown in **Fig. 9.**

This study applied the sliding window for time series forecasting for multi-variable learning. Two water-level stations' data were restructed. The water level at the upper stream station was the variable (X) to predict the water level at Thung Song Municipality, which was the variable (y), multiple steps ahead, as shown in **Fig. 10.**

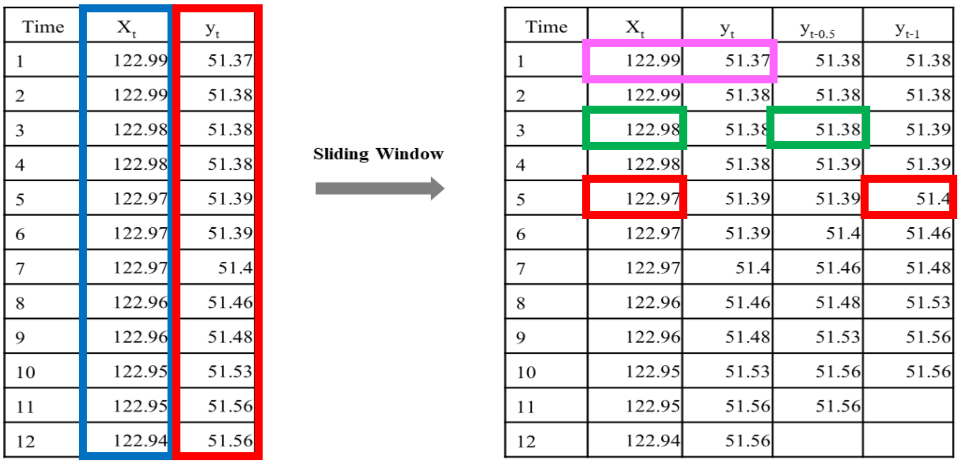


Time	Measure
1	20
2	50
3	70
4	80
5	120
6	100
7	70
8	50

Sliding Window

X_t	Y_{t-1}	Y_{t-2}
	20	50
20	50	70
50	70	80
70	80	120
80	120	100
120	100	70
100	70	50
70	50	
50		

Fig. 9. Sliding Window for Time Series Forecasting Data Frame.



Time	X_t	y_t
1	122.99	51.37
2	122.99	51.38
3	122.98	51.38
4	122.98	51.38
5	122.97	51.39
6	122.97	51.39
7	122.97	51.4
8	122.96	51.46
9	122.96	51.48
10	122.95	51.53
11	122.95	51.56
12	122.94	51.56

Sliding Window

Time	X_t	y_t	$y_{t-0.5}$	y_{t-1}
1	122.99	51.37	51.38	51.38
2	122.99	51.38	51.38	51.38
3	122.98	51.38	51.38	51.39
4	122.98	51.38	51.39	51.39
5	122.97	51.39	51.39	51.4
6	122.97	51.39	51.4	51.46
7	122.97	51.4	51.46	51.48
8	122.96	51.46	51.48	51.53
9	122.96	51.48	51.53	51.56
10	122.95	51.53	51.56	51.56
11	122.95	51.56	51.56	
12	122.94	51.56		

Fig. 10. Sliding Window with Two Variables in this Study.

The water level at SWR025 (X_t) was used to supervise the water level at NKO001, which was restructured to arrange a sequence for prediction every 30 minutes, and five hours ($y_t, y_{t-0.5}, y_{t-1}, \dots, y_{t-5}$) ahead. Plotting the water level of SWR025 with sliding window multi-step forecasting water level of NKO001 is shown in **Fig.11**. After arranging the sequence for prediction, every forecasting timestep of the data frame has to start and end at the same period, and the row of missing data must be removed. The total timestep of the data frame for machine learning is 51,109 timesteps. Data have been divided into 60%, 20%, and 20% for training, validation, and testing. Then the data was input into five machine-learning regression algorithms.

2.3.2. Machine Learning Flood Forecasting Model Setting up

Mosavi et al., 2018 defined ML as a field of artificial intelligence (AI) used to induce regularities and patterns, providing more straightforward implementation with low computation cost, as well as fast training, validation, testing, and evaluation, with high performance compared to physical models, and relatively less complexity. Furthermore, several ML algorithms, e.g., artificial neural network (ANNs), neuro-fuzzy, support vector machine (SVM), and support vector regression (SVR), were reported as practical flood forecasts. However, the capability of forecasting depends on their training, whereby the system learns the target task based on past data. Géron, 2019 classified ML systems according to the amount and type of supervision they get during the training into four categories: supervised learning, unsupervised learning, semisupervised learning, and reinforcement learning.

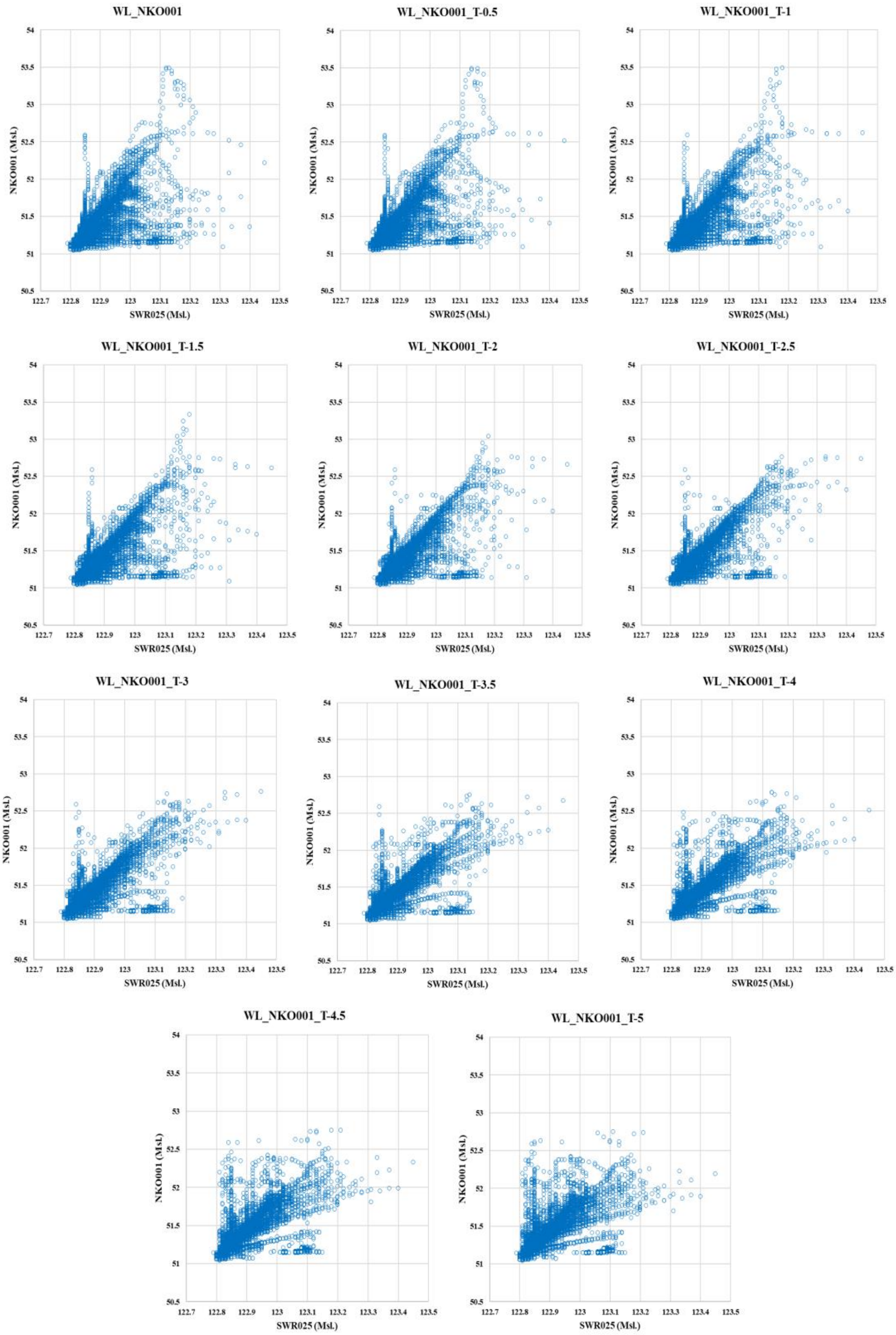


Fig. 11. Distribution of Water Level data of SWR025 and NKO001 after arranging the sequence for prediction.

In this study, five supervised learning regression algorithms, which are Linear Regression (LR), Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forest model (RF), have been selected. The detail of each algorithm is as briefly described.

1) The Linear regression model (LR) predicts by simply computing a weighted sum of the input features plus a constant called the *bias term*, as Linear regression model prediction Equation (1) shown in **Fig. 12** (Géron, 2019).

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{1}$$

$\hat{y}$ is the predicted value,
n is the number of features,
x_1 is the number of features, and
θ_j is the jth model parameter (including the *bias term* θ_0 and the feature weights θ_1, θ_2, …, θ_n)

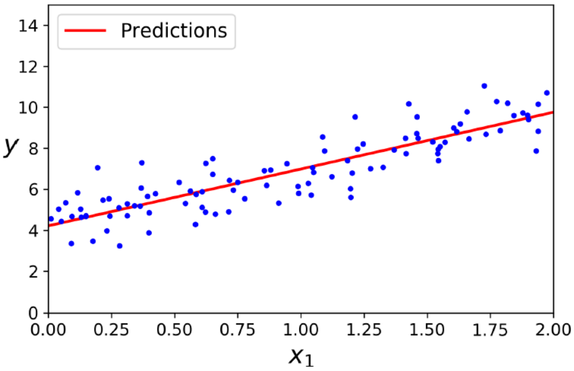


Fig. 12. Linear Regression and Model Prediction (Géron, 2019: p.118)

2) The Support Vector Machine model (SVM) finds hyperplanes by the maximum linear margin of support vectors. Suppose data cannot classify as linear. Then, kernel function handling with nonlinear datasets adds more features, such as polynomial, sigmoid and radial basis functions (RBF). **Fig. 13** shows two linear SVM Regression models train on random linear data, one with a large margin (ϵ = 1.5) and the other with a small margin (ϵ =0.5) (Géron, 2019).

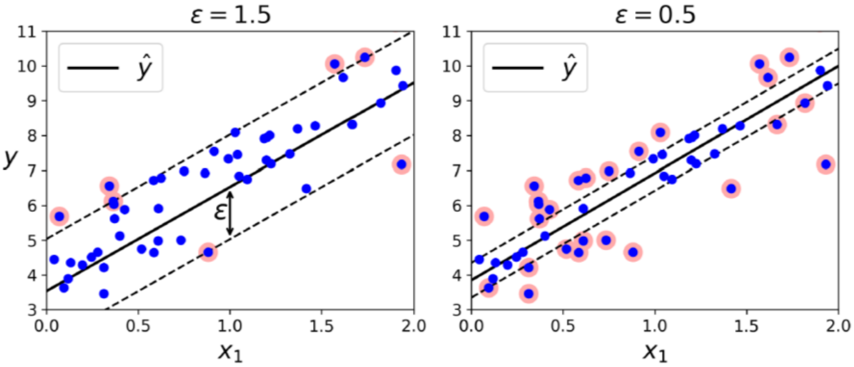


Fig. 13. SVM Regression (Géron, 2019: p.165).

3) K-Nearest Neighbors model (KNN) predicts the category of test samples according to training sample *k*, which is the closest neighbor to the test sample and inserts it into a category with the greatest

probability. Near or far distances to neighboring points can be calculated using the Euclidean distance equation (2) (Andrian, 2019). **Fig. 14** shows a sample of new instance in the K-Nearest Neighbors model (Géron, 2019).

$$D(a,b) = \sqrt{\sum_{k=1}^{d} (a_k - b_k)^2} \tag{2}$$

D is the distance between points
a is the known point
b is the unknown point
d is the dimension of the point being measured
k is the value of neighboring data measured

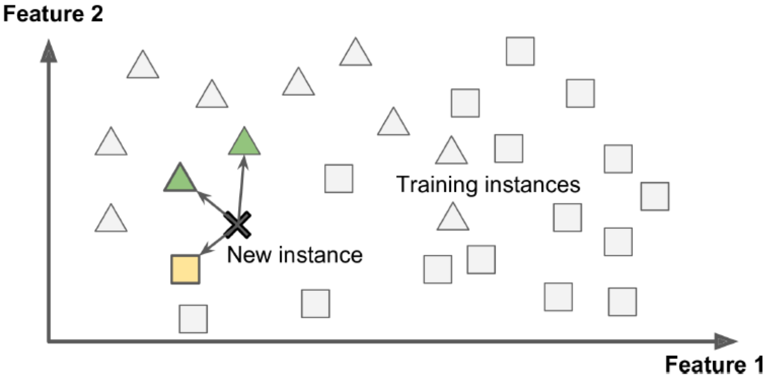


Fig. 14. K-Nearest Neighbors Model (KNN), (Géron, 2019: p.19).

4) The Decision Tree model (DT) expresses data with a tree-like graph based on the rules or conditional statements of the variables, subdivides them into similar data types by separation rules, and continues this classification until the final classification criteria are satisfied. In a DT, through a binary recursive partitioning process, split variables and split points that minimize the mean squared error (MSE) are identified in each step. In addition, "pruning" is performed to determine the tree size that minimizes the MSE using cross-validation to prevent overfitting. A Decision Tree (DT) conceptual diagram is shown in **Fig. 15 (**Lee et al., 2020).

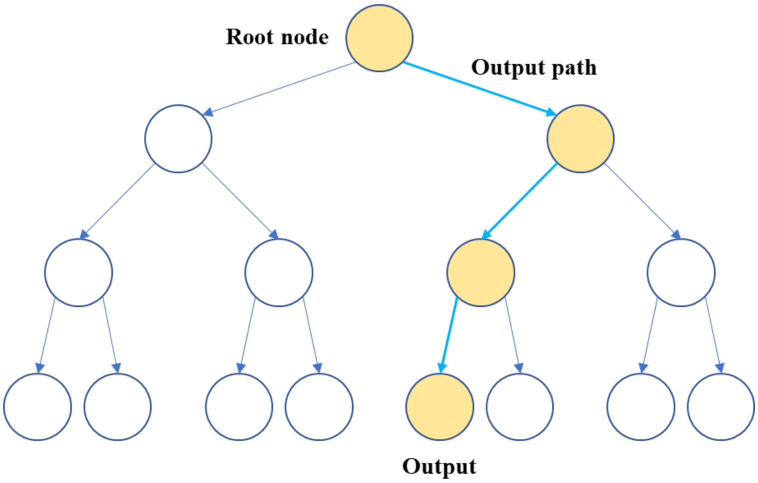


Fig. 15. Conceptual Diagram of a Decision Tree (DT); modified from: (Lee et al., 2020: p.4).

5) The Random Forest model (RF) is an ensemble of Decision Trees, generally trained via the bagging method. Random Forest increases the number of Decision Trees and creates multiple training data sets from one data set. It has improved predictive power as a result of creating multiple DTs through multiple learnings and then combining multiple DTs. The Conceptual diagram of a Random Forest (RF) is shown in **Fig. 16** (Lee et al., 2020).

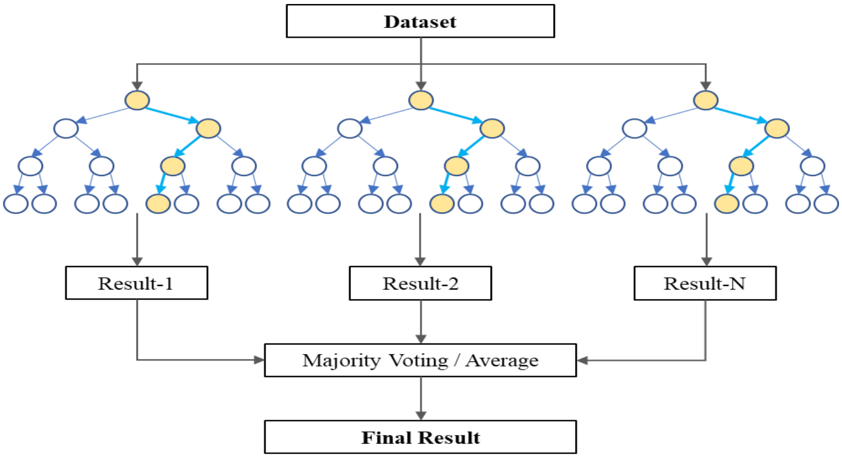


Fig. 16. Conceptual Diagram of a Random Forest (RF); modified from: (Tbico, 2021).

Regression is a modeling method that learns relationships within provide and predicts using this relationship. The kind of model that learns and predicts continuous data is called a regression problem. Regression was typically found in the following scenario (Haykin, 2008). One of the random variables is considered to be of particular interest; that random variable is referred to as a dependent variable or *response*.

The remaining random variables are called independent variables or *regressors*. Their role is to explain or predict the response's statistical behavior and the response's dependence on the regressors. It includes an additive *error* term to account for uncertainties in how this dependence is formulated. The error term is called the expectational or *expectational error*, which is used interchangeably.

2.3.3. Five Algorithms Forecasting Efficiency Comparison

As previously illustrated how various algorithms function; LR, SVM, KNN, DT, and RF. LR often solves regression problems by examining relationships between two or more variables. It supports only linear solutions, while SVM, KNN, DT, and RF support both linear and non-linear solutions. Therefore, when there is a large number of features with fewer data-set (with low noise), LR may outperform DT/RF. In this study, five algorithm performances of forecasting have been evaluated by three parameters: MSE, MAE, and R-squared. Mean Square Error (MSE) is an average squared deviation of the predicted value, Mean Absolute Error (MAE) indicates how the predicted values are far from the measured values, and the R-squared is the strength of the relationship between the predicted values and actual values (Yafouz et al., 2021), as equations shown below:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4}$$

$$R - squared = 1 - \frac{(y_i-\hat{y}_i)^2}{(y_i-\bar{y}_i)^2} \tag{5}$$

where N is the number of observations, y is the vector of the actual values, and $\hat{y}$ is the vector of the predicted values (Surakhi et al., 2021).

2.3.4. Inundation Map Generating by Applying ML Flood Forecasting with GIS

Once the best algorithm for the ML Flood Forecasting model has been obtained, the model is input with water level data at the upstream station, SWR025, to forecast the water level rise at municipal sites. Next, the water level overflows of the bank will be analyzed for the extent of the flood area by Mike 11HD, Mike21FM, and Mike Flood mathematical models.

MIKE 21 Flow Model (Mike21FM) is a modeling system for 2D free-surface flows. It is applicable to the simulation of hydraulic and environmental phenomena in lakes, estuaries, bays, coastal areas, and seas (DHI, 2017b). It requires information on bathymetry. Thus, a Digital Elevation Model (DEM) resolution of 20 m^2 covering all river networks was used to generate a bathymetry file in this study, as shown in **Fig.17.** Mike Flood is coupling Mike11HD and Mike21FM to simulate flood maps in hourly steps. Its linkages set up to Lateral Link allow a string of MIKE21FM cells/elements to be laterally linked to a given reach in MIKE11HD, either a branch section or an entire branch. Flow through the lateral link is calculated using a structural equation. This link is particularly useful for simulating overflow from a river channel onto a flood plain, as the concept shown in **Fig.18** (DHI, 2017c). Further, the extent of the inundation area is illustrated with GIS techniques.

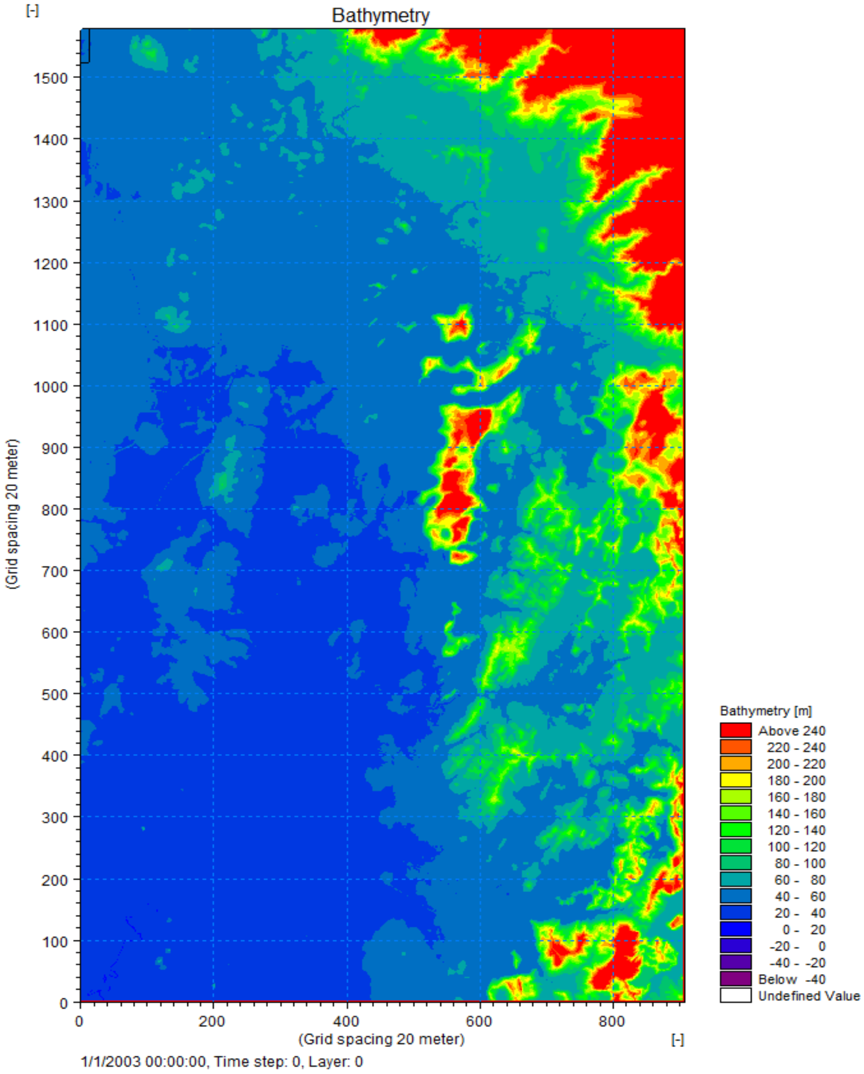


Fig. 17. Bathymetry of the Study Area.

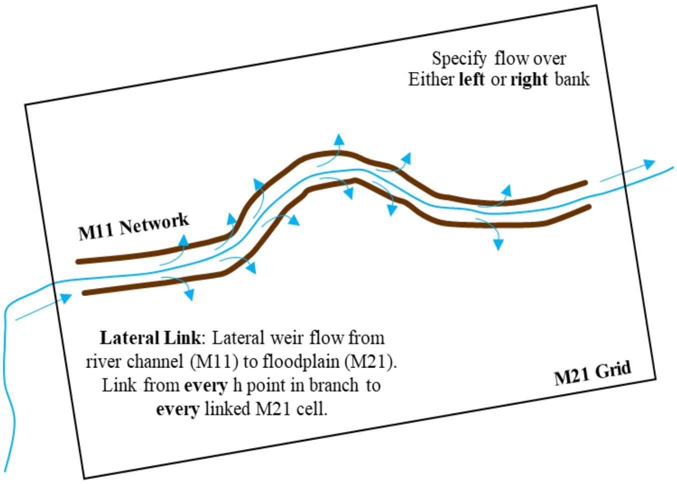

Specify flow over
Either left or right bank
M11 Network
Lateral Link: Lateral weir flow from
river channel (M11) to floodplain (M21).
Link from every h point in branch to
every linked M21 cell.
M21 Grid

Fig. 18. Application of Mike 11 and Mike 21 on Flood Area Analysis: modified from: (DHI, 2017c: p.17).

3. RESULTS AND DISCUSSIONS

3.1. Data Scattering Analyze by Sliding Window Technique

To figure out the scattering pattern of the water level data at SWR025 and NKO001, the data at these two stations at present and multiple future steps have been analyzed and viewed by scatter plot, as shown in **Fig. 19**. The relationship shows a linear trend with many noises. The noise of the data was largely found below the 45° line from the Timesteps 0-3 hours, and a large amount of noise above the 45 ° line in the Timesteps 3.5-5 hours, while the noise at the Timesteps of 3 hours is comparatively less than the other Timesteps. It indicates that the rise and fall of the water level at NKO001 at 3 hours is close to the current SWR025 water level.

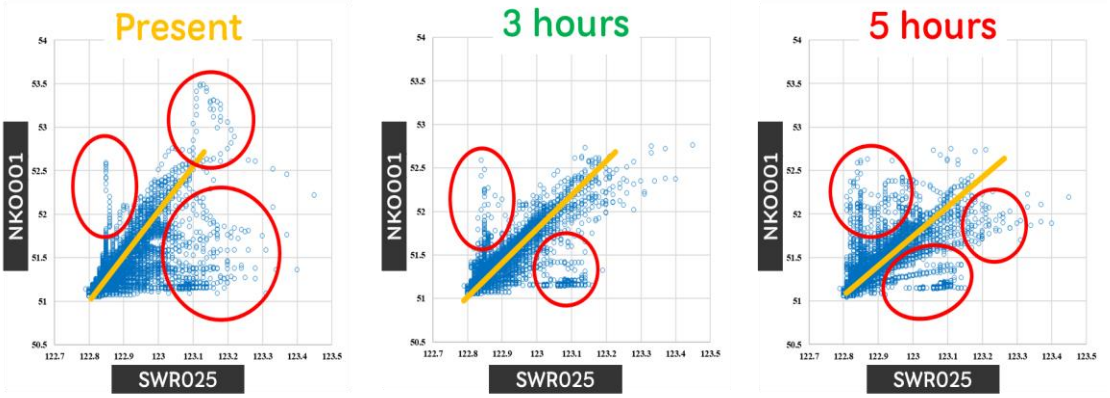


Fig. 19. Water Level Data Scattering Patterns in Two Stations-Scattering Analyzes.

To finish the sequence for prediction, every forecasting timestep of the data frame has to start and end at the same period, and the row of missing data must be removed. The total timestep of the data frame for machine learning is 51,109 timesteps. Later, the data were divided into 60%, 20%, and 20% for training, validation, and testing.

3.2. Machine Learning Forecasting Models Performance

Since the appropriate time step forecasting was decided according to the scattering pattern of the water level data, five algorithms have been supervised to learn with water level data of SWR001 and NKO001 with multi-step forecasting. In addition, MSE, MAE, and R^2 compared the algorithms' performance.

The performance of five algorithms on multi-step water level forecasting has been evaluated, and the results are presented in **Table 1**. The Random Forest algorithm performs best among other algorithms by MSE 0.006, MAE 0.044, and R^2 0.75. Further, most algorithms show that the best forecast time is 3 hours ahead, except for K-Nearest Neighbors, whose best forecast time is 2.5 hours. It is consistent with the flow time from station SWR025 to station NKO001, which takes 3 hours.

Table 1.

Performance Comparison among Five Algorithms on Multi-step Water Level Forecasting.

Algorithm	**Model Performance**	**Forecasting time (hr.)**										
		+0	**+0.5**	**+1.0**	**+1.5**	**+2.0**	**+2.5**	**+3.0**	**+3.5**	**+4.0**	**+4.5**	**+5.0**
Linear Regression	MSE	0.013	0.012	0.012	0.011	0.01	0.01	**0.01**	0.01	0.01	0.01	0.011
	MAE	0.063	0.061	0.06	0.058	0.056	0.055	**0.054**	0.054	0.054	0.055	0.056
	R^2	0.518	0.54	0.56	0.581	0.594	0.603	**0.604**	0.598	0.584	0.567	0.548
Support Vector Machine	MSE	0.01	0.009	0.009	0.008	0.008	0.007	**0.007**	0.007	0.007	0.007	0.008
	MAE	0.052	0.05	0.053	0.052	0.052	0.052	**0.051**	0.051	0.051	0.052	0.053
	R^2	0.641	0.66	0.67	0.693	0.7	0.706	**0.708**	0.704	0.697	0.685	0.671
K-Nearest Neighbors	MSE	0.01	0.009	0.008	0.008	0.007	**0.007**	0.007	0.007	0.007	0.007	0.008
	MAE	0.054	0.052	0.051	0.05	0.049	**0.047**	0.047	0.046	0.046	0.047	0.049
	R^2	0.628	0.657	0.676	0.697	0.713	**0.724**	0.722	0.719	0.708	0.694	0.676
Decision Tree	MSE	0.009	0.008	0.007	0.007	0.007	0.006	**0.006**	0.006	0.006	0.007	0.007
	MAE	0.051	0.049	0.048	0.047	0.046	0.045	**0.044**	0.044	0.045	0.045	0.046
	R^2	0.672	0.697	0.718	0.734	0.738	0.746	**0.749**	0.744	0.733	0.719	0.704
Random Forest	MSE	0.009	0.008	0.007	0.007	0.006	0.006	**0.006**	0.006	0.006	0.007	0.007
	MAE	0.051	0.049	0.048	0.047	0.046	0.045	**0.044**	0.044	0.045	0.045	0.046
	R^2	0.672	0.697	0.719	0.736	0.741	0.748	**0.75**	0.744	0.734	0.719	0.704

Additionally, the forecasted water level at the NKO001 station from Random Forest Algorithm shows good agreement with the observed water level, as shown in **Fig. 20.** Thus, the developed ML model's accuracy of 3 hours ahead forecasting is validated.

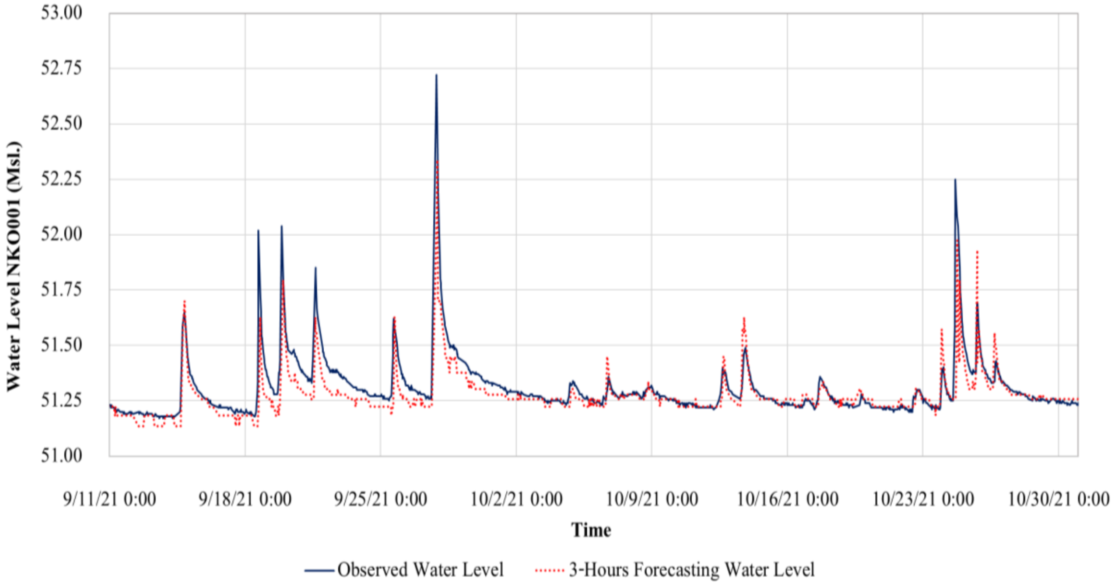


Fig. 20. Comparison between the Observed Water Level and the 3-hours Forecasting at Station NKO001; flood events in 2021.

Then, based on the water level at the upstream station, SWR025, the forecasting of the water level at the surveillance station, NKO001, can be conducted.

3.3. Dataset of Flood Maps

Since the Random Forest algorithm showed the best performance in ML flood forecasting, as previously described in 3.2, it was then applied to another set of water levels that caused the flood in June 2022. The comparison between the forecasted water level and the observed at the downstream station, NKO001, is shown in **Fig. 21.** The forecasting accuracy was then reconfirmed with the MSE 0.009, MAE 0.056, and R^2 0.75.

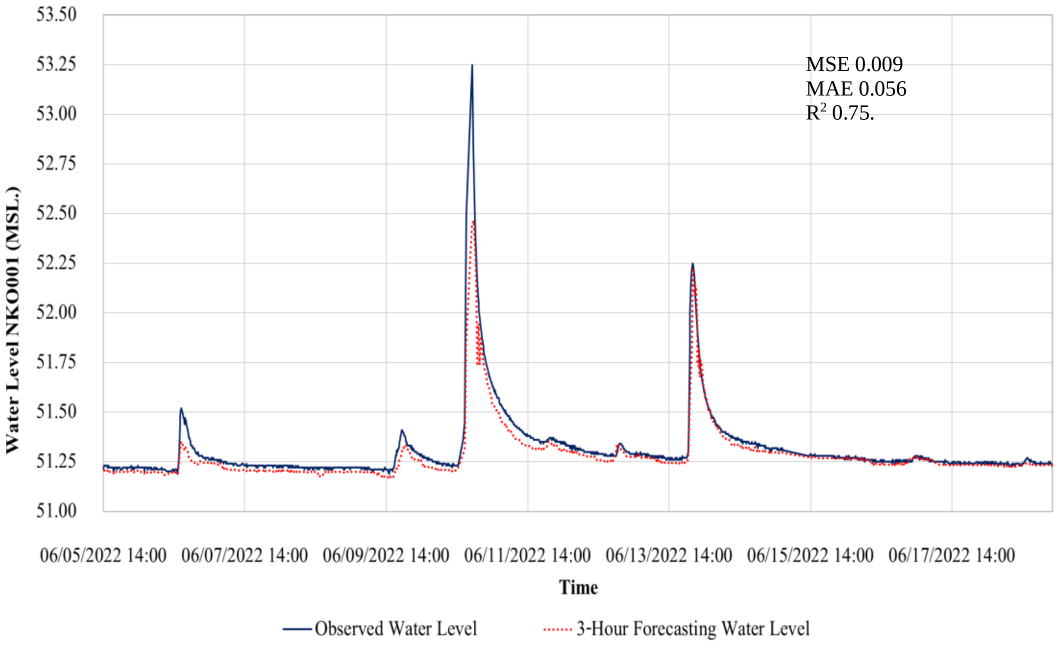


Fig. 21. Comparison between the Observed Water Level and the 3-hours Forecasting at Station NKO001, flood events in June 2022.

As an example of applying the results of flood forecasting to create flood maps, further, the extent of the flood area was evaluated by generating a flood map with the Mike Flood, coupling 1D and 2D models where the Mike11HD models the river and Mike21FM models the floodplain with lateral links.

The simulation consumed the amounts of input data and time processing the flood maps; the water depth and surface elevation are obtained in the ".dfs2" file type. It is shortly transformed into a ".KML" file type by Mike to Google Earth; to store the flood maps with a resolution of 20x20 m. in the database. It can be updated in real-time according to the reported water level upstream frequency. Thus, the people in flood-prone areas can properly take flood alleviation action in time.

The generated flood map is overlayed with the actual inundation area obtained from the municipality, as shown in **Fig. 22.** The agreement means that whenever the water level is reported at the upstream station, the inundation area can be evaluated. Then, the flood warning procedures can proceed based on the generated flood map.

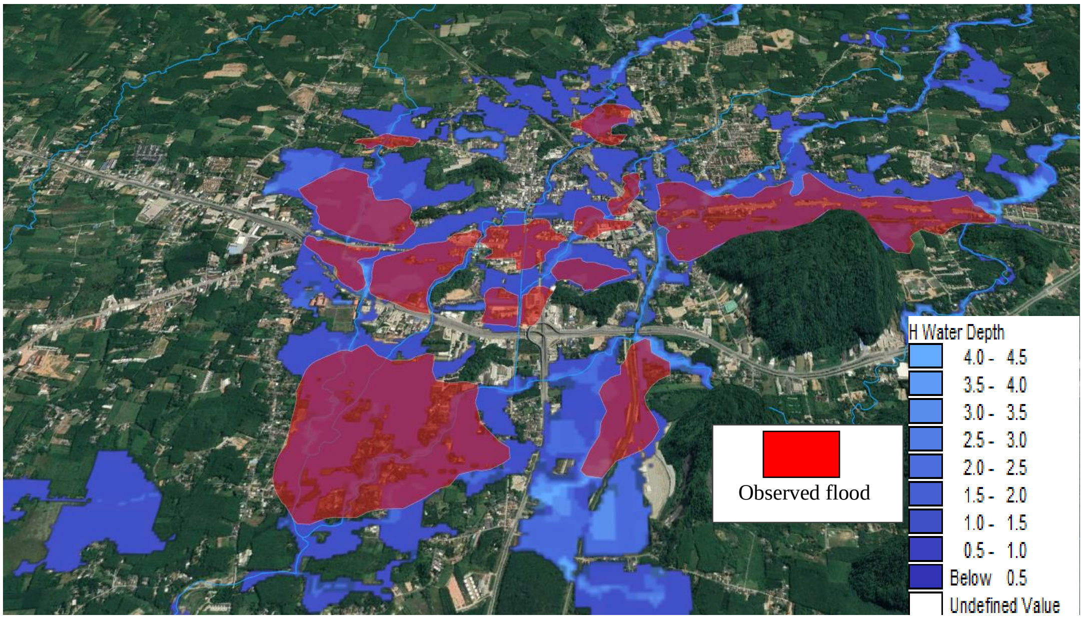


Fig. 22. Flood Map of Thung Song Municipality due to the peak flow on June 10, 2022.

3.4. Flood Forecasting System

Finally, the Machine Learning Forecasting Model and the Flood Maps Database have been gathered as the Flood Forecasting System. Since the real-time water level data of SWR025 and NKO001 are input into the Flood Forecasting System, the Machine Learning Forecasting Model forecasts the water level +0, +0.5, +1.0, … ,+5.0 hours ahead. The forecasted result will be matched with the flood map from the database at the same water elevation at NKO001. Whether the flood maps indicate the risk area in Thung Song Municipality, the system will send an alarm to the Thung Song Municipality officers and related agencies.

The procedure for surveillance and alerting people at risk of flooding is shown in **Fig. 23.**

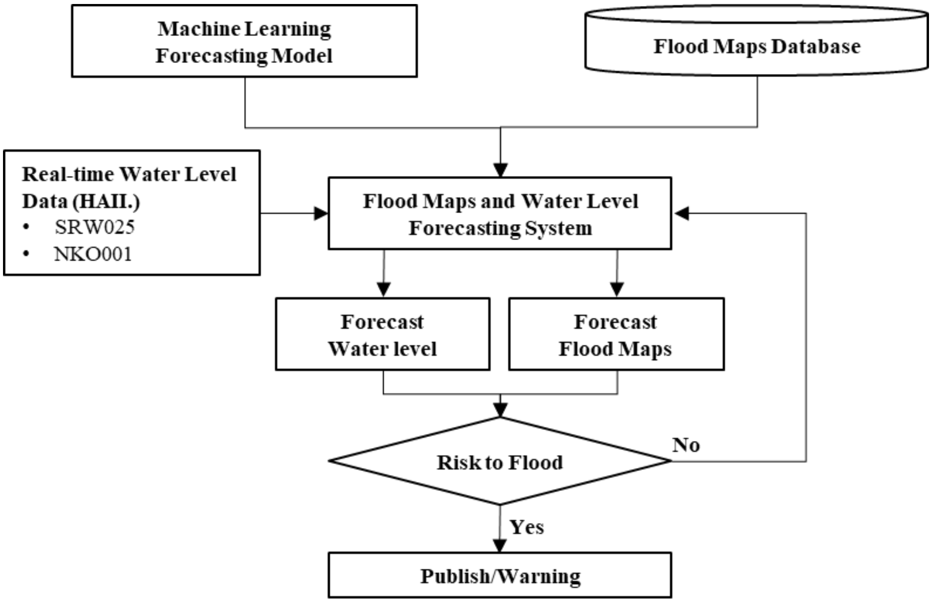

Machine Learning Forecasting Model
Flood Maps Database
Real-time Water Level Data (HAII.)
• SRW025
• NKO001
Flood Maps and Water Level Forecasting System
Forecast Water level
Forecast Flood Maps
Risk to Flood
No
Yes
Publish/Warning

Fig. 23. Overall Procedure of Flood Forecasting and Surveillance.

4. CONCLUSION

The short-term flood warning information in remote areas can be problematic due to insufficient rainfall data and the limitation of water level observed stations. In addition, most numerical flood forecasting models require detailed physical data and are highly computational and time-consuming. Thus, the related agencies may not be able to warn people in risk areas accurately and timely.

This study developed a machine learning-based model to cope with the future flood situation in Thung Song Municipality in Nakhon Si Thammarat under data scarcity. Using the water level relationship between the upper station, SWR025, and the downstream station, NKO001, a 2-variable sliding window technique has been applied. First, the data has been arranged and restructed; the water level at the upstream station was the variable input to predict the water level downstream. After the water level data scattering has been analyzed, the best prediction sequence can be decided. Then, five Machine Learning algorithms which are Linear Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Random Forest, were supervised to learn with water data with multi-step forecasting. Their performance was compared based on the MSE, MAE, and R^2. As a result, the Radom Forest performed the best by MSE 0.006, MAE 0.044, and R^2 0.75, under 3 hours ahead of forecasting.

Consequently, the developed ML flood forecasting model has been validated by inputting the water level at the upstream station, which is 10 km from the municipality. The 3-hour forecasted water level at Thung Song Municipality station showed good agreement with the observed flood in November 2021. Furthermore, the "Flood Forecasting System" can be conducted in the next step as the extent of the flood area has been evaluated by generating a flood map with the Mike Flood mathematical model. The water depth and surface elevation are shortly transformed by Mike to Google Earth. As a result, the 3-hour forecasted inundation area under that specific rainfall, which can be real-time updated, is helpful for flood warning and disaster alleviated procedures.

In summary, even in remote areas with limited water level measurement stations and incomplete data, Flood warnings can also be notified by a developed machine learning-based model. The introduced 2-variable sliding window technique and the algorithm must be selected to suit the data set and then applied with the Mike Flood mathematical model. Further, the extent of flood-prone areas can be shown in conjunction with the GIS system. The information under specific rainfall is promptly notified in advance.

ACKNOWLEDGEMENT

This research is supported in part by the Graduate Program Scholarship from the Graduate School, Kasetsart University.

REFERENCES

Andrian, R., Naufal, M.A., Hermanto, B., Junaidi, A., and Lumbanraja, F.R. (2019). k-Nearest Neighbor (k-NN) Classification for Recognition of the Batik Lampung Motifs. *Journal of Physics*. 1338(2019) 012061. https://doi:10.1088/1742-6596/1338/1/012061

Brownlee, J. (2020). *Introduction to Time Series Forecasting with Python, How to Prepare Data and Develop Models to Predict the Future*. Machine Learning Mastery.

Chang, L.C., Amin, M.Z.M., Yang, S.N., and Chang, F.J. (2018). Building ANN-Based Regional Multi-Step-Ahead Flood Inundation Forecast Models. *Water*. 10. 1283. https://doi:10.3390/w10091283

Chang, D.L., Yang, S.H., Hsieh, S.L., Wang, H.J., and Yeh, H.C. (2020). Artificial Intelligence

Methodologies Applied to Prompt Pluvial Flood Estimation and Prediction. *Water*. 12. 3552. https://doi:10.3390/w12123552

DHI. (2017a). *MIKE 11 A modeling system for rivers and channels*, User Guide. Denmark.

DHI. (2017b). *MIKE 21 Flow model FM, hydrodynamic module*, User Guide. Denmark.

DHI. (2017c). *MIKE FLOOD, 1D-2D Modelling*, User Manual. Denmark.

Dietterich, T.G. (2002). Machine Learning for Sequential Data: A Review. *LNCS*. 2396.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.)*. Penguin Books. O'Reilly Media.

Haykin, S. (2008). *Neural Networks and Learning Machines (3rd ed)*. McMaster University. Pearson Prentice Hall.

Hydro-Informatics Institute (HII.), (2021, November 25). Flood Forecasting System. http://www.thaiwater.net/floodforecast

Kim, H. I., and Han, K.Y. (2020). Inundation Map Prediction with Rainfall Return Period and Machine Learning. *Water*. 12. 1552. https://doi:10.3390/w12061552

Lee, J.Y., Choi, C., Kang, D., Kim, B.S., and Kim, T.W. (2020). Estimating Design Floods at Ungauged Watersheds in South Korea Using Machine Learning Models. *Water*. 12. 302. https://doi:10.3390/w12113022

Rafiei Emam, A., Mishra, B., Kumar, P., Masago, Y., & Fukushi, K. (2016). Impact Assessment of Climate and Land-Use Changes on Flooding Behavior in the Upper Ciliwung River, Jakarta, Indonesia. *Water*, 8(12). doi:10.3390/w8120559

Surakhi, O., Zaidan, M.A., Fung, P.L., Motlagh, N.H., Serhan, S., Khanafseh, M.A., Ghoniem, R.M., and Hussien, T. (2021). Time-Lag Selection for Time-Series Forecasting Using Neural Network and Heuristic Algorithm. MDPI. *Electronics*. 10. 2518. https://doi:10.3390/electronics10202518

Tibco. (2021). *What is a Random Forest?* [Online] Available from: https://www.tibco.com/reference-center/what-is-a-random-forest [Accessed April 20, 2022]

Thairath. (2021). *Flooding on November 29, 2021, in Nakhon Si Thammarat Province*. [Online] Available from: https://www.thairath.co.th/news/local/south/2253247 [Accessed April 25, 2022]

Yafouz, A., Ahmed, A.N., Zaini, N., Sherif, M., Sefelnasr, A., and Shafie, A.E. (2021). Hybrid deep learning model for ozone concentration prediction: comprehensive evaluation and comparison with various machine and deep learning algorithms. *Engineering Applications of Computational Fluid Mechanics*. 15. 1. (902-933) https://doi.org/10.1080/19942060.2021.1926328