JOURNAL OF FLOOD ENGINEERING
Vol. 10 No. 2 (July-December, 2019)

Received: 14th July 2019 Revised: 17th August 2019 Accepted: 21st September 2019

ADVANCES IN DATA-DRIVEN FLOOD FORECASTING USING RADAR DATA

Pijush Samui[1], Linda See[2, 3], Tawee Chaipimonplin[3, 4] and Pauline Kneale[5]

[1]*Centre for Disaster Mitigation and Management, VIT University, Vellore - 632 014, India. E-mail: pijush.phd@gmail.com*

[2]*International Institute of Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria. E-mail: see@iiasa.ac.at and geolms@leeds.ac.uk*

[3]*School of Geography, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK E-mail: chaipimonplin@hotmail.com*

[4]*Department of Geography, Faculty of Social Sciences, Chiang Mai University, Chiang Mai, 50200, Thailand*

[5]*Teaching and Learning Directorate, 3 Endsleigh Place, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, UK. E-mail: pauline.kneale@plymouth.ac.uk*

ABSTRACT: Artificial Neural Networks (ANNs) and other data-driven methods are appearing with increasing frequency in the literature to forecast river levels or discharge at a point in the future. However, many of these data-driven models are developed for predicting only short lead times, e.g. 1 hour ahead, where unsurprisingly they perform very well. There have been much fewer documented attempts at predicting floods at longer, more useful lead times from a flood warning and civil protection perspective. In this paper ANN and Support Vector Machine (SVM) flood forecasting models are developed for the Upper Ping catchment at Chiang Mai in Thailand. Raw radar reflectively data are used as inputs to these models to see whether the lead time of the prediction can be increased beyond 12 hours. The models without radar data can produce reasonable forecasts up to a maximum of 15 hours ahead but the addition of radar data extends the lead time up to 36 hours ahead in predicting when the water will overflow the banks of the river and the flood peak. This study shows that the inclusion of spatially-distributed raw radar reflectivity data in data-driven models opens up a new, promising area for neuro-hydrological research.

Keywords: Neural networks, Support vector machines, Flood forecasting, Radar data

1. INTRODUCTION

Flooding is a major, recurring problem in many monsoon countries. Over the last 10 years, flooding has occurred in Thailand during most monsoon seasons, where the city of Chiang Mai in particular experienced a severe flood during the summer of 2005. To minimise the potential loss of life and the damage caused by flooding, early warning systems are needed that can provide timely and accurate forecasts. These systems require data for their development. Unfortunately, the historic flood record at Chiang Mai is limited in both record length and in terms of the number of gauging stations across the catchment so forecasting floods is a challenge. The Hydrology and Water Management Centre for the Upper Northern region (2005; 2007a, b) has responsibility for flood warning in the Upper Ping catchment. The technique currently in

use is based on a correlation between the water stage at an upstream station (P67) and the downstream station (P1) at Chiang Mai. The maximum time for flood warning using this method is currently 6-7 hours. In addition to this approach, the Natural Disasters Research Unit in the Civil Engineering Department of Chiang Mai University (CENDRU) uses a support vector machine coupled with a hydrodynamic model to predict stage 7 hours ahead at P1 station (Natural Disasters Research Unit, 2007b). The results are very good for this lead time. However, with severe floods like that experienced in 2005, longer lead times are needed. There are a number of conceptual and physical hydrological models in existence for the Ping Catchment but most of them predict either monthly or daily discharge (Schreider *et al*., 2002; Vongtanaboon *et al*., 2008; Taesombat and Sriwongsitanon, 2006, 2010; Mapiam and Sriwongsitanon, 2009). Hourly forecasts are needed if an effective operational flood forecasting system is to be implemented and developed.

Data-driven methods (e.g. Artificial Neural Networks (ANNs), fuzzy logic, support vector machines, etc.) offer an alternative approach that does not require knowledge of the physical relationships in the catchment, but instead learns these relationships from the data. Two recent reviews of the literature reveal a plethora of papers which demonstrate the successful application of ANNs for rainfall-runoff modelling and other hydrological applications (Abrahart *et al*., 2010; Maier *et al*., 2010). These reviews also cover examples of hybrid, data fusion or soft computing applications, where different technologies are used together to produce a better forecast than an individual model (see e.g. See, 2008; Solomatine *et al*., 2008). There are a few examples of the use of ANNs in Thai catchments but most have been developed for daily rather than hourly forecasting such as the models developed by Thaisawasdi *et al*., (2007). Another example is the study by Tingsanchali and Gautam (2000), which involved the development of an ANN to predict floods 1 day ahead using the average daily rainfall of 10 stations, evaporation based on 4 meteorological stations and runoff data as the inputs. However, the model underestimated the peak of the flood. This was attributed to the rainfall data, which the authors argued was not representative of the actual values across the catchment. Sukka (2005) also used an ANN trained with backpropagation to predict daily water inflows to a reservoir one and two days ahead using daily precipitation and discharge. However, the results were poor and there was again a large underestimation of the peak. A finer temporal resolution was employed by Patsinghasanee *et al*., (2004), who developed ANNs for a 12 hour lead time but the same problem with an underestimation of the flood peaks occurred. Longer lead times of up to 72 hours were predicted using ANNs developed by Ninprom and Chumchean (2009). However, they did not show any graphical results or provide any information about the performance of the model in terms of peak prediction. The most relevant piece of research is the study by Chidtong *et al*., (2009), who built a series of hybrid forecasting models (i.e. neuro-fuzzy models optimised with a genetic algorithm) to predict the flood at Chiang Mai in 2005 using hourly river level data. The models were also applied to a large flood in Koriyama in Japan. However, for Chiang Mai, the authors used only daily rainfall as an input as hourly was not available. The results showed that the hybrid model outperformed the other models to which it was compared (i.e. neuro-genetic and an ANFIS model), and that the hybrid system could produce a good forecast with a lead time of 12 hours.

The review above has shown that there are clearly some applications of both physical/ conceptual models and ANNs in the development of hydrological models for the Upper Ping catchment. However, these models either forecast daily data, or they predict at lead times of 12 hours or less. Here we ask the question of whether longer lead times than those previously reported in the literature can be predicted for this catchment, in particular for large storms the size of that encountered in 2005. The main addition is raw radar reflectivity data, which has not been used before as a direct input to an ANN. In addition, this paper also uses support vector machines (SVMs) to predict the river level at Chiang Mai for lead times of 18 to 36 hours ahead. The first set of experiments uses river level data from stations at Chiang Mai and upstream to predict at a lead time of 18 hours, extending upon that reported in Chidtong *et al*., (2009), to determine whether useful results can be obtained using this data source alone. The second set of experiments then uses raw radar reflectivity data to see whether the lead time of the forecast can be usefully extended.

2. METHODS

2.1 Artificial Neural Networks

ANNs have been used in hydrology for more than two decades (Abrahart *et al*., 2009) as well as in a range of other domains such as financial forecasting and pattern recognition (Hu and Hwa, 2002; Kamruzzaman *et al*., 2006). Their roots can be traced back to the work of McCulloch and Pitts in the early forties and research into Artificial Intelligence (Russell and Norvig, 1995), where the architecture is loosely based on the human brain. ANNs are comprised of individual processors called neurons or nodes which contain a simple transformation function. The nodes are interconnected in layers where the most common arrangement is a single hidden layer. Each node in the input layer corresponds to a single input variable. These input nodes are then fully connected via weights to each hidden layer node, where the determination of the optimal number is often undertaken via trial and error although various heuristics exist (e.g. Minns and Hall, 1996; Walczak and Cerpa, 1999). The hidden layer nodes are then interconnected by weights to one or more output nodes. The ANN essentially performs a mapping of an input vector, x, to an output vector, y, as the input data are fed forward through the network in a forward pass. Once a network is fully trained, the ANN is a deterministic model. Training involves giving the network a data set with inputs and known outputs and iteratively adjusts the weights until the network is capable of predicting the desired output. The network can be trained to learn the relationship in the data using many different algorithms, where backpropagation of error is the most common (Rumelhart *et al*., 1986). A testing data set is then used to determine how determine how well the NN predicts or forecasts using data it has not seen before. Bishop (2005) provides a good reference source for further detailed information.

2.2 Support Vector Machines (SVMs)

SVMs allow one to carry out regression without the need for knowing the form of the regression equation beforehand. As with ANNs, the main aim of a SVM is to find a function $f(x)$ that

approximates y_i with $f(x_i)$. This study uses the SVM as a regression technique by introducing an ε-insensitive loss function, $L_\varepsilon(y)$, where

$$L_\varepsilon(y) = 0 \quad \text{for} \quad |f(x) - y| < \varepsilon \quad \text{otherwise} \quad L_\varepsilon(y) = |f(x) - y| - \varepsilon \tag{1}$$

This defines a tube, ε, (Figure 1) so that if the predicted value is within the tube, the loss is zero, while if the predicted point is outside the tube, the loss is the magnitude of the difference between the predicted value and the radius, ε, of the tube.

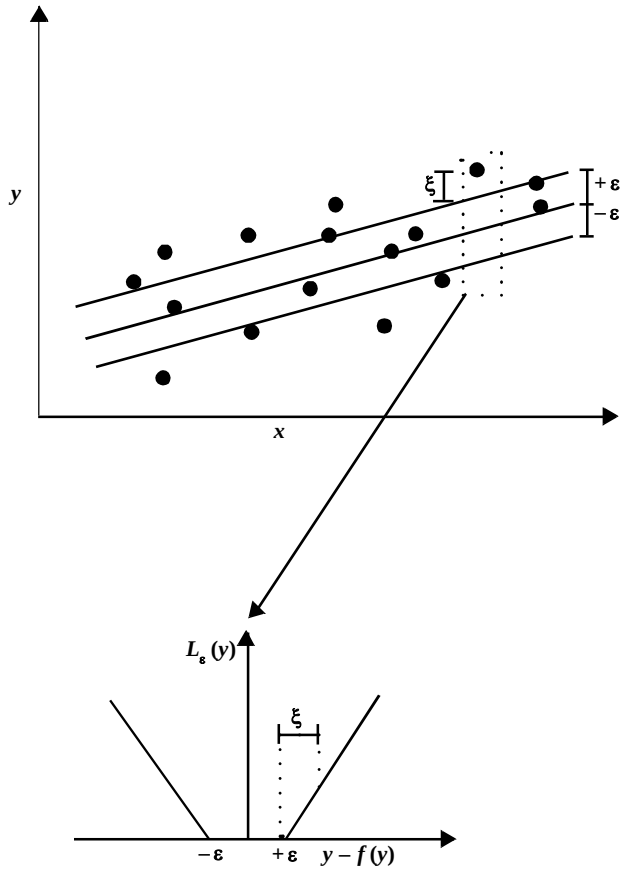

Figure 1: Prespecified Accuracy ε and Slack Variable ξ in Support Vector Regression. Taken from Scholkopf (1997)

In the simplest case, the solution is linear:

$$f(x) = \langle w, x \rangle + b \; w \in R^n, \quad b \in R \tag{2}$$

where w is a weight vector, $\langle w, x \rangle$ represents a dot product, and b is a scalar. Geometrically, $f(x)$ represents a hyperplane, w is its normal vector and b is an intercept. One seeks the smallest possible w among all potential solutions by minimizing the Euclidean norm $\|w\|^2$. This condition

is often referred to as a requirement for the flatness in *f*. The solution for $f(x)$ is obtained through solving a convex optimization problem for *n* data points as follows:

$$\begin{aligned} &\text{Minimize:} && \frac{1}{2}\|w\|^2 \\ &\text{Subject to:} && y_i - (\langle w, x_i\rangle + b) \le \varepsilon, \quad i = 1, 2, ..., n \\ & && (\langle w, x_i\rangle + b) - y_i \le \varepsilon, \quad i = 1, 2, ..., n \end{aligned} \tag{3}$$

In order to allow for regression errors, slack variables ξ_i and ξ_i^* (Figure 1) are introduced in Eq. (4). Only if a data point *i* lies outside a distance of ε from the positive side of $f(x_i)$, then the slack variable ξ_i is nonzero. ξ_i^* is defined in a similar manner for the negative side. To allow for these errors, additional constraints are introduced and the formulation can then be restated as:

$$\begin{aligned} &\text{Minimize:} && \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\text{Subject to:} && y_i - (\langle w.x_i\rangle + b) \le \varepsilon + \xi_i, \quad i = 1, 2, ..., n \\ & && (\langle w.x_i\rangle + b) - y_i \le \varepsilon + \xi_i^*, \quad i = 1, 2, ..., n \\ & && \xi_i \ge 0 \quad \text{and} \quad \xi_i^* \ge 0, \qquad i = 1, 2, ..., n \end{aligned} \tag{4}$$

The constant $0 < C < \infty$ determines the trade-off between the flatness of *f* and the level of deviation greater than ε that is to be tolerated (Smola and Scholkopf, 2004). In practice, the *C* value is selected by trial and error. The above constrained optimization problem of Eq. (4) is often solved by the method of Lagrange multipliers. A Lagrangian function is constructed in the following way:

$$\begin{aligned} L(w, \xi, \xi^*, \alpha, \alpha^*, \gamma, \gamma^*) = \frac{\|w\|^2}{2} + C\left(\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) - \sum_{i=1}^{n}\alpha_i[\varepsilon + \xi_i - y_i + \langle w.x_i\rangle + b] \\ - \sum_{i=1}^{n}\alpha_i^*[\varepsilon + \xi_i^* + y_i - \langle w.x_i\rangle - b] - \sum_{i=1}^{n}(\gamma_i\xi_i + \gamma_i^*\xi_i^*) \end{aligned} \tag{5}$$

where α, α^*, γ and γ^* are the Lagrangian multipliers. The solution to the constrained optimization problem is determined by the saddle point of the Lagrangian function $L(w, \xi, \xi^*, \alpha, \alpha^*, \gamma, \gamma^*)$, which has to be minimized with respect to *w*, *b*, ξ and ξ^* given by:

$$\text{Condition 1:} \qquad \frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} x_i(\alpha_i + \alpha_i^*)$$

$$\text{Condition 2:} \qquad \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i = \sum_{i=1}^{n}\alpha_i^*$$

Condition 3: $$\frac{\partial L}{\partial \xi} = 0 \Rightarrow \sum_{i=1}^{n} \gamma_i = \sum_{i=1}^{n} (C - \alpha_i)$$

Condition 4: $$\frac{\partial L}{\partial \xi^*} = 0 \Rightarrow \sum_{i=1}^{n} \gamma_i^* = \sum_{i=1}^{n} (C - \alpha_i^*) \quad (6)$$

Substituting (6) into (5) yields the dual optimization problem:

Maximize: $$-\varepsilon \sum_{i=1}^{n} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{n} y_i (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i . x_j)$$

Subject to: $$\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \alpha_i^*, \ 0 \le \alpha_i^* \le C \ \text{ and } \ 0 \le \alpha_i \le C \quad (7)$$

The solution of the above optimization problem gives the coefficients α_i^* are α_i. From the Karush-Kuhn-Tucker (KKT) optimality condition, it is known that some of α_i, α_i^* will be zero. The non-zero α_i, α_i^* are called support vectors. After substituting the equation for *w* from Eq. (5), the $f(x)$ of Eq. (2) can be written as

$$f(x) = \sum_{\text{support vectors}} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b \quad (8)$$

where $b = -\left(\frac{1}{2}\right) \sum_{i=1}^{n} x_i (\alpha_i - \alpha_i^*)[x_r + x_s]$ and x_r and x_s are any two support vectors.

From Eq. (5) it is clear that *w* has been completely described as a linear combination of the training data. Thus, the complexity of the function representation by support vectors is independent of the dimensionality of the input space, and it depends only on the number of support vectors. In this study, linear regression is not adequate. For non-linear regression, the SVM maps the input data into a higher dimensional feature space through some nonlinear mapping (Boser *et al*., 1992) where the relationship then becomes linear. Instead of *x*, the mapped $\Phi(x)$ is now used in the optimization formulation of Eq. (7). This gives a revised formulation as follows:

Maximize: $$-\varepsilon \sum_{i=1}^{n} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{n} y_i (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(\Phi(x_i).\Phi(x_j))$$

Subject to: $$\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \alpha_i^*, \ 0 \le \alpha_i^* \le C \ \text{ and } \ 0 \le \alpha_i \le C \quad (9)$$

The computation involving $\Phi(x_i) . \Phi(x_j)$ is costly so is replaced with a well-behaved kernel function $K(x_i, x_j)$ (Cristianini and Shwae-Taylor, 2000; Cortes and Vapnik, 1995). Mapping to a higher dimensional space is not unique, and different kernel functions may yield different results. The selection of a proper kernel functions lies at the core of SVM application. After the mapping, $f(x)$ becomes:

$$f(x) = \sum_{\text{support vectors}} (\alpha_i - \alpha_i^*) K(x_i . x) + b \tag{10}$$

where $b = -\left(\frac{1}{2}\right) \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) [K(x_i, x_r) + K(x_i, x_s)]$, and x_r and x_s are again any two support vectors. Some commonly used kernels are polynomials of various degrees, radial basis functions, Gaussian functions, splines and sigmoid functions. Figure 2 illustrates the typical architecture of a SVM.

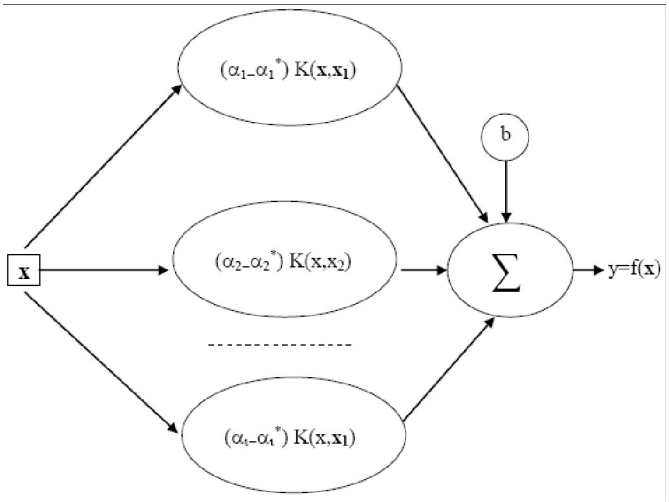

Figure 2: Architecture of a Support Vector Machine

3. STUDY AREA AND DATA

The Ping catchment is located in the Northern part of Thailand and covers 5 provinces: Chiang Mai, Lamphun, Kamphaengphet, Tak and Nakhonsawan. The annual average rainfall varies from 900 to 1,900 mm with an average of 1,125 mm (Rodratana and Piamsa-nga, 2008). The Ping River is the main river in this catchment with a length of 740 km (Mapiam and Sriwongsitanon, 2009). The entire Ping catchment covers approximately 33,898 km^2 and is mainly covered by forest (46.5%), agriculture (31.2%) and paddy fields (12.6%). The Ping catchment is divided into two parts: the Upper and the Lower Ping. The Upper Ping is a large complex river basin covering two provinces (17° 14′ 30″ – 19° 47′ 52″ N, 98° 4′ 30″ – 99° 22′ 30″ E): Chiang Mai and Lam Phun (Mapiam and Sriwongsitanon, 2009). It has an area of approximately 23,600 km^2 with 15 sub-catchments (Figure 3). The distance from the source of the river to Chiang Mai city is 190 km (Hydrology and Water Management Centre for Upper Northern Region, 2007b).

Monsoon conditions in Thailand come from northeastern weather systems (November to February), which bring moisture from the South China Sea as well as the southwest monsoon (May to September), which brings rain from the direction of the Indian Ocean (Boochabun

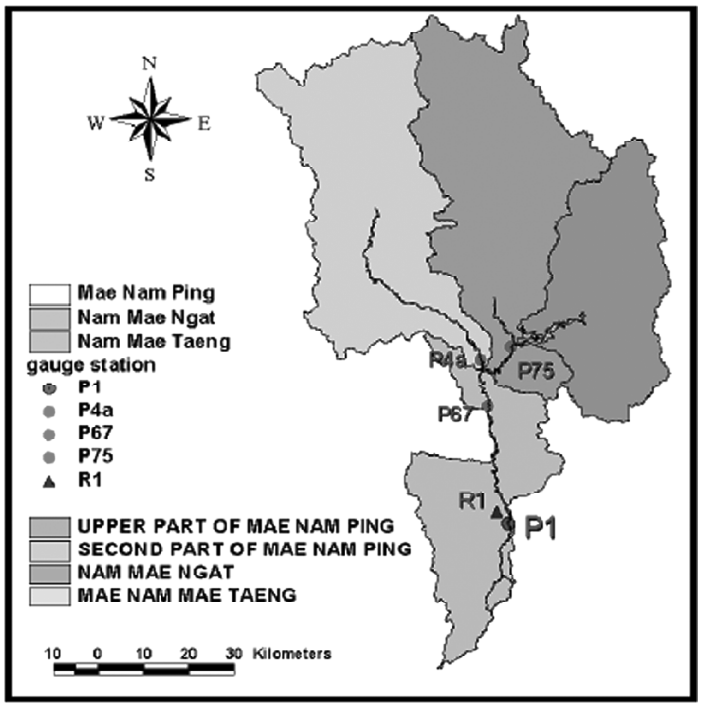

Figure 3: Location of the River Ping and Chiang Mai

et al., 2004). In this catchment the average regional temperature is approximately 25.4°C, the maximum is 41.4°C in May and the minimum is 3.7°C in January (Natural Disasters Research Unit, 2007c). The wettest month is August, which has an average rainfall of approximately 224.4 mm, whilst the driest month is January with 7.7 mm. The elevation of this basin ranges from 380m to 2,275 m above sea level.

The flooding in Chiang Mai at P1 station is recorded by the Hydrology and Water Management Centre for the Upper Northern Region. In addition to the main Ping River channel, there are seven minor rivers in the Chiang Mai area, two of which feed into the Ping above station P1, where the average annual discharge is 59.38 m^3/s. Flooding in the city occurs when water discharge is greater than 460 m^3/s and water stage level in the main channel exceeds 3.70 m above the local datum (304.2 msl) (Hydrology and Water Management Centre for Upper Northern Region, 2007a). The main causes of flooding in this catchment are considered to be meteorological, fed by monsoonal rainfall. According to the flood records for the past 50 years in the Chiang Mai city area, the four highest monsoon flood events occurred in 1987, 1994, 1995 and 2005 with water levels of 4.53, 4.43, 4.27 and 4.93 m, respectively. Sophhonphattanakul *et al.*, (2009) investigated the effect of changes in land use on stream flow in this catchment and found that changing land use through urbanization, industrialization and deforestation have also contributed to flooding in the Upper Ping catchment. Chiang Mai's land use has changed rapidly in response to the National Economic and Social Development

Plan volume 5 (1982-1986) for developing Chiang Mai into a 'Primate city' (Chatchawan, 2005). As a result of development, there has been deforestation in the catchment, and the building of infrastructure in the city and along the Ping River has increased. Engineering work on the Ping channel as part of flood control works has also changed the flooding level. In the past the flooding level at P1 was 3.40 m. After excavation of the Ping River channel in 2004, the flooding level increased to 3.70 m (Chatchawan, 2005). Flood events more recently have been higher when compared with previous decades although no flood events occurred in 2007.

The 2005 flood event was triggered by heavy and prolonged rainfall. The biggest flood event of the year took place between 13-16 Aug 2005. There was heavy rain on 12 August, with an average of 128 mm within a 24 hour period in the north Chiang Mai region. It started to flood on 13 August with the river rising at a rate of 12-14 cm/hr. The maximum water level was 4.9 m on 14 August and this elevated level remained for 8 hours. As a result, the water covered a very wide area of the city for up to 51 hours. Smaller flood events then followed in the month of September.

The input data used in the modelling includes three water level gauging stations (P1, P75 and P67) and radar images, all of which were available at an hourly time scale. The locations are shown in Figure 4. In the first set of modelling experiments, only water level is used. The second half of the paper then focuses on the use of input data from radar images.

4. MODEL DEVELOPMENT

4.1 ANN/SVM Models for a Lead Time of 18 hours

Input variables from river gauging stations at Chiang Mai (P1) and upstream (P67 and P75) were used to develop the NN and SVM models. The training dataset contained storm events from 2001 to 2004, in particular: 1/08 – 31/10 for 2001 and 2002; 1/09 – 31/10 for 2003; and 1/05 – 31/10 for 2004. Data for 2005 (1/08 – 1/09) were used to test the performance of the models. A range of potential input variables was derived from the three stations including levels at time t, t-3, t-6 continuing at 3 hour intervals to t-24 and moving averages over the previous 6, 12 and 24 hours. Stepwise linear regression was then employed to reduce the number of input variables from 36 to 12. This method was shown to be an effective input determination method in previous NN modelling experiments in Chiang Mai when compared with several other approaches (Chaipimonplin, 2010). There is one rainfall station near P1. However, experimentation revealed that this input had little effect on the ability of the models to make more accurate forecasts (Chaipimonplin, 2010).

Once the inputs were selected, feedforward networks with 10 hidden nodes were trained with backpropagation and Bayesian Regularisation (MacKay, 1992). The advantage of using this algorithm is that a validation data set is not required for stopping the training process. Therefore, more data can be used in the training data set, which is particularly relevant to this case study as the amount of data available for model development is relatively small. Simple trial and error revealed little difference when a larger number of hidden nodes was used. For each experiment, 50 models were developed and the average was used as the model prediction

based on suggestions by Anctil (2007). For the SVM, C was set to 100, a radial basis function was used as the kernel (with $\sigma = 0.1$) and ε was set to 0.01.

The Root Mean Squared Error (RMSE), the difference in the peak prediction (PDIFF) and the Coefficient of Efficiency (CE) were calculated for the model predictions using Hydrotest (Dawson *et al.*, 2007), to match those chosen by Chidtong *et al.*, (2009) so that a direct comparison could be made. These measures are calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Q_i - \hat{Q}_i)^2}{n}} \tag{11}$$

$$PDIFF = \max(Q_i) - \max(\hat{Q}_i) \quad [\text{for } i = 1 \text{ to } n] \tag{12}$$

$$CE = 1 - \frac{\sum_{i=1}^{n} (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n} (Q_i - \bar{Q}_i)^2} \tag{13}$$

where Q_i is the actual value, $\hat{Q}_i$ is the model prediction (where $i = 1$ to n data points), $\bar{Q}$ is the mean of the observed data, and $\tilde{Q}$ is the mean of the modelled values. The *RMSE* and *PDIFF* are expressed in meters while the *CE* is a dimensionless coefficient. In addition, hydrographs showing the model predictions were examined.

4.2 Addition of Radar Data

Radar data covering the Chiang Mai area were obtained from the bureau of Royal Rainmaking and Agriculture Aviation, which operates five radar stations across Thailand. The CAPPI (Constant Altitude Plan Position Indicator) method is used to detect precipitation from the images, which is mainly from the southwest and northeast monsoon in Thailand from convective activity. The spatial resolution of each image is 1 km, with a ground coverage radius of 240 km. Radar data are often used to estimate rainfall. However, the radar data require calibration and there is only one rainfall station available, which is located near Chiang Mai. For this reason, an alternative approach was employed in which raw radar reflectivity values across the image were used as inputs to the NN and SVM models.

Radar images were available at 1 hr intervals. A 30 × 50 km square north of Chiang Mai was chosen from the radar image to provide sufficient coverage of the river on both sides as shown in Figure 4. The image was sampled at 12 points covering the river with a distance of 10 km between points. The points are labeled as *Z*11, *Z*12, etc. to reflect the row and column. The 3 × 3 pixels directly surrounding each of the 12 points were also extracted in order to create an average at that point. Each row of points was then further averaged, i.e. Z1 is the

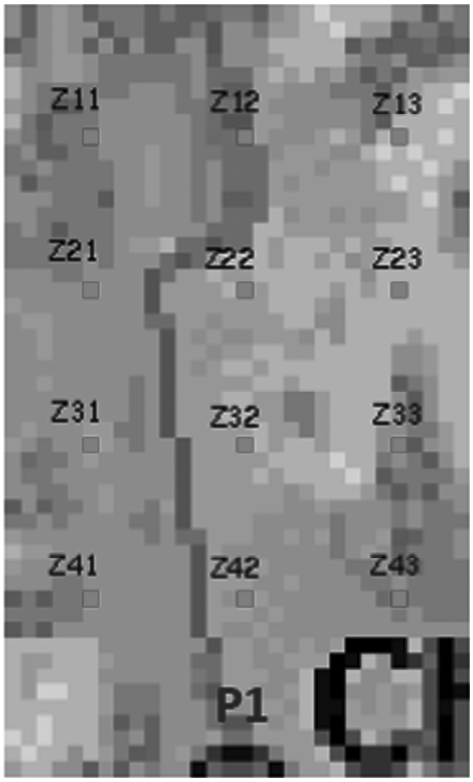

Figure 4: A 30 × 50 km Section of the Radar Image with 12 Sample Points

average of the points *Z*11, *Z*12 and *Z*13. The P1 gauging station is located at the bottom of the image so the travel time from row *Z*1 to P1 is the longest. Values were extracted from the radar images going 4 days backwards in time. The inputs to the *ANN* and *SVM* were then chosen by selecting radar values at those times when the correlation was the highest with the water level at P1 in order to develop models for predicting the water level at lead times of 24, 30 and 36 hours ahead. Previous values of water level at P1 were also used as inputs. The purpose of this set of experiments was simply to determine the feasibility of using raw radar data for extending the lead time of the forecast. The same *NN* and *SVM* settings were used as in the previous experiment.

5. RESULTS

5.1 Model Results for a Lead Time of 18 hours

Table 1 shows the goodness-of-fit statistics for a lead time of 18 hours in order to compare them with the results of Chidtong *et al.*, (2009), which was for a lead time of 12 hours. It is not surprising that the results are worse as the attempt to extend the lead time is a harder problem to forecast using only water level data. The *SVM* outperforms the *NN* in terms of the *RMSE* and the *CE* but the difference between the peak and the maximum value prediction is better for the *NN*.

Table 1
Performance Statistics for the *SVM* and *NN* for a Lead Time of 18 hours Compared to the Performance Reported in Chidtong *et al.*, (2009) for a Lead Time of 12 hours

Model	*RMSE (m)*	*PDIFF (m)*	*CE*
NGO (Chidtong *et al.*, 2009)	0.154	-0.316	0.963
ANFIS (Chidtong *et al.*, 2009)	0.109	-0.246	0.982
Hybrid (Chidtong *et al.*, 2009)	0.100	-0.165	0.982
SVM	0.199	-0.419	0.936
NN	0.213	-0.024	0.927

Figure 5 shows the results of the *NN* and the *SVM* for the highest event during the summer of 2005 in the month of August. These results in the form of hydrographs are actually much more informative than the global performance statistics. Both the *SVM* and the *NN* are quite late in predicting the start of the flood event. They are also approximately 3 hours late in predicting a level of 3.7 m or the level at which the water overtops the banks of the river. The *SVM* then overpredicts the peak while the *ANN* hits the peak (which explains the *PDIFF* statistic) although it generally underpredicts the upper part of the hydrograph so the *PDIFF* statistic is somewhat misleading without looking at the performance on the hydrograph. The falling limb, on the other hand, is predicted well by both models. The *SVM* generally exhibits a smoother behaviour while the *ANN* shows more erratic predictions, especially as evidenced on the lower levels after the storm. From an operational perspective, both models were late in predicting the overtopping of the bank but the *SVM*, with its overprediction of the peak, is probably a better model. Without rainfall data available to drive the rising limb of the hydrograph, the maximum lead time that can, therefore, be achieved using water level data from upstream gauging stations is approximately 15 hours.

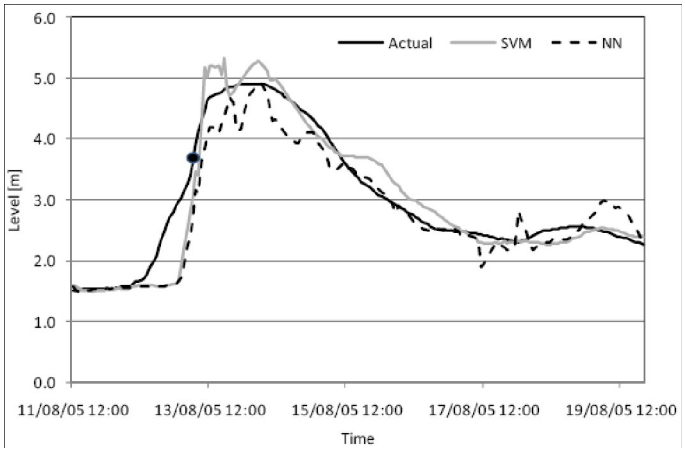

Figure 5: *SVM* and *NN* Forecasts Compared to the Observed Values for a Lead Time of 18 hours Using Water Level Data as Inputs. The Dot Denotes the Level at which the Water Overtops the Bank (3.7 m)

5.2 Model Results for Longer Lead Times using Radar Data

Figures 6 to 8 show the results of the NN and SVM models for lead times of 24, 30 and 36 hours ahead compared to the observed river levels. No performance statistics are provided since they do not reveal anything useful in relation to the way the models predict the river level using the radar data other than to suggest that the models perform very badly with the exception of *PDIFF*. However, the hydrographs reveal very interesting patterns. Once again, both the *SVM* and *ANN* predict a late start to the flood event but for lead times of 24 and 30 hours, both models predict a level of 3.7 m (or the level at which the river flows over the banks) on time. The rising limb of the hydrograph (with the exception of the start) is actually very well represented. The *SVM* then does a better job by predicting the peak very well while the *ANN* drops off very rapidly. At a lead time of 36 hours, the *SVM* continues to perform will with respect to the rising limb and the peak but the *ANN* is late and overpredicts the peak. The rest of the hydrograph is very poorly predicted and the behaviour of the *SVM* and *ANN* is very erratic. This may be due to noise in the raw radar reflectivity data that would normally be removed during the calibration process when estimating rainfall from rain gauges, i.e. the more usual way in which radar data are utilized. However, it is clear that using the spatial extent of the raw radar reflectivity data as an input to the *SVM* and *ANN* models has the potential to extend the lead time of the forecast considerably. The *SVM* shows better performance when compared to the *ANN* in terms of predicting the peak of the flood event. However, both models are able to predict a crucial element needed for flood forecasting and early warning, i.e. the time at which the water will flow over the banks of the river, with a considerable enough lead time for civil protection activities to be implemented.

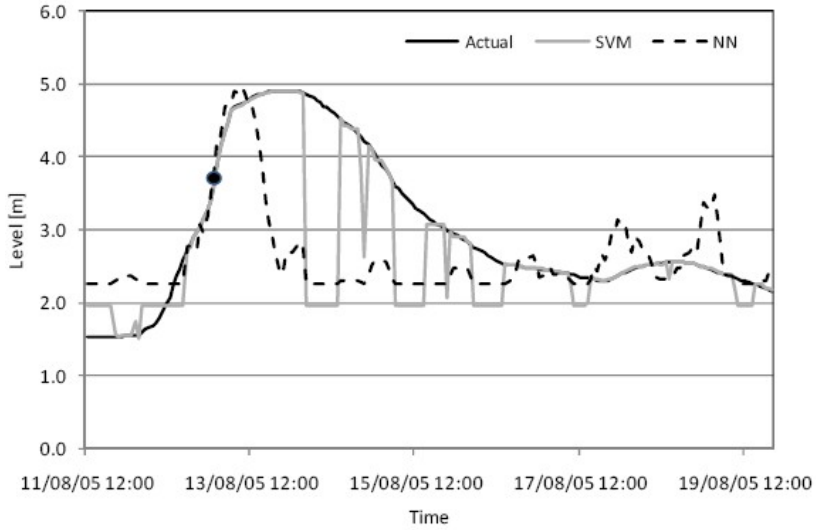

Figure 6: *SVM* and *NN* Forecasts Compared to the Observed Values for a Lead Time of 24 hours Using Radar Data as Inputs. The Dot Denotes the Level at which the Water Overtops the Bank (3.7m)

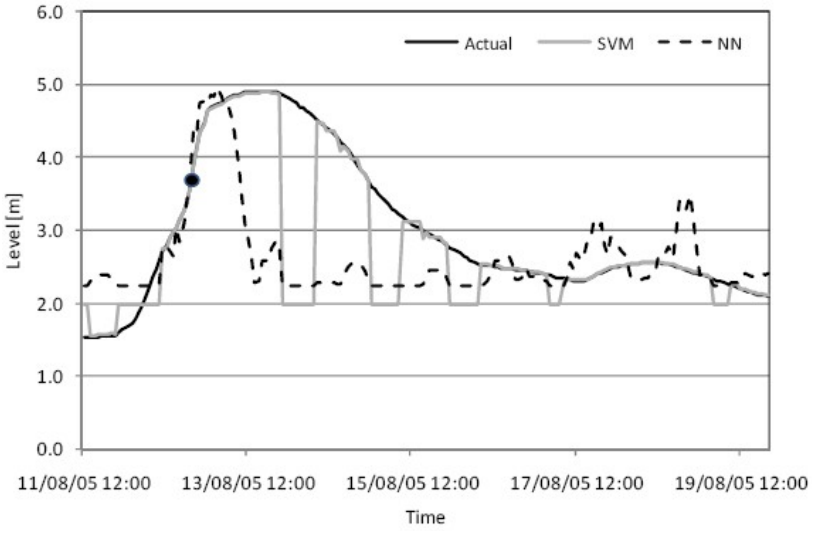

Figure 7: *SVM* and *NN* Forecasts Compared to the Observed Values for a Lead Time of 30 hours Using Radar Data as Inputs. The Dot Denotes the Level at which the Water Overtops the Bank (3.7m).

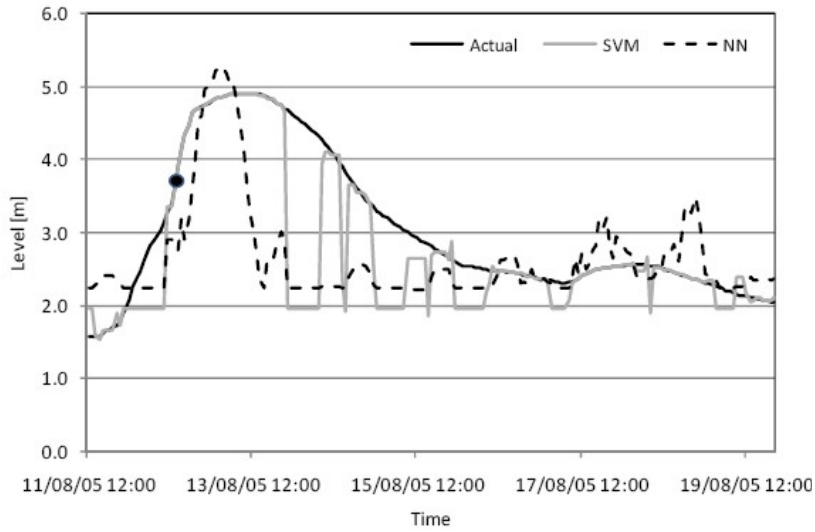

Figure 8: *SVM* and *NN* Forecasts Compared to the Observed Values for a Lead Time of 36 hours Using Radar Data as Inputs. The Dot Denotes the Level at which the Water Overtops the Bank (3.7 m)

6. CONCLUSIONS

Data-driven methods are increasingly being reported as successful methods for rainfall-runoff modelling or routing applications. However, many of these studies report short lead times that

have little relevance from an operational flood forecasting and flood warning perspective. For the Upper Ping catchment, the existing operational models work well up to a lead time of 7 hours. Chidtong *et al*., (2009) extended this further to 12 hours and showed excellent results. Experimentation in this paper attempted to extend this lead time further to 18 hours using level data from stations at Chiang Mai and upstream. However, it was clear that 15 hours is probably the maximum lead time at which good results can be achieved given water level inputs, unless investment in additional rain gauging stations is made in the upper part of the catchment. It is understood that additional gauging stations are now being put in place which could be used for model development in the future. However, given this lack of data, the question was posed as to whether radar information could be used to extend the lead time. Models were developed for 24, 30 and 36 hours ahead using both SVM and NN models. Although the general hydrograph prediction for the largest flood event in 2005 was erratic, the level at which the river overtops the banks and the peak prediction was extremely good. This study indicates that using raw radar reflectivity data as a model input provides a vast data rich potential for improving the ability to warn and defend against large floods in the future. Extension to other catchments will be the subject of future research.

Acknowledgements

We would like to give our thanks to the Bureau of Royal Rainmaking and Agricultural Aviation in Thailand for the radar images and to Mr. Chanti Deiyothin who developed the program for extracting the radar data from the images. We would also like to thank the Hydrology and Water Management Center for the Upper Northern Region for providing the water stage data.

References

[1] Abrahart R. J., See L. M., Dawson C. W., Shamseldin A. Y., and Wilby R. L., (2010), Nearly Two Decades of Neural Network Hydrological Modeling, In: Sivakumar, B. and Berndtsson, R. (Eds.) *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting*. World Scientific.

[2] Anctil F., (2007), Tools for the Assessment of Hydrological Ensemble Forecasts, In: *Proceedings of the International Workshop on Advances in Hydroinformatics 2007*, Coulibaly P., (Eds.), 4-7 June, Niagara Falls, Canada.

[3] Bishop C. M., (2005), Neural Networks for Pattern Recognition, Oxford, Oxford University Press.

[4] Boochabun K., Tych W., Chappell N. A., Carling P. A., Lorsirirat K., and Pa-Obsaeng S., (2004), Statistical Modeling of Rainfall and River Flow in Thailand, *Journal of the Geological Society of India*, **64**: 503-515.

[5] Boser B. E., Guyon I. M., and Vapnik V. N., (1992), A Training Algorithm for Optimal Margin Classifiers, In D. Haussler, Editor, *5th Annual ACM Workshop on COLT*, p. 144-152, Pittsburgh, PA, ACM Press.

[6] Chaipimonplin T., (2010), *An Exploration of Neural Network Modelling Options for the River Ping, Thailand*, Unpublished PhD Thesis, School of Geography, University of Leeds.

[7] Chaipimonplin T., See L. M., and Kneale P. E., (2010), Using Radar Data to Extend the Lead Time of Neural Network Forecasting on the River Ping, *Disaster Advances*, **3**(3).

[8] Chatchawan S., (2005), The Report of Flood's Situation and Resolving at Chiang Mai City 2005 (Published in Thai), *2005 Annual Conference*, Chonburi.

[9] Chidthong Y., Tanaka H., and Supharatid S., (2009), Developing a Hybrid Multi-Model for Peak Flood for Forecasting, *Hydrological Processes*, **23**: 1725-1738.

[10] Cortes C., and Vapnik V. N., (1995), Support Vector Networks, *Machine Learning*, **20**: 273-297.

[11] Cristianini N., and Shawe-Taylor J., (2000), *An Introduction to Support Vector Machine*, Cambridge, Cambridge University Press.

[12] Dawson C. W., Abrahart R. J., and See L. M., (2007), HydroTest: A Web-Based Toolbox of Statistical Measures for the Standardised Assessment of Hydrological Forecasts, *Environmental Modeling & Software*, **27**: 1034-1052.

[13] Drucker H., Donghui W., and Vapnik V. N., (1999), Support Vector Machine for Spam Categorization, *IEEE Transactions on Neural Networks*, **10**(5): 1048-1054.

[14] Haykin S., (1999), Neural Networks, 2nd Ed., Prentice-Hall, Upper Saddle River, N.J.

[15] Hu Y. H., and Hwa J.-N., (2002), *Handbook of Neural Network Signal Processing*, CRC Press, Boca Raton, FL.

[16] Hydrology and Water Management Center for Upper Northern Region, (2005), Report of Flooding in Chiang Mai City: 14-16 August 2005 (Published in Thai). Chiang Mai.

[17] Hydrology and Water Management Centre for Upper Northern Region, (2007a), *Ping Report* [Online], [Accessed 16 November 2007], Available at *http://www.hydro-1.net/DATASHOW/dspic/ping-report.html*

[18] Hydrology and Water Management Centre for Upper Northern Region, (2007b), *Upper Northern Region Flood Warning Brochures* [Online], [Accessed 16 November 2007], Available at *http://www.hydro-1.net/*

[19] Kamruzzaman J., Begg R. K., and Sarker R. A., (2006), *Artificial Neural Networks in Finance and Manufacturing*, Idea Group Publishing, Hershey, PA.

[20] Mackay D. J. C., (1992), A Practical Bayesian Framework for Back Propagation Networks, *Neural Computation*, **4**(3): 415-447.

[21] Maier H. R., Jain A., Dandy G. C., and Sudheer K. P., (2010), Methods Used for the Development of Neural Networks for the Prediction of Water Resource Variables in River Systems: Current Status and Future Directions, *Environmental Modeling & Software*, **25**: 891-909.

[22] Mapiam P. P., and Sriwongsitanon N., (2009), Estimation of the URBS Model Parameters for Flood Estimation of Ungauged Catchments in the Upper Ping River Basin, Thailand. *Sciences Asia*, **35**: 49-56.

[23] Minns, A.W. and Hall, M.J. (1996) Artificial Neural Networks as Rainfall-Runoff Models, *Hydrological Sciences Journal*, **41**: 399-417.

[24] Natural Disasters Research Unit (2007b) *Flood Forecasting System* [Online], [Accessed 16 November 2007], Available at *http://www.cendru.eng.cmu.ac.th/engflood/*

[25] Natural Disasters Research Unit (2007c) *General Information of Upper-Ping Basin* [Online], [Accessed 16 November 2007], Available at *http://www.cendru.eng.cmu.ac.th/flooding/?name=/chapter1/cp1_4/artical4*

[26] Ninprom S., and Chumchean S., (2009), Effectiveness of Bias Adjustment Technique to Reduce Errors in MM5 Rainfall Forecasting on Flood Forecasting Results of the Upper Ping River Basin, *The 14th National Convention on Civil Engineering ({ublished in Thai)*, Suranaree University of Technology, Nakhon Ratchasima.

[27] Patsinghasanee S., Lipiwattanakarn S., and Sriwongsitanon N., (2004), An Optimized Back Propagation Neural Network for Flood Forecasting in The Ping River, *The 9th National Convention on Civil Engineering (Published in Thai)* Thailand.

[28] Rodratana P., and Piamsa-Nga N., (2008), 'The Study of Flood Alleviation of Chiang Mai City Area', *The 13th National Convention on Civil Engineering (Published in Thai)*, Phataya, Thailand.

[29] Rumelhart D. E., Hinton G. E., and Williams R. J., (1986), Learning Internal Representations by Error Propagations, In: Rumelhart D. E., and McClelland J. L., (Eds.), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, (MIT Press, Cambridge, Massachusetts, USA), **1**: 318-362.

[30] Russell, S. and Norvig P., (1995), *Artificial Intelligence: A Modern Approach*, Prentice-Hall, Saddle River, NJ.

[31] Scholkopf B., (2007).

[32] Schreider S. Y., Jakeman A. J., Gallant J., and Merritt W. S., (2002), Prediction of Monthly Discharge in Ungauged Catchments Under Agricultural Land Use in the Upper Ping Basin, Northern Thailand, *Mathematics and Computers in Simulation*, **59**: 19-33.

[33] See L. M., (2008), Data Fusion Methods for Integrating Data-Driven Hydrological Models, *Studies in Computational Intelligence (SCI)*, **79**: 1-18.

[34] Smola A. J., and Scholkopf B., (2004), A Tutorial on Support Vector Regression, *Statistics and Computing*, **14**: 199-222.

[35] Solomatine D., See L., and Abrahart R. J., (2008), Data-Driven Modeling: Concepts, Approaches and Experiences, In: Abrahart R. J., See L., and Solomatine D. P., (Eds), *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*, Heidelberg, Springer-Verlag, 17-30.

[36] Sophhonphattanakul S., Wangwongwiroj N., and Israngkura U., (2009), Effect on Change in Land Uses on Stream Flow in the Upper Ping River Basin, *The 14th National Convention on Civil Engineering (Published in Thai)*, Nakhonratchasima, Thailand.

[37] Sukka P., (2005), Forecasting of Daily Inflow to Large Reservoir in Upper Ping River Basin Using Artificial Neural Network, Master Thesis, *Civil Engineering*. Chiang Mai, Chiang Mai University.

[38] Taesombat W., and Sriwongsitanon N., (2006), An Evaluation of the Effectiveness of Hydrodynamic Models Application for Flood Routing Investigation in the Upper Ping River Basin, *Engineering Journal Kasetsart*, **20**: 74-82.

[39] Taesombat W., and Sriwongsitanon N., (2010), Flood Investigation in the Upper Ping River Basin Using Mathematical Models, *Kasetsart Journal Natural Science*, **44**: 152-166.

[40] Thaisawasdi O., Sriwongsitanon N., and Lipiwattanakarn S., (2007), Daily Flow Estimation for Small Ungauged Basins Using Artificial Neural Network Models, *The 12th National Convention on Civil Engineering (Published in Thai)* Phitsanulok, Thailand.

[41] Tingsanchali T., and Gautam M. R., (2000), Application of Tank, NAM, ARMA and Neural Network, Models to Flood Forecasting, *Journal of Hydrological Processes*, **14**: 2473-2487.

[42] Vongtanaboon S., Lim H. S., and Richards K., (2008), Data-Based Mechanistic Rainfall-Runoff Modeling for a Large Monsoon Dominated Catchment in Thailand, *Silpakorn University Sciences and Technology Journal*, **2**: 14-28.

[43] Walczak S., and Cerpa N., (1999), Heuristic Principles for the Design of Artificial Neural Networks, *Information and Software Technology*, **41**: 107-117.