Comparison of the Efficiency of Input Determination Techniques with LM and BR Algorithms in ANN for Flood Forecasting, Mun Basin, Thailand

Tawee Chaipimonplin and Thaveesak Vangpaisal

***Abstract*—The aims of this study are to improve the model perforemance with input selection and comparision fast learning and automated learning algorithms. Therefore, the methodology was, first, investigation cross correlation, stepwise regression, cross correlation with stepwise regression, genetic algorithm, supervise and all input) then comparing the learning algorithms: Levenberg Marquardt-LM and Baysian Regularization-BR). The ANN was used to forecast the water level at a gauge station M.7. The results showed the selecting input from genetic algorithm gives the best result for forecasting the flood peak and BR provided better results than LM particulaly at the flood peak.**

***Index Terms*—Neural network, flood forecasting, Levenberg-Marquardt, Bayesian Regularization, Mun Basin.**

I. Introduction

Flooding is a serious problem almost every year in some areas of Thailand. In 2011, the tropical storm NOCK-TEN passed Thailand, bringing with it torrential rains, in which over 30 provinces were flooded, including Chiang Mai and Ubon Ratchathani. For effective flood prevention, an early warning system is necessary. Hydrological model such as MIKE 11 [1], TANK [2] and Artificial Neural Network (ANN) [2], [3], have been developed and applied for flood forecasting. Notably, the ANN model, which is a black box model, makes use of a data driven method. The advantages of the ANN model are that it does not require physical data or field data, and has less computation time than the other approach models. It is also easy to update when new data become available. Therefore, the Hydrology Division, Royal Irrigation Department, Thailand, has chosen to apply the ANN model for flood forecasting for over 11 basins in Thailand [4]. For improving the ANN model for its better and more effective use in for flood forecasting, there are several methods, for example, selecting the input variable from the input determination techniques [5], [6], adding extra input variables [6], [7], and selecting different transfer functions [8], selecting different learning algorithms [9]. However, it has to be borne in mind that different basins may require different ANN models because of different runoff behaviors in the basin and different data available. In addition, it is obvious that the performance of the ANN model depends on learning algorithms, input variables and number of hidden nodes [10]. Learning algorithms in the Matlab software for example, Scaled Conjugate Gradient (SCG), Levenberg-Marquardt (LM) and Bayesian Regularization (BR) have different approaches, SCG is good for pattern recognition, LM is the fastest algorithm and BR is an automate regularization for improving generalization [11]. Chaipimonplin [10] reviewed many input determination techniques that have been used for selecting the input variable in the ANN model for example, cross correlation, stepwise regression, genetic algorithm, PMI, *etc*. He also found that the suitable techniques for forecasting water level in the Ping Basin, Thailand, included cross correlation, stepwise regression, cross correlation plus stepwise regression and genetic algorithm. Thus, this study was conducted to determine the efficiency of the two learning algorithms; LM and BR, for forecasting the water level at the specified case study, Mun Basin, Thailand. The inputs for the modeling were obtained by using four input determination techniques, as well as supervision selection, and by using all the input variables. The trial and error of the number of hidden nodes was also take into account.

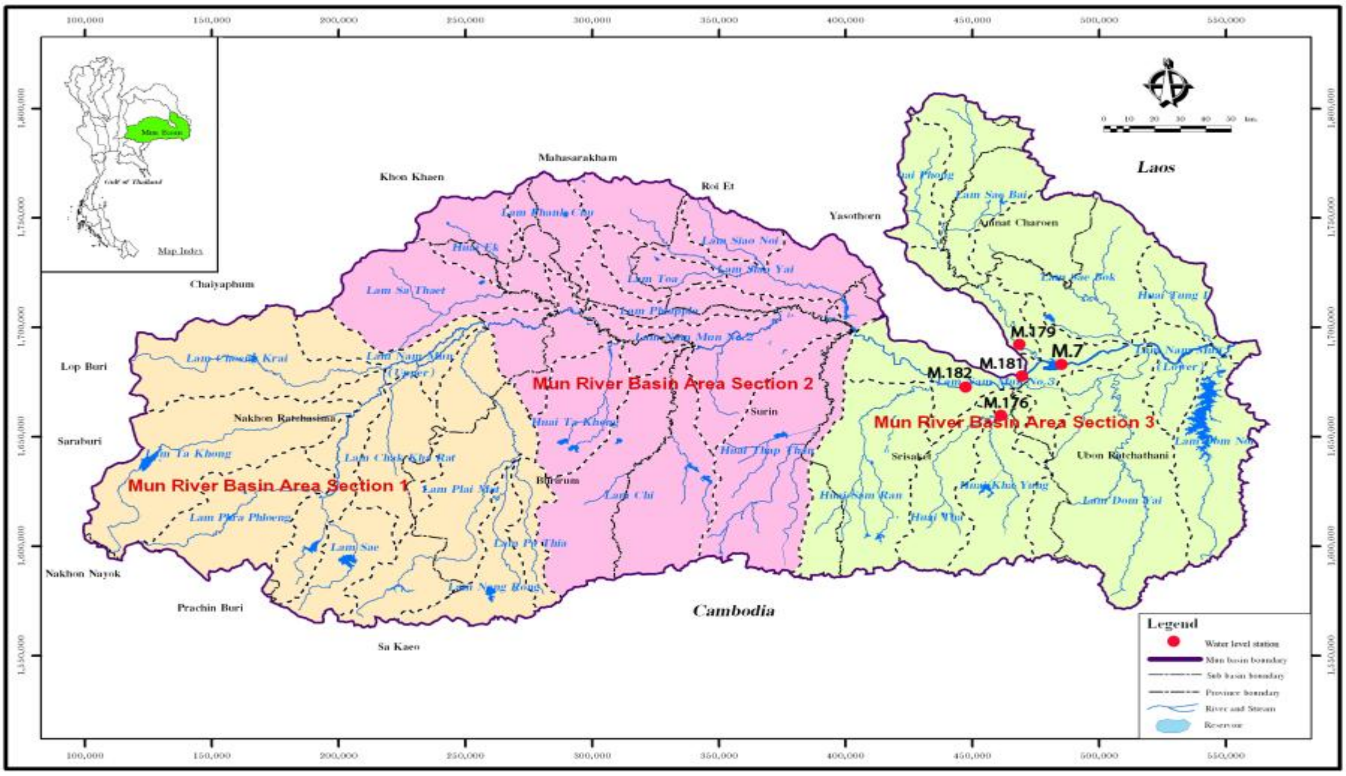
Manuscript received May 25, 2013; revised July 19, 2013. This work was supported in part by National Research Council of Thailand.

Tawee Chaipimonplin is with the Department of Geography, Chiang Mai University, Chiang Mai, Thailand (e-mail: tawee.c@cmu.ac.th, chaipimonplin@hotmail.com).

Thaveesak Vangpaisal is with the Civil Engineering Department, Ubon Ratchathani University, Ubon Ratchthani, Thailand (e-mail: thaveesak.v@bub.ac.th).

II. Study Area

Mun Basin is the largest river basin in Thailand, covering an area of approximately 71,000 square kilometers or 14% of the country's land area (Fig. 1a). Inundation problem, which occurs frequently in the Mun Basin area, often takes place in the area along the riverside of the Mun River in the lower part of the basin, especially in the Ubon Ratchathani province. The major causes of flood related problems in the Mun Basin include (1) inadequate water storage capacity, preventing water flow from retarding from the upper part of the basin, (2) limited drainage capability due to natural obstacles in the lower section of the Mun River, and (3) expansion of settlers into flood threatened areas. The Ubon Ratchathani province is 630 kilometers east of Bangkok. In the past 50 years, 23 flood events have been recorded for the area. The highest level of river runoff is recorded usually between September and October. The flood problem takes place in the community areas along the river banks in the city area of Ubon Ratchathani because these areas are located in the lower part of the interception point of the Mun River, adding to the woes is the high rainfall intensity of the area. It is a well-known and accepted fact that the stream water level of +112 meters above mean sea level, measured at the M.7 in Ubon Ratchathani, is the cause for the onset of the flood problems in the area. The highest water level of +117.7 meters above mean sea level (12.76 meters) was recorded in the year 1978, and this flooding caused extensive damage to the communities, environment, and economy. It was stated in the water-resources management plan for the Mun Basin [12] that trying to solve the flood problems in the Mun Basin by resorting to constructional measures was uneconomical, and that the promulgation of non-constructional measures such as flood forecast and warning systems, land utilization control, and evacuation of flood threatened areas are what must be implemented.



(a)

(b)

Fig. 1. Mun basin (edited from [12])

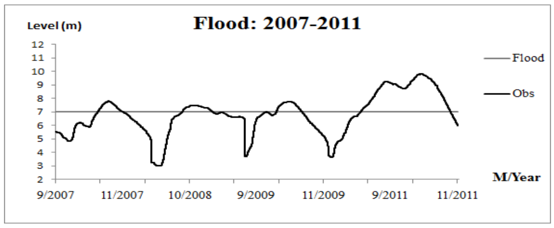


Fig. 2. Hydrographs at the M.7 station of floods between 2007- 2011

III. Methodology

Inputs for the model were the recorded water levels at the gauge stations in the study area. Four upper stations (M.181, M.179, M.176 and M.182) were used to forecast the water level at the M.7 gauge station (Fig. 1b). The M.181 and M.182 are the upstream gauge stations of M.7 in the main river, while the M.176 and M.179 are the gauge stations in the tributary of the main river. The longest distance of 76 kilometers is from M.176 to M.7, followed by 72 kilometers, which is from M.182 to M.7. The available data in the study were hourly water levels. The available data from the five stations covered the five year period from 2007 to 2011, in which four flood events had occurred including the most devastating flood in 2011 (Fig. 2). At the M.7, the water level of 7 meters denotes the onset of flood in the study area. In order to explore the efficiency of the learning algorithms, the experiment was divided into five for the different input variables. The model performances, between LM and BR, were compared.

A. *Input Determination Techniques*

For this study, four different approaches of the input determination technique (cross correlation-*AC*, stepwise regression-*AS*, cross correlation and stepwise regression-*ACS*, genetic algorithm-*AG*), supervise selection-*ASp* and all the input variables-*AA* were explored to indicate the difference in the input variable selection. Stepwise regression, which is the multiple regression method, removes the less correlation variable *AC*, and is the method used to calculate the relationship between the input variables, for this study only those input variables with correlation greater than 0.9 were selected. *AS* is the multiple regression method, by which the less correlation variables were removed and the input variables that remained after the removal were selected. *ACS* selected the input variables from those variables with correlation greater than 0.9, *AG* is based on biological evolution and natural selection and was developed by Holland [13] and *ASp* selected only the input variable time t of each station. The cross correlation and stepwise regression were calculated from the SPSS software and The WEKA software was used to calculate the genetic algorithm.

Three input variables were used for each station, which is the variable time *t*, time step back of 12 hours (*t*-12) and 24 hours (*t*-24). The total number of input variables was 15. Table I presents the selected input variables for each technique. *AC*, *AS* and *ASC*, which did not selected variables from the M.176 showed similar input variables, also the cross correlation and the stepwise regression provided the same input variables (12 inputs).

TABLE I: INPUT VARIABLES

Input	Input Determination Techniques					
	AC	***AS***	***ACS***	***AG***	***ASp***	***AA***
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
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		X		X
M.176					X	X
M.176_12						X
M.176_24				X		X
M.182	X	X	X	X	X	X
M.182_12	X	X	X	X		X
M.182_24	X	X	X			X
Total	12	12	10	11	5	15

B. *Artificial Neural Network Model*

Models were developed to compare the two learning algorithms; LM and BR. The number of hidden nodes for the model depended on the number of input nodes, which varied according to the five techniques of input determination. The hidden nodes were set from 1 to 2*n*+1[10] (*n* was the number of input variables), therefore, the numbers of the hidden node of the *AC*, *AS*, *ACS*, *AG*, *Asp* and AA models ranged from 1-25, 1-21, 1-23, 1-11 and 1-31, respectively. The result of this study was the water level at the M7 at the 24 hours in advance. For the available dataset, the data in the period 2007 – 2009 was used for the model learning, and the data in the year 2011 was used for the model testing. The final results were obtained from the average of the 50 loop calculations.

To assess the model performances, Peak Difference PDIFF (1), Root Mean Square Error-RMSE and Coefficient of Efficiency (Nash-Sutcliffe efficiency)-CE were applied [14].

$$PDIFF = max(Q'i) - max(Qi) \qquad (1)$$

where $Q'i$ is the modeled value at time *i*, and Qi is the observed value at time *i*. If the result of PDIFF is positive value, it means the model forecasting is over the actual peak, while, negative value means the model forecasting is under the actual peak.

IV. RESULTS AND DISCUSSION

The study results are presented separately corresponding to the input determination techniques as followed.

A. *Cross Correlation (AC) and Stepwise Regression (AS)*

Both the techniques; cross correlation and stepwise regression, selected all the input variables from all stations except the station M.176. It may be because of this station's distance to M.7, which is the longest distance, also it is not located at the main river. The hydrograph (Fig. 3a) shows that the forecast results of water levels agree well with the observed data. The LM and BR models show very similar performances. Nonetheless, the BR model (red line) seems to be better than the LM model as with increasing hidden node, it decreases the accuracy of peak forecasting (Fig. 3b), the CE values slightly decrease and the errors of RMSE rise up when the number of hidden nodes increases (Fig. 3c, 3d).

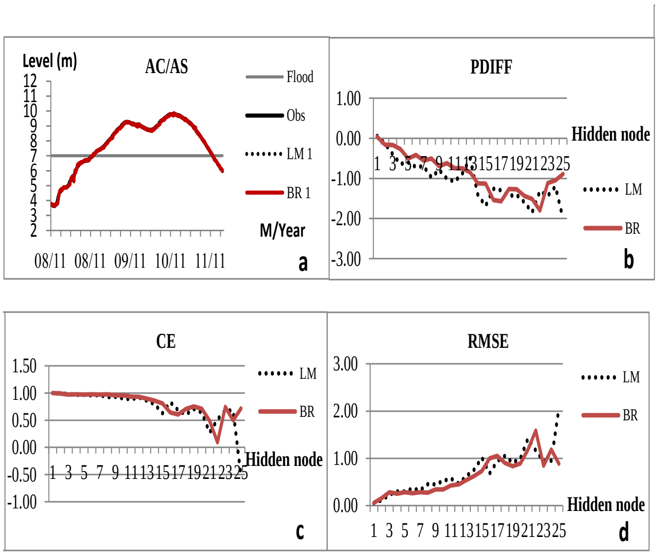


Fig. 3. Results of the models with the inputs from AC and AS

Therefore, the best hidden node is the hidden node which is indicated as 12:1:1. The performances of the LM and BR models are similar but that the BR model (in which the 0.037 meters error (PDIFF) and the CE value of 0.999 were obtained) gives a better peak water level forecasting than the LM model. All the same the RMSE values of the LM and BR models seem to be similar as there exists a difference of only 0.003.

B. *Cross Correlation and Stepwise Regression (ACS)*

The combination techniques (cross correlation and stepwise regression) required only two input variables less than the cross correlation. The RMSE between the two learning algorithms has a difference of 0.02 and the CE values of the LM and BR are 0.990 and 0.999 respectively. However, the BR model still provides a better result at the peak water level than the LM model (0.048/0.061). In addition, both the models hydrographs present good results (Fig. 4a). The effect of the number of hidden nodes are still similar with AC as BR is better than LM with more numbers of hidden node result in the poorer model performances (Fig. 4b, 4c, 4d).

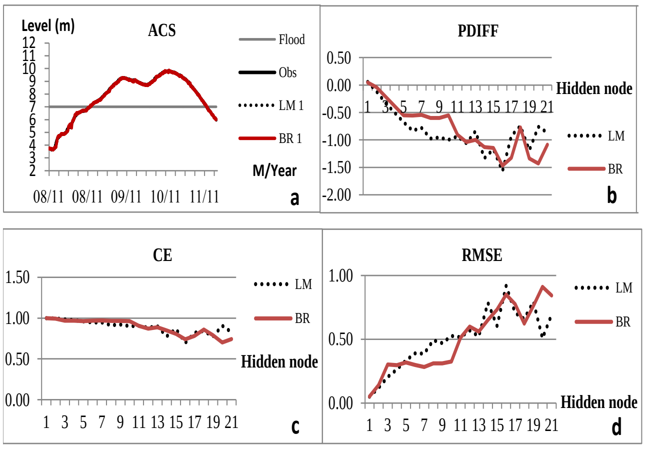


Fig. 4. Results of the models with the inputs from ACS

C. Genetic Algorithm (AG)

The model with the input selected from the genetic algorithm technique seems to provide the best performance for forecasting the water level at the peak. The modeling results with the BR and LM have errors of only 0.001 and 0.007 meters, which is the best score for each of the learning algorithms. On the other hand, the RMSE (0.073/0.075) and CE (0.998/0.998) values of AG_LM and AG_BR are the worst when compared with the other three input determination techniques. The reason might be, first, the AG chooses M.176_24, which has less correlation but may have a great effect on the peak water level of the M.7 gauge station, and second, it ignores the M.181_24 whereas the other three techniques do not ignore it (TableI) as the M.181 gauge station is located at the main river, and so the accuracy of forecasting the water level at M.7 may depend on it (Fig. 1b). Again, it can be seen that the performances of both the models are similar (Fig. 5a) and that the LM and BR models provide less accuracy at greater numbers of hidden nodes (Fig. 5b, 5c, 5d).

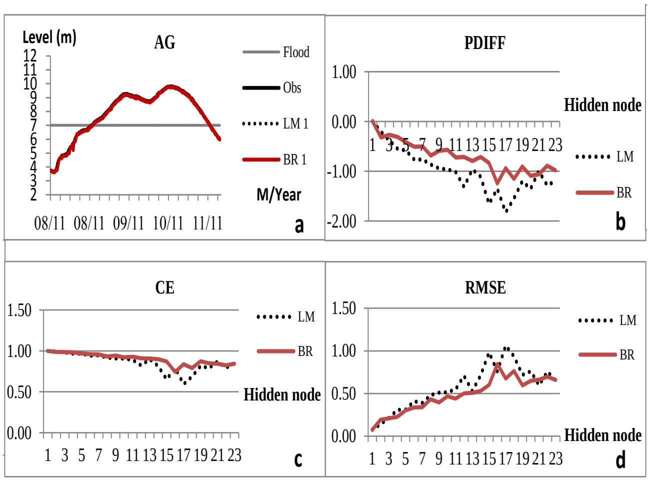


Fig. 5. Results of the models with the inputs from AG

D. Supervise Selection (ASp)

The idea of *ASp* is to reduce the number of input variables as much as possible, hence this technique tried to select only the data at time t of the five stations, Unfortunately, the forecasting results of both the LM and BR models were the poorest, in which the values of PDIFF, RMSE and CE were 0.172 meters lower than the peak, 0.192 and 0.987, respectively. This is because only five input variables were selected and that may not be enough information for ANN to learn and forecast. This indicated the influence of the number of inputs on the model performance, that is a fewer number of input variables leads to a reduction in the model's investigative capability. Fig. 6a presents the hydrographs of ASp both LM and BR algorithms underestimate with observation hydrograph. Performances of the models with the different numbers of hidden nodes drop dramatically at greater numbers of hidden nodes (Fig. 6b, 6c, 6d).

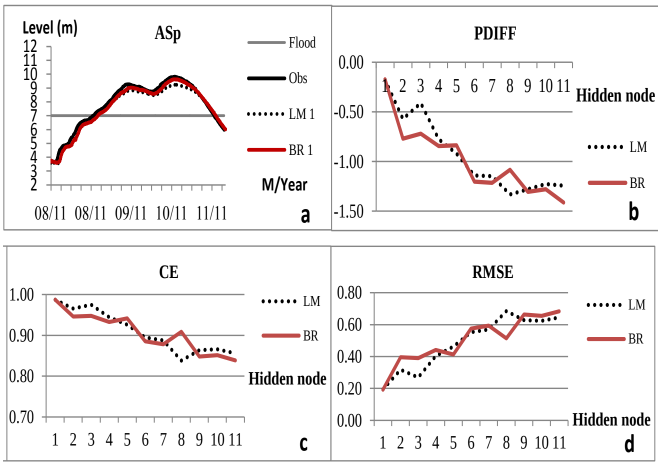


Fig. 6. Results of the models with the inputs from ASpAll Inputs (AA)

All the inputs consisted of 31 hidden nodes, which was the maximum number considered for this study. The hydrographs in Fig. 7a present the mode performance training with the LM and BR algorithms using all 15 input variables, it seems to be good results but when looks at the graphs at PDIFF, CE and RMSE (Fig. 7b, 7c, 7d), there are large errors occurring when the number of hidden nodes increases. Moreover, better results are obtained when training with the BR algorithm. The practice of using all the input variables gives better results at the peak than the other input determination techniques except the genetic algorithms as AA includes an input variable at M.176_24. However, too many input variables, of which some are unsuitable reduce the overall model performance as the CE of LM/BR and the RMSE of LM/BR are 0.998/0.998, 0.080/0.082 respectively, also because of the large size of the data set the model needs more time particularly for the learning process.

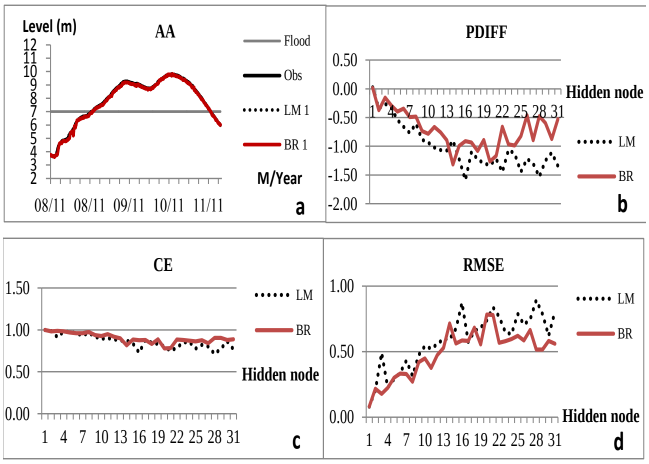


Fig. 7. Results of the models with the inputs from AA

V. Conclusion and Recommendation

To sum up, model training with the LM and BR shows similar performances in forecasting the water level at M.7 station but for the peak water levels, it is obvious that forecasting with the BR algorithm provided better results than with the LM. However, the major disadvantage of the BR learning algorithm is that it takes a long time to finish the learning process, particularly with larger numbers of hidden nodes. Additionally, the ANN models forecast water levels at the M.7 24 hours in advance from selected input variables from six techniques. The overall results are quite similar but *AG* seems to be the best technique for selecting the input variable for peak forecasting. In contrast, it is obvious that the insufficient input variable (*ASp*) could lead to the worst forecasting performance. These techniques also show that M.7, M.181 and M.169 are important input variables for flood forecasting at the M.7. As for finding the best number of hidden nodes for this study area, it can be pointed out that only one hidden node is the perfect number so it seems easy or not complex for forecasting the water levels 24 hours a head at M.7. Therefore, the recommendation for future study is to extend the forecast period to more than 24 hours or to use testing models with different input determination techniques for small flood events.

Acknowledgment

Thanks to the National Research Council of Thailand for the financial support, the Hydrology and Water Management Center for the Lower Northeastern Region, Thailand, for the water level data and 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, May 2012.

[2] T. Tingsanchali and M. R. Gautam, “Application of TANK, NAM, ARMA and neural network model to flood forecasting,” *Hydrological Processes*, vol. 14, no. 14, pp. 2473-2487, Nov 2000.

[3] S. Supharatid, “Application of a neural network model in establishing a stage-discharge relationship for a tidal river,” *Hydrological Processes*, vol. 17, no. 15, pp. 3085-3099, Aug 2003.

[4] Hydrology Division. (August 2012). Prediction water discharge 1-3 days in advances in major catchments by ANN, Royal Irrigation Department [Online]. Available: http://water.rid.go.th/hyd/. (Thai).

[5] 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, no. 1-4, pp. 93-107, Jan 2005.

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

[7] 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, no. 3, pp. 35-43, July 2010.

[8] H. Yonaba, F. Anctil, and V. Fortin, “Comparing sigmoid transfer function for neural network multistep ahead streamflow forecasting,” *Journal of Hydrologic Engineering*, vol. 15, no. 4, pp. 275-283, April 2010.

[9] 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, no. 1, pp. 12-27, Sep 2011.

[10] T. Chaipimonplin, “An exploration of neural network modelling options for the upper River Ping,” Ph.D. dissertation, Department of Geography, The University of Leeds, UK, 2010.

[11] M. H. Beale, M. T. Hagan, and H. B. Demuth, *Neural Network Toolbox™ 7 User’s Guide*, The Math Works, Inc., Natick, MA, 2011, ch 2, pp. 2:1-2:32.

[12] DWR, “Intergrated water resources management plan for the Mun Basin: Excutive summary,” Department of Water Resources, Ministry of Natural Resources and Environment, 2005. [Thai]

[13] J. Holland, *Adaptation in Natual and Artificial Systems*, University ofMichigan Pres, Ann Arbor, 1975.

[14] C. W. Dawson. (March 2013). HydroTest: statistical assessment of hydrological forecasts. [Online]. Available: http://co-public.lboro.ac.uk/cocwd/HydroTest/index.html.

Tawee Chaipimonplin graduated in B.Sc (soil science), KhonKaen University, KhonKaen, Thailand in 1997, M.Ap.Sc (Geospatial Information), RMIT University, Melbourne, Australia in 2002, Certificate of Processing the Radar Data of Earth Surface Remote Sensing, NPO Mashinostroyenia, Moscow, Russian Federation in 2004, and Ph.D. (physical geography), 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, Thailand. During working at Chiang Mai University, his training experience are “GIS and Remote Sensing for Natural Hazard and Risk Assessment” at ITC, The Netherlands, “CASITA final workshop” at IIRS, India and “Integration of Geo-Informatics for Environment and Disaster Management” at GISTDA, Thailand. His area of interest is using GIS, Remote Sensing and Artificial Neural Network for natural hazard; flood forecasting and landslide mapping

Thaveesak Vangpaisal received B. Eng (civil engineering) from KhonKaen University, KhonKaen, Thailand in 1989, M. Eng.Sc. (water engineering) from University of New South Wales, Sydney, Australia in 1998 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.