The Efficiency of Input Determination Techniques in ANN for Flood Forecasting, Mun Basin, Thailand

Tawee Chaipimonplin[1], Thaveesak Vangpaisal[2]

[1]Department of Geography, Chiang Mai University, Chiang Mai, Thailand

[2]Civil Engineering Department, Ubon Ratchathani University, Ubon Ratchathani, Thailand

[*1]tawee.c@cmu.ac.th; [2]thaveesak.v@ubu.ac.th

***Abstract*-The selection of input variables from different techniques can provide different artificial neural network (ANN) model performances. This study utilizes an ANN model to forecast the water level at the M.7 gauge station for t+48 hour. Three objectives to be investigated are: (1) to compare the efficiency of four input determination techniques (cross correlation, stepwise regression, cross correlation with stepwise regression, and genetic algorithm); (2) to investigate the number of hidden nodes from 1-2n+1 node; and (3) to compare two different learning algorithms (Levenberg Marquardt-LM and Baysian Regularization-BR). Results demonstrate that the cross correlation and the cross correlation with stepwise regression techniques are best for selecting input variables to forecast water levels at t+48 hours at the M.7 gauge station. Additionally, the use of only one hidden node is sufficient for the ANN model, and LM and BR learning models perform similarly.**

Keywords- Input Determination; Mun Basin; Flood Forecasting; Artificial Neural Network

I. INTRODUCTION

Hydrological models such as MIKE 11 [1], TANK [2], IHACRES [3], URBS [4] and Artificial Neural Network (ANN) [5] have been developed and applied to flood forecasting. In addition, the ANN model is an example of a data driven method. The advantages of the ANN model are that it does not require physical data or field data, and requires less computation time than other approaches. There are several methods for improvement to the typical ANN model such as the selection of input variables from input determination techniques [6, 7], the addition of extra input variables [8], the selection of different transfer functions [9], and the selection of different learning algorithms [10-12]. In addition, input variable determination techniques can be classified into six approaches: (1) trail and error or an ad hoc approach [13]; (2) sensitivity analysis approaches [14, 15]; (3) statistical approaches, such as correlation analysis [16], multiple regression [17], and Partial Mutual Information-PMI [18, 19]; (4) neural network-specific approaches, such as pruning algorithms [20] and saliency analysis [6]; and (5) data driven approaches such as Self Organizing Maps (SOMs); and (6) genetic algorithms [21].

ANN model performance also depends on learning algorithms, input variables and the number of hidden nodes; when there are too many hidden nodes, the ANN model overfits the training data and is not able to generalize to new data. Therefore, the suitable number of hidden node is generally determined by trail and error [22].

The Matlab software includes many type of learning algorithms such as Scaled Conjugate Gradient (SCG), Levenberg-Marquardt (LM) and Bayesian Regularization (BR); each algorithm offers a different approach. For example, SCG is exemplary for pattern recognition, LM is the fastest algorithm, and BR provides an automate regularization for improving generalization [23].

The objectives of this study are: (1) to determine the efficiency of input determination techniques, and (2) to investigate the effective number of hidden nodes by the comparison of two learning algorithms (LM and BR) to forecast the water level at the Mun Basin, Thailand. The inputs for the models are obtained from four input determination techniques, also supervision selection and the use of all input variables.

II. METHODOLOGY

The Mun Basin is the largest river basin in Thailand, with an area of approximately 71,000 km^2. The basin is located in the north eastern region of Thailand, adjacent to the Chi and Khong-Isan Basins in the north. The Mun Basin is divided into 31 sub-basins, which expand across 15 provinces in the region (Fig. 1).

DOI: 10.5963/JWRHE0402002

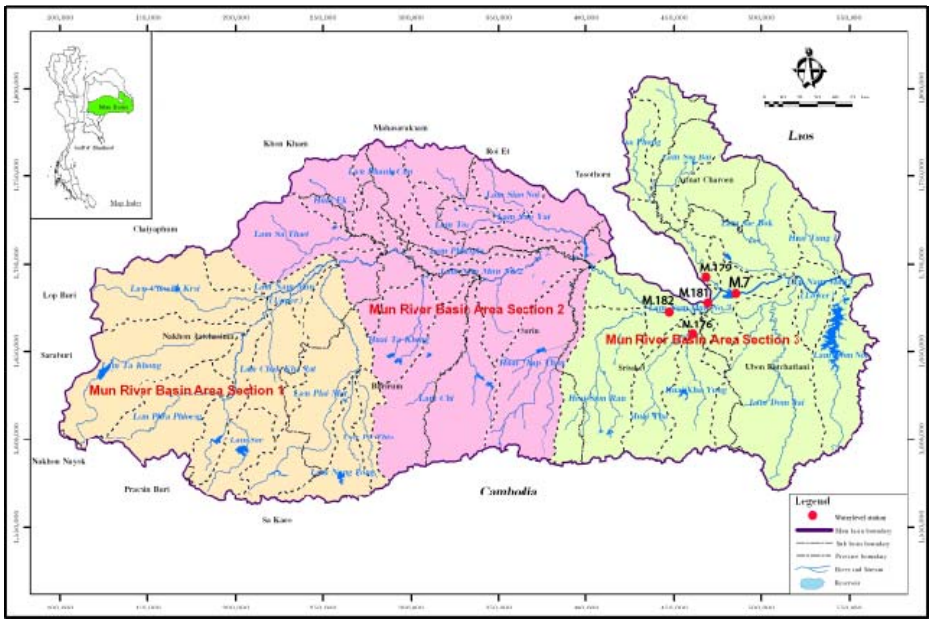

Fig. 1 Mun Basin [27]

The Mun Basin can be divided into three sub-areas according to differences in topographical conditions and land use: the upper, middle and lower parts of the basin. The annual average rainfall is 1,399 mm. The annual average rainfall is highest in the east and the northeast areas of the basin, and lower in the inner area. The annual runoff in the rainy season between May and October is approximately 16,300 million m^3 or 85.9% of the total annual runoff; the annual runoff in the dry season between November and April is approximately 2,680 million m^3 or 14.1% of the annual runoff. The Mun River is located in the lower part of the basin, particularly in Ubon Ratchathani province. Ubon Ratchathani province is located at the lower part of the Mun Basin, and is approximately 630 km east of Bangkok. In the past 50 years, there were 23 flood events (occurrences during which the water level exceeded 7.0 m) in this area. The highest recorded level of river runoff was between September and October. The highest water level of +117.7 m above mean sea level (12.76 m) was recorded in the year 1978, which had dramatic effects on the communities, environment, and economy.

Inputs for the ANN model are the recorded water level from four upper stations (M.181, M.179, M.176 and M.182) and the M.7 station, to forecast the water level at M.7. The M.181 and M.182 stations are the gauge stations located upstream of M.7 on the main river, while the M.176 and M.179 are gauge stations located on the tributary of the main river (Fig. 2). The longest distance of 76 km is from M.176 to M.7, followed by 72 km from M.182 to M.7.

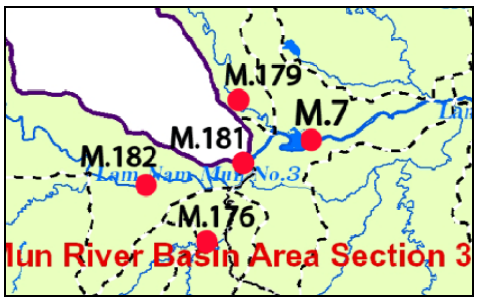

Fig. 2 Location of five gauge stations

The data in the study is recorded per hour. However, the available data from all five stations covers only the time frame 2007-2011, with four total flood events and no floods in the year 2010 (Fig. 3). At the M.7 station, a water level of 7 m represents the onset of flood in the study area. Data from five gauge stations can enhance the variable data from time t to time t-12 hour (12 hours before time t: t_12) and t-24 hour (24 hours before time t: t_24); total input consists of 15 input variables.

In order to explore the efficiency of the learning algorithms and the number of hidden nodes, the experiments are divided into five different input variables. Performances of the LM and BR models were compared.

DOI: 10.5963/JWRHE0402002

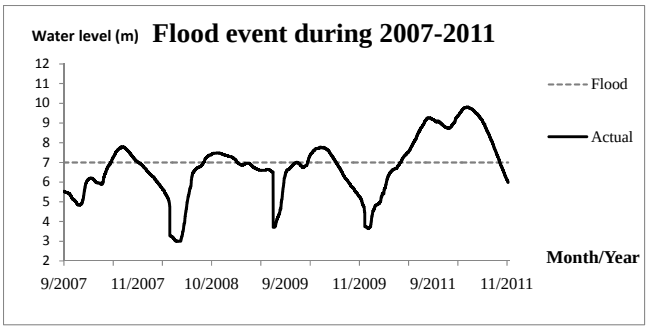

Fig. 3 Flood events during 2007-2011

Four different input determination techniques were investigated: (1) cross correlation (BC) is a method of calculating the relationship between input variables and output variables greater than 0.9; (2) stepwise regression (BS) is a multiple regression method, which removes the smaller correlation variables and selects from the remaining input variables; (3) cross correlation and stepwise regression (BCS) is a combination of BC and BS which first selects the input variable by cross correlation and then uses stepwise regression to select from the remaining inputs; and (4) genetic algorithm (BG) is based on biological evolution and natural selection and was developed by Holland [24]. Additionally, supervised selection (BSp) consists of input variables at time t of each station; all input variables (BA) are explored to indicate the differences in input variable selection (Table 1). The cross correlation and stepwise regression were calculated from SPSS software, and WEKA software was used to calculate the genetic algorithm.

To compare the four input determination techniques, all 15 input variables are significant as 12, 13, 11 and 9 are selected by cross correlation, stepwise regression, combination techniques and the genetic algorithm, respectively. The data from the M.176 station at time 12 and 24 hours seem to be less significant input variables.

TABLE 1 INPUT VARIABLES OF T+48 HR

Input	Input Determination Techniques					
	BC	*BS*	*BCS*	*BG*	*BSp*	*BA*
M.7	X	X	X	X	X	X
M.7_12	X	X	X	X		X
M.7_24	X	X	X	X		X
M.181	X	X	X	X	X	X
M.181_12	X	X	X	X		X
M.181_24	X	X	X			X
M.179	X	X	X	X	X	X
M.179_12	X	X	X	X		X
M.179_24	X					X
M.176		X		X	X	X
M.176_12		X				X
M.176_24						X
M.182	X	X	X		X	X
M.182_12	X	X	X	X		X
M.182_24	X	X	X			X
Total	12	13	11	9	5	15

ANN Models were developed to compare the two learning algorithms, LM and BR. The number of hidden nodes for each model depends on the number of input nodes, which varied according to the five input determination techniques. The hidden nodes are set from 1 to 2n+1 (n representing the number of input variables). The output of this study is the forecasted water level at the M.7 gauge station at t+48, anextension of the previous study that forecasted water level at t+24 hours [25]. For the available dataset, the ANN model will learn from small flood events and try to forecast a big flood event. Therefore the data from the years 2007-2009 is used as learning data, and the data in the year 2011 is used for testing the model performance. The final results were obtained from the average of 50 loop calculations.

DOI: 10.5963/JWRHE0402002

To assess the model performances, Peak Difference PDIFF (Eq. 1), Root Mean Square Error-RMSE (Eq. 2) and Coefficient of Efficiency (Nash-Sutcliffe efficiency)-CE (Eq. 3) were applied [26].

$$PDIFF = max(Q'i) - max(Qi) \tag{1}$$

In Equation 1, $Q'i$ is the modelled value at time i, and Qi is the observed value at time i. If the result of PDIFF is a positive value, it indicates that the model forecasting is greater than the actual peak, while a negative value indicates that the model forecasting is less than the actual peak.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Q'i - Qi)^2}{n}} \tag{2}$$

A value of RMSE close to zero indicates good model performance.

$$CE = 1 - \frac{\sum_{i=1}^{n}(Q'i - Qi)^2}{\sum_{i=1}^{n}(Q'i - \bar{Q})^2} \tag{3}$$

In Equation 3, $\bar{Q}$ represents the average of observer data, and a CE value close to 1 indicates good model performance.

III. RESULTS AND DISCUSSION

All model performances will be decreased with an increasing number of hidden nodes, as demonstrated by CE, RMSE and PDIFF values (Fig. 4).

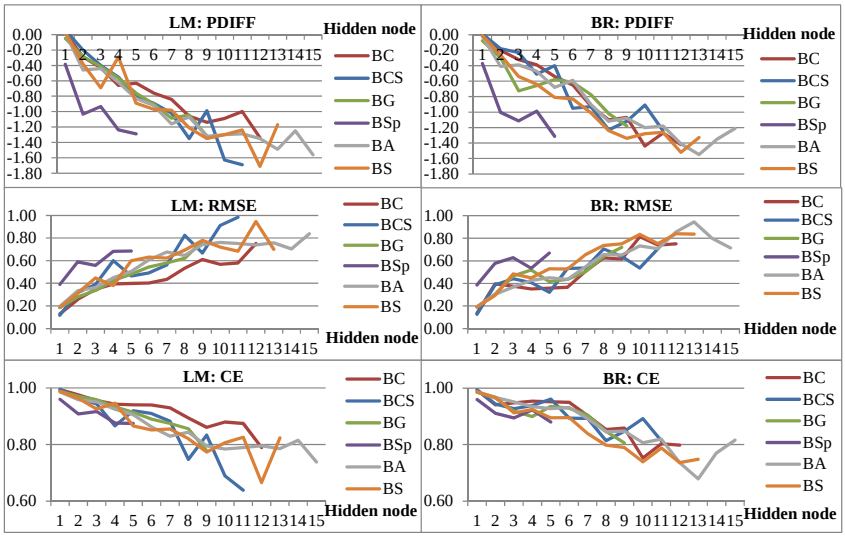

Fig. 4 Comparison of all models with PDIFF, RMSE and CE values

A hydrograph of all models, with LM and BR algorithms, used different input determination techniques. All hydrographs demonstrated similar results that forecasted water level at t+48 hour close to the actual water level at t+48hours, with the exception of the BSp model which used time t as the input variable (Fig. 5).

DOI: 10.5963/JWRHE0402002

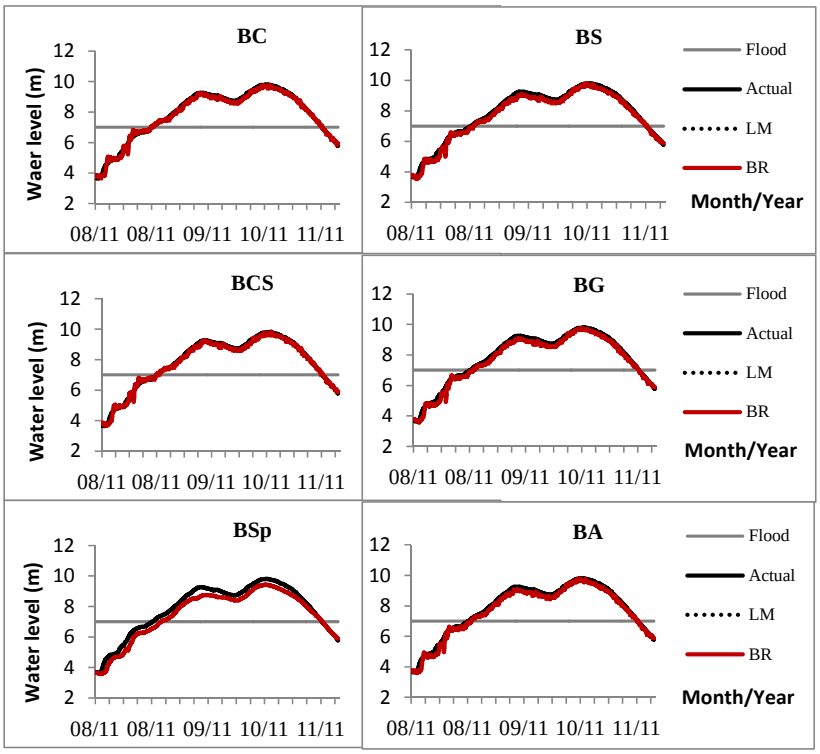

Fig. 5 Hydrographs of all models

The hydrographs of five models (BC, BS, BCS, BG and BA) provided similar results and statistic values (CE and RMSE) except for the BSp model, which provided the least accurate results (Table 2). It may be that using only time t as the input variable is not suitable for water level forecasting. Moreover, training algorithm BR tends to underestimate at the peak, demonstrated by the negative PDIFE values (Table 2). BS_BR and BCS_BR models demonstrated best performances to forecast the peak water level with only 0.01 m error (Table 2).

TABLE 2 MODEL PERFORMANCE EVALUATION

Models	PDIFF	RMSE	CE
BC_LM	0.02	0.13	0.99
BC_BR	-0.02	0.14	0.99
BS_LM	0.03	0.19	0.99
BS_BR	-0.01	0.19	0.99
BCS_LM	0.06	0.12	1.00
BCS_BR	0.01	0.13	0.99
BG_LM	-0.05	0.19	0.99
BG_BR	-0.08	0.19	0.99
BSp_LM	-0.37	0.39	0.94
BSp_BR	-0.38	0.40	0.94
BA_LM	0.02	0.20	0.99
BA_BR	-0.02	0.20	0.98

Models with different input determination techniques demonstrate different results, particularly regarding the flood situation. As can be seen in Fig. 6, models BC and BCS were the only two models to forecast the water rising to the flood level slightly before the actual time, a key component for flood warnings. On the contrary, models with other input determination techniques (BS, BG, BA and BSp) do not forecast with appropriate warning time.

DOI: 10.5963/JWRHE0402002

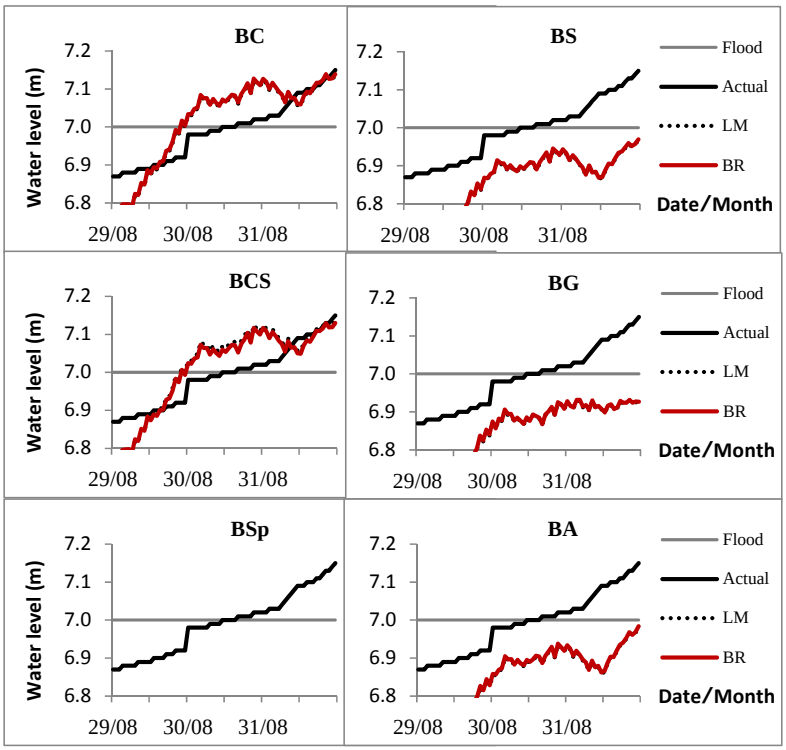

Fig. 6 Hydrograph of all models when flood begins

IV. CONCLUSIONS AND RECOMMENDATIONS

Learning algorithms LM and BR provide similar performances. The suitable number of hidden nodes is 1 for flood forecasting, particular this study area. The reason may be that all learning data sets represent small flood events, but the testing data set represents a big flood event. In addition, selecting input determination by cross correlation, and by cross correlation with stepwise regression, provide the best techniques for flood forecasting and flood warning at t+48 hour.

ACKNOWLEDGMENT

Thanks to the National Research Council of Thailand for the financial support, secondly to the Hydrology and Water Management Center for the Lower Northeastern Region, Department of Water Resources, Thailand, for the water level data. Thanks to the authors affiliation institutions for supports. Finally, thanks to Mrs. Phitchaya Chaipimonplin for editing the study area map.

REFERENCES

[1] S. Kure and T. Tebakari, "Hydrological impact of regional climate change in the Chao Phraya River Basin, Thailand," *Hydrological Research Letters*, vol. 6, pp. 53-58, 2012.

[2] T. Tingsanchali and M. R. Gautam, "Application of TANK, NAM, ARMA and neural network model to flood forecasting," *Hydrological Processes*, vol. 14, pp. 2473-2487, 2000.

[3] S. Y. Schreider, A. J. Jakeman, J. Gallant and W. S. Merrit, "Prediction of monthly discharge in ungauged catchments under agricultural land use in the Upper Ping basin, northern Thailand," *Mathematic and Computers in Simulations*, vol. 59, pp. 19-33, 2002.

[4] P. P. Mapiamand and N. Sriwongsitanon, "Estimation of the URBS model parameters for flood estimation of ungauged catchment in the upper Ping river basin, Thailand,"*Science Asia*, vol. 35, pp. 49-56, 2009.

[5] S. Supharatid, "Application of a neural network model in establishing a stage-discharge relationship for a tidal river," *Hydrological Processes*, vol. 17, pp. 3085-3099, 2003.

[6] G. J. Bowden, H. R. Maier and G. C. Dandy, "Input determination for neural network models in water resources application. Part 2. Case study: forecasting salinity in a river," *Journal of Hydrology*, vol. 301, pp. 93-107, 2005.

[7] T. Chaipimonplin, L. M. See and P. E. Kneale, "Improving neural network for flood forecasting using radar data on the Upper Ping River," in *Proceeding of 19th International Congress on Modelling and Simulation*, 2011, pp. 1070-1076.

[8] T. Chaipimonplin, L. M. See and P. E. Kneale, "Using radar data to extend the lead time of neural network forecasting on the River Ping," *Disaster Advances*, vol. 3, pp. 35-43, 2010.

[9] H. Yonaba, F. Anctil and V. Fortin, "Comparing sigmoid transfer function for neural network multistep ahead streamflow forecaasting," *Journal of Hydrologic Engineering*, vol. 15, pp. 275-283, 2010.

[10] M. H. Fun and M. T. Hagan, "Levenberg-marquardt training for modular networks," in *Proceeding of the 1996 International*

DOI: 10.5963/JWRHE0402002

Conference on Neural Network, 1996, pp. 468-473.

[11] A. P. Piotrowski and J. J. Napiorkowski, "Optimizing neural network for river flow forecasting-Evolutionary Computation methods versus the Levenberg-Marquardt approach," *Journal of Hydrology*, vol. 407, pp. 12-27, 2011.

[12] T. Chaipimonplin, "Investigation internal parameter of neural network model for flood forecasting at Upper River Ping, Chiang Mai," *KSCE Journal of Civil Engineering*, 2015, in press.

[13] A. Y. Shamseldin, "Application of a neural network techniques to rainfall-runoff modeling," *Journal of Hydrology*, vol. 199, pp. 272-294, 1997.

[14] K. P. Sudheer, "Knowledge extraction from trained neural network river flow models," *Journal of Hydrologic Engineering*, vol. 10, pp. 264-269, 2005.

[15] A. Kalra and S. Ahmad, "Using oceanic atmospheric oscillations for long lead time streamflow," *Water Resources Research*, vol. 45, pp. 1-18, 2009.

[16] S. Kim, J. H. Kim and K. B. Park, "Neural network models for the flood forecasting and disaster prevention systems in the small catchment," *Disaster Advances*, vol. 2, pp. 51-63, 2009.

[17] D. B. May and M. Sivakumar, "Prediction of urban stormwater quality using artificial neural network" *Environmental Modeling & Software*, vol. 24, pp. 296-302, 2009.

[18] A. Sharma, "Seasonal to inter annual rainfall probabilistic forecasts for improved water supply management: Part 1 A strategy for system predictor identification," *Journal of Hydrology*, vol. 239, pp. 232-239, 2000.

[19] T. M. K. G. Fernanda, H. R. Maier and G. C. Dandy, "Selection of input variables for data driven models: an average shifted histogram partial mutual information estimating approach," *Journal of Hydrology*, vol. 367, pp. 165-176, 2009.

[20] S. T. Chen and P. S. Yu, "Pruning of support vector networks on flood forecasting," *Journal of Hydrology*, vol. 347, pp. 67-78, 2007.

[21] G. S. Heo and I. S. Oh, "Simultaneous node pruning of input and hidden layers using genetic algorithms," in *Proceeding of 2008 International Conference on Machine Learning and Cybernetics*, 2008, pp. 3428-3433.

[22] T. Chaipimonplin, "An Exploration of neural network modelling options for the Upper River Ping," Ph.D. thesis, The University of Leeds, Leeds, UK, 2010.

[23] M. H. Beale, M. T. Hagan and H. B. Demuth, *Neural Network Toolbox™ 7 User's Guide*, Natick, MA: The Math Works, Inc, 2011.

[24] J. Holland, *Adaptation in Natural and Artificial Systems*, Ann Arbor: University of Michigan Pres, 1975.

[25] T. Chaipimonplin and T. Vangpaisal, "Comparison of the efficiency of input determination techniques with LM and BR algorithms in ANN for flood forecasting, Mun Basin, Thailand," *International Journal of Computer and Electrical Engineering (IJCEE)*, vol. 6, pp. 90-94, 2014.

[26] C. W. Dawson, HydroTest: Statistical Assessment of Hydrological Forecasts website. [Online]. Available: http://co-public.lboro.ac.uk/cocwd/HydroTest/index.html.

[27] Department of Water Resources, *Intergrated Water Resources Management Plan for the Mun Basin: Executive Summary*, Department of Water Resources, Ministry of Natural Resources and Environment, 2005. [Thai]

Tawee Chaipimonplin graduated in B.Sc (Soil Science) from Khon Kaen University, Khon Kaen in 1997, Thailand, M.Ap.Sc (Geospatial Information) from RMIT University, Melbourne, Australia in 2002, Certificate of Proceeding the Radar Data of Earth Surface Remote Sensing from NPO Mashinostroyenia, Moscow, Russian Federation in 2004 and Ph.D. (Physical Geography) from the University of Leeds, Leeds, UK in 2010.

He is a member of the Climate and Environmental Research Group (CERG) and a lecturer at Department of Geography, Chiang Mai University. His training experiences are "3 months course of GIS and Remote Sensing for Natural Hazard and Risk Assessment" at ITC, The Netherlands and "1 week of CASITA final workshop" at IIRS, India. His researches are using Artificial Neural Network (ANN), Geographic Information System (GIS) and Remote Sensing for natural hazard.

Thaveesak Vangpaisal received B.Eng (Civil Engineering) from Khon Kaen University, Khon Kaen in 1989, M. Eng.Sc (Water Engineering) from University of New South Wales, Sydney, Australia in1998 and Ph.D. (Civil Engineering) from Monash University, Melbourne, Australia in 2003.

He is an Assistant Professor at Department of Civil Engineering, Faculty of Engineering, Ubon Ratchathani University, Thailand. His research interest covers small watershed management, water quality, groundwater hydraulics and ground water contamination.

DOI: 10.5963/JWRHE0402002