Configuring ANN for Inundation Areas Identification based on Relevant Thematic Layers

Supattra Puttinaovarat[1], **Paramate Horkaew**[2], and **Kanit Khaimook**[3], Non-members

ABSTRACT

Flooding is one of the major natural disasters that can cause not only disruptions to daily lives but also damages to our properties. The study of flood models to determine inundation areas has therefore become increasingly more important for decision makers and the authorities. Being readily accessible, the Artificial Neural Network (ANN) scheme was frequently adopted for hydrology and flood modeling. Most ANN techniques primarily take into account rainfall data and then predict runoff consequences. Despite convincing successes, they usually neglect other causative factors. This study thus focuses on configuring and improving generic ANN for inundation areas identification using various flood deterministic attributes. Accordingly, an ANN was developed using nine flood causative factors, derived from relevant thematic layers. They consist of flood plain in the past, height above sea level, water density, water blockage, sub basin areas, soil drainage capability, land uses and monthly rainfall, whose prognostic values were previously reported and assessed against the full scale census and comprehensive GIS survey with satisfying cogency. The guidelines and precautions suggested in this paper may be applied to various ubiquitous ANN based frameworks for flood forecasting and related risks assessment.

Keywords: GIS, Flood Prediction, ANN

1. INTRODUCTION

Floods are one of the most frequent natural disasters that cause severe damage to both human lives and properties. It is estimated that out of global disasters, 41.4% are due to flooding [1]. Thailand, for instance, has been subject to recurring floods, whose incidents reported in the official documents, dated since as far as early 90's. More recently, according to the Department of Disaster Prevention and Mitigation (DPM), Thailand has experienced the major floods in 2011, which severely affected residences around Chaopraya river basin. The total estimated cost in terms of property lost due to the event amounted to 1.44 trillion Baht, hence considered as one of the most costly catastrophes in the nation's history. This striking event as a result was placed 4$^{\text{th}}$ in the most severe disasters world-wide ranking table [2]. Pathumthani had also been greatly suffered by this event. The economic affliction are so far estimated to be over 0.1 trillion Baht. Its extent covered 7 districts, 60 subdistricts and 522 (or 98.7%) villag-es.

Forecasting the event prior to its actual occurrence is an attempt to prepare for the inevitable. To this end, Hydrology, whose doctrine could be dated back to as early as the pre-historic era, based on for example, natural observations and more often than not, with religious beliefs affiliation. Nonetheless, its more modernized research has evolved into scientifically constructing a model of hydrological processes, with the goal to offer timely and accurate estimates of future discharge at specific watershed locations. A wide variety of forecasting techniques avail-able typically adopted rainfall-runoff modeling (Fig 1) and many models have been proposed with differing modifications. Thus far, there remain deficits: these models were unable to generalize well and thus barely overcome higher-than-average accuracy limit, thanks to intransigent over other spatial causative factors. More recent years, there have been many studies on flood susceptibility and flood prediction using Geographical Information System (GIS). They commonly rely on analyzing relevant predicting factors, associated with weights and rates specified by human experts, normally are employed from variety of fields. Furthermore, collecting these information is not only time consuming but also impractical and inappropriate in many cases. Land slope, for instance, is a prime factor, as it retains high significance in flood predictability for Angthong province [3]. This same factor however, is not eligible for those areas on river basin such as Pathumthani province, whose slopes fall within as narrow as 0-5% range [2].

To elucidate various flood prediction models, a more recent investigation [4] studied prominent tech-

Manuscript received on March 31, 2013 ; revised on June 26, 2013.

Final manuscript received September 2, 2013.

[1,3] The authors are with School of Information Technology,Institute of Social Technology, Suranaree University of Technolo-gy, Nakhon Ratchasima, Thailand., E-mail: supattra.p@psu.ac.th and kkanit@sut.ac.th

[1] The author is with School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand., E-mail: phorkaew@sut.ac.th

[2] The author is with Faculty of Science and Industrial Technology, Prince of Songkla University, Suratthani Campus, Suratthani, Thailand., E-mail: supattra.p@psu.ac.th

[2] Corresponding author.

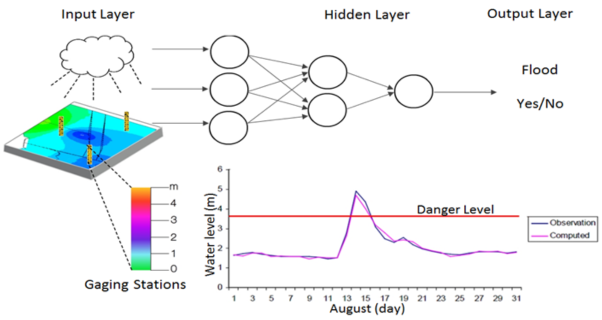

Fig.1: *Graphical illustration of typical ANN for discharge prediction.*

niques, *i.e.*, Artificial Neural Network (ANN), Genetic Algorithm (GA), Fuzzy Logic and Self Organizing Map (SOM) and as-sessed their applicability on predicting flood. Bench-marked against the indicators given by 100 human ex-perts, the ANN was found the most consistent and promising, despite however some concerns impeding it from being readily applicable.

Inspired by [4] and in the quest to remedy the limitations associated with the scheme, this paper focuses on further tuning the ANN model. Specifically, based on one of the most extensive and recent GIS data available, the Multi Layer Perceptron (MLP) and the Radial Basis Function (RBF) structures were experimented. The underlying rationale regarding each constituent factor will be discussed. In addition, guidelines and insight precautions on applying the ANN on predictions of flood and probably similar disasters were also addressed.

This paper focuses on configuring ANN for inundation areas identification based on various flood causative factors (Fig 2), which were selected based on knowledge acquired from during our preliminary on-site survey [4]. There were indeed many recent studies that adopted MLP and RBF in flood prediction [5-9]. What these techniques had in common is that they built a model which predicts run-off (or discharge) from rainfall, without taking into account other flood causative factors. Moreover, their common deficit is that prediction could only be made at watersheds and where gauging stations exist (typically sparse). This limited coverage as a result constrains the inundation areas being identifiable merely at locations where water level can be measured. Nonetheless in their own right, both MLP and RBF are versatile and have a strong merit as prediction network if properly implemented. Compared with other ANN based methods, many studies [10-13] have confirmed MLP and RBF superior accuracy over, for instance, LVQ and others. The Adaptive Resonance Theory (ART) neural network, despite its simplicity, are not fully developed and currently criticized for its statistical inconsistency and high de-pendency on the order in which the data are processed. Furthermore, we have previously reported [4] that when benchmarked against prominent learning based systems, the ANN outperformed GA, SOM and Fuzzy Logic in terms of flood prediction accuracy and model generalization. Consider the Self Organizing Map (SOM) for example; it could predict the flood no better than an educated guess, and hence not particularly useful. After closer inspection into our raw data, it was revealed that in those cases the over-fitting had overruled model generalization. More specifically, SOM ranked *Terrain Height* as almost the least significant factor due to the fact that Pathumthani terrain is generally flat; therefore flood probability was considered by SOM to be more dependent on other factors. This is however, not the case in some areas that survived the flood, mostly due to its relatively raised level. Discussion on shortcomings of the remaining techniques can also be found in [4]. Last but not least, research in MLP and RBF are relatively matured and as such they are currently widely accessible to hydrological organizations, whose prolonged learning curve in adopting cutting edge AI paradigm could cause an adverse effect for unhindered risk management.

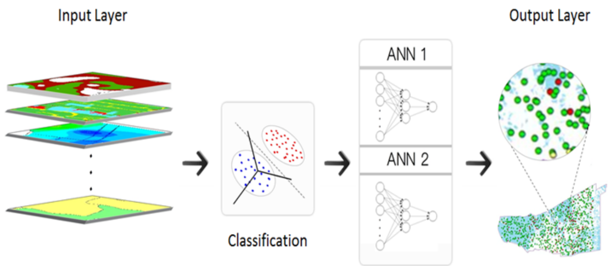

Fig.2: *ANN for flood prediction proposed in this study.*

Following aforementioned rationales, the aim of this study is to investigate the readily available techniques and provide the guideline design a properly implementation of the flood prediction model. To this end, MLP and RBF were chosen as our primary schemes. Their merits were elucidated based on GIS data recorded during the actual flooding in 2006, 2010 and 2011. Their configurations were then adjusted accordingly to best reflect and predict the actual events as they occurred. The resultant model may serve as a tool in flood hazards and risks assessment. Moreover, its deriving process could be considered as the guidelines and precautions for adopting an upcoming learning based flood prediction paradigm in the future.

2. THEORY AND RELATED MATERIALS

This section discusses related learning techniques, adopted previously in the literatures namely GA, Fuzzy Logic, SOM and ANN [14-19]. Overviews of the former three systems and their usages in flood predictions are briefly mentioned, while more emphasis will be put on detailed analyses of ANN variants and their applications. Subsequently, Geographical Information System (GIS) framework will be brought into this context and com-ments on related materials are provided.

2.1 GA, Fuzzy Logic and SOM

Genetic algorithm is search and optimization techniques based on the principles of national selection and genetics. In solving a hydrological problem, GA has been used in determining fountainhead [15]. In that study, it serves as an automatic fuzzy rule selector, with no expert input required. The resultant rules were then executed to estimate rainfall, based on temperature, humidity, wind speed and solar radiation. The derived model was able to predict the rainfall a day in advanced with 20-80% error, depending on how early the event was participated [16].

Fuzzy logic is a decision making assisting tool that can tackle a problem at hand quickly and efficiently and cope well with imprecision and nonlinearity. The generation of a fuzzy forecast models were based on both expert knowledge and historical data [15]. The model was found in hydrological forecasting such as real-time flood prediction by taking into account hourly rainfall and discharge. It was able to make prediction 12 hours prior to the event, with 63-98% accuracy, depending on how early it was predicted [17].

Finally, SOM is an ANN that can learn without explicit training [18]. It was incorporated into the ANN to determine water, using hourly rainfall gauged from nine stations. It can make prediction 2 hours before hand, with a promising 94% accuracy [19].

2.2 Artificial Neural Networks (ANN)

ANN is a mathematical model of human perception that can be trained to perform a particular task based on available empirical data. Particularly, when the relationships between data are unknown, the ANN can efficiently establish their underlying connections. There exist many ANN variants, which can be categorized into 6 architectures, *i.e.*, Multi Layer Perceptron (MLP), Radial Basis Function (RBF), Learning Vector Quantization (LVQ), Adaptive Resonance Theory (ART), Auto-Associative NNs and Self Organizing Map (SOM) [18-22]. This study focuses particularly on MLP and RBF types

MLP is a configuration of an ANN that has multiple layers and is suitable for complicate scenario. The configuration employs Back Propagation learning scheme that consists of 2 reciprocal procedures, *i.e.*, Forward Pass and Backward Pass. The former passes the data presented at the input layer through one hidden layer after an-other, while the later adjusts the connecting weights ac-cording to an error correction rule, whereby the error signal which is the difference between the target and actual responses is fed back into the network in the opposite direction to the forward links, whose weights are then iteratively tuned until the actual response reach the specified target.

RBF ANN is also considered of multilayer and forward feeding configuration. It has the ability to simplify complex problems, by means of non-linear mapping. As the complexity increases, adding more perceptron layers to ANN could lead to longer computational time and hence slower convergence. Moreover, discrimination lines are unable to separate among clusters that are naturally. In these scenarios however, it may be more tangible to incorporate into ANN a circular shape discriminator, *e.g.*, Gaussian, whose extent covers most of the plausible region in each cluster. This configuration consists of 3 layers, *i.e.*, n-neural input, one hidden, and n-neural output layers. The connectivity between input and hidden layers are hypothetical, while the one between hidden and out-put layers is specified with a set of adaptive weights that are adjusted during the neural training process.

The current trend in GIS is adopting the ANN in fore-casting rain volume, runoff and water level for flood warning systems [5-7, 23-24]. The techniques employed either MLP or RBF in their implementations. Their result-ant predictability was 70-95% accurate and was able to forecast hydrological events 3-7 hours in advance. The factors incorporated were local slope, height, accumulated flow, hourly water level at gauging stations, rainfall measure, land utilization and soil and other geological characteristics, *etc.* [25].

In our implementation, the MLP was defined as per equations (1) and (2)

$$x_j(p) = \sum_{i=1}^{n} x_i(p)w_{ij}(p) + b_j(p)w_j(p) \quad (1)$$

$$y_j(p) = \frac{1}{1+e^{-x(p)}} \quad (2)$$

where p is the data set, n is the total number of input nodes, $x_i(p)$ is a sampled data present at the i^{th} node, $w_{ij}(p)$ is the weight assigned to the link between the i^{th} and the j^{th} nodes, $b_j(p)$ and $w_j(p)$ are the bias and weight linking between bias and the j^{th} node and $y_j(p)$ is the response at node the j^{th} node. As for the RBF configuration, equation (3) was employed.

$$F(x_j; c_i) = \sum_{i=1}^{p} w_i \exp\left(-\frac{N}{d^2}\|x_j - c_i\|^2\right) \quad (3)$$

where p is the number of kernels, N is the number of hid-den layer nodes, d is the maximum distance between each data point to the mean, x_j is a sampled data present at the j^{th} node. c_i is the kernel value and wi is the weights linking between input nodes.

2.3 Geographical Information System (GIS)

GIS is a computerized process that has been recognized as a powerful means of integrating and analyzing data from various sources in the context of comprehensive floodplain management. Adequate information and prediction capability is vital to evaluate alternative scenarios for flood mitigation policies and to improve decision making processes associated with flood management, causes and factors contributing to floodplain mapping [3], such as rainfall, land slope, height above sea level, sub-basin size and density, water passage blockage and soil drainage [25 - 28]. Furthermore, GIS has been incorporating with other schemes, for instance, satellite imaging, in flood management.

3. METHODOLOGY

In the preliminary study [4], where prominent learning based prediction systems, *i.e.*, ANN, GA, Fuzzy Logic and SOM were benchmarked, generic ANN was found to be the most superior in terms of flood predictability. In that study, key factors which were incorporated into the model were rainfall, flood plain in the past, height above sea level, water density, water blockage, sub-basin area, soil drainage ability, and land utilization. Based on the total of 100 spatial data, which were uniformly sampled from the flood event recorded in 2011, several statistical analyses were employed to validate and compared those learning models [4]. It was reported that, for instance, ANN and Fuzzy Logic performed equally well, with the best accuracy of 92%. However, when considering their generalization ability, using Leave-one-out cross validation (LOOC) on 6 but 1 district, and taking into account both false positive and negative, ANN could predict the event with the least error of 30.2%, as shown in Table 1.

Table 1: *Comparison of ANN, GA, Fuzzy Logic and SOM.*

Method	*ANN*	*GA*	*Fuzzy Logic*	*SOM*
Accuracy	92%	88%	92%	88%
Error	30.2%	57.9%	77.2%	52.7%

Following this and other findings [4], concerns that were needed to be raised when building (training) such prediction models included that data sampling should be able to cover possible variations without leaping into over fitting, where more points were unnecessarily included, thus making the model specific yet unable to generalize. In addition, each factor should be quantized, normalized and rated properly so as to avoid variations in some factors being submerged under other more dominants' noise.

The aim of this paper therefore is to further improve ANN based flood prediction system, thereby configuring it into MLP and RBF types, bearing in mind those issues. The predictability of the ANN in both configurations will be assessed and compared against the flood data as it occurred in Pathumthani, 2011. The work presented here differs from others in that, unlike [5-7, 23-24], where flood peak was predicted hourly in advanced under the hypothesis that the water levels at gauge stations were provided, it employed 9 factors (an additional factor than [4] was added) into the MLP and RBF ANNs for flood prediction, from which an optimal approach is deduced.

3.1 Data Preparation

By relevant thematic layers, this study means rainfall, flood plain in the past, terrain elevating, water block-age, soil drainage capability, and monthly rainfalls, *etc.*, which were recorded in the geographical form.

In order to develop the flood model, understanding and determination of flood causative factors are crucial. These factors are selected based on the knowledge acquired from collected during our recent on-site survey. A public hearing was organized and 100 local administrative governors, sub-districts headmen, village chiefs, academic and local representatives in the area at risk were participated [2]. This information formed the bases of the overlay functions, constituting to nine relevant factors, which were sampled and analyzed on a reference frame, *i.e.*, Digital Elevation Model (DEM) digitized from elevating contours [2, 4].

3.2 Deterministic Attributes

Rainfall (referred herein as P1) as depicted in **Fig 3**, has a direct relationship with the amount of water in the water passage, which in turns has an influence upon flood happening in respective area.

Some areas were repeatedly flooded as they are practically more prone to reoccurring of the event than others. Flood plain in the past (P2) as mapped in **Fig 4**, therefore is another key attribute that may be taken into account to determine whether a location is likely to be affected.

Terrain elevation (P3) is evidently one if not the most prominent attribute that has strong determinant on flood. Public generally assume that the higher the area, the less likely it would be affected. It is normally measured as the height above sea level. In this study, the information was uniformly digitized from elevation contours, as in **Fig 5**.

Water density (referred as P4) of the sub-basin facilitates drainage of areas nearby, making them less

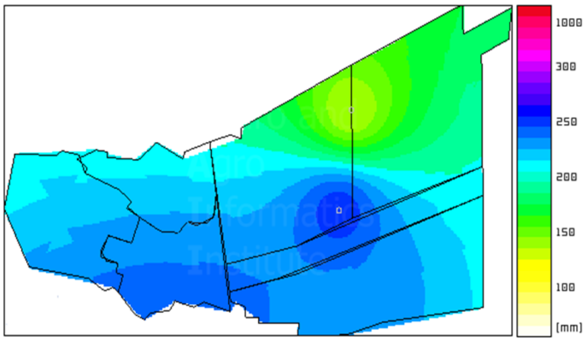

Fig.3: *Averaged Rainfall in mm unit.*

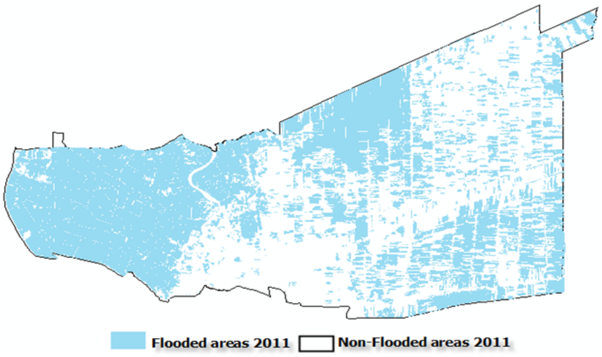

Fig.4: *Flood Plain in the past.*

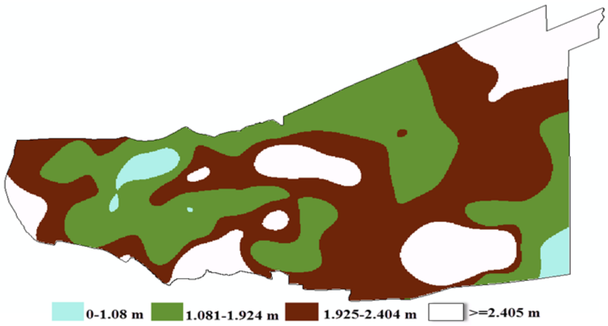

Fig.5: *Height above Sea Level.*

likely to be affected. The water density is computed as the ratio between the total length of water passage and the coverage basin area, as described in Equation 4.

$$D_d = L/A \tag{4}$$

where D_d is the water density (1/km) and L is the total length of the water passage in that basin, whose area is A.

Water blockage (P5), e.g., land transits and dyke, on the other hand, impedes and slows down the overflow to be discharged from the area, putting it at higher risk. The blockage is similarly computed as the water density, but the length of the blockage replaces that of the passage. The map of blockage in the area is illustrated in **Fig 6**.

Fig.6: *Water Blockage in Pathumthani.*

The extents of sub-basin areas (P6) play a major part in the outgoing flow through water passages. Prolonged flood is almost inevitable, if the sub-basin is larger than the river and cannel flow capacity. **Fig 7** illustrates sub-basins, where the larger areas roughly coincide with those greatly affected during the 2011 deluge [2].

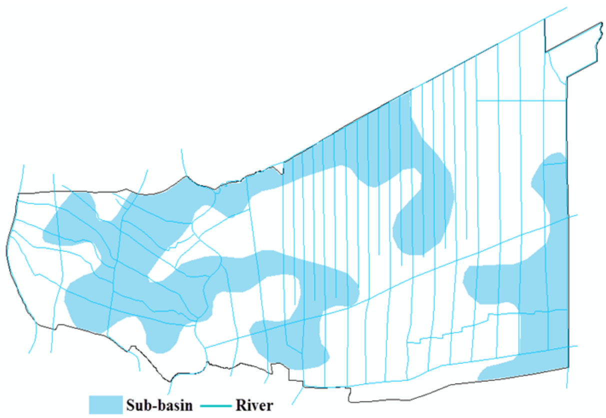

Fig.7: *Sub basin areas.*

Soil texture is another factor proven to be connected with flood. Soil with fine texture such as that with a high proportion of clay particles may disrupt the water drain-age from the area, causing swamps. Soil with a lower clay proportion, on the contrary, accelerates the draining. This attribute was expressed in our experiment in terms of soil drainage capability (P7), and was categorically divided per administrative district into 2 levels (*i.e.*, bad and control) as illustrated in **Fig 8**.

Land use (P8) is another factor that determines the possibility of a given area taking flooded. Florae covering the area (*e.g.*, forest and perennial) could reduce that risk, as they could absorb water and hence delaying the flush. The map labelled with land uses is depicted in **Fig 9**.

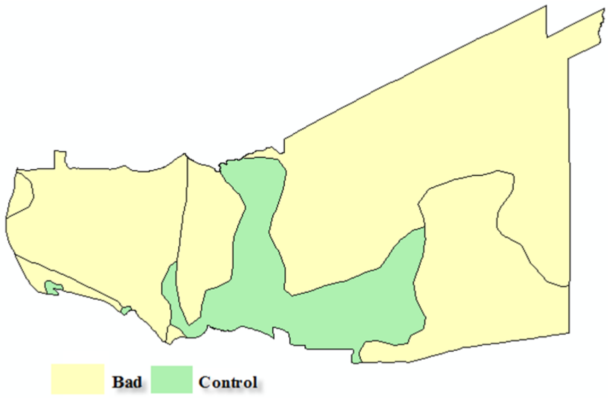

Fig.8: *Soil Drainage Capability.*

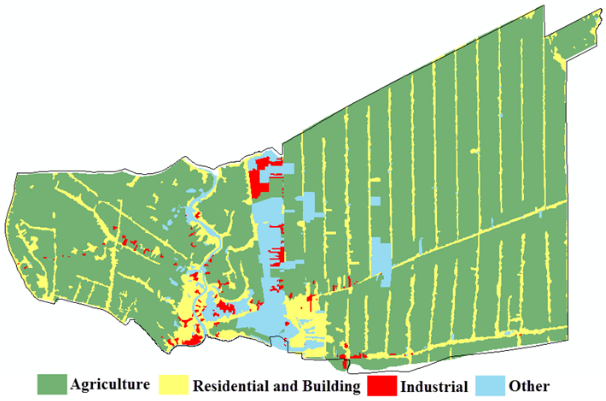

Fig.9: *Land Uses.*

Monthly rainfall (P9) is the last but not least factor that could improve flood predictability, enabling the pin point of the event at a given time frame. It was neglected before in the previous study, but now is considered here.

Since Pathumthani is repeatedly flooded, the factors mentioned above will be sampled from the GIS recorded in 2011, 2006 and 2010, when the city was in respective order, the most severely affected by flood. The reason for those 3 years were taken into account instead of a specific one alone is to remove irregularities that might have been caused by other data irrelevant matters, for in-stance, politics, and administrative ambiguity and disagreements, *etc.* The sampling was made uniformly 1000 instances/ year.

3.3 Data Preprocessing

The relevant thematic layers were derived from various sources and differed in units. It was therefore difficult to interpret and probably led to missed classification. To ensure homogeneity, each attribute was prepared by first rated its raw value into discrete levels (bands) as tabulated in Table 2, and later applied min-max normalization, so as to avoid the effect of quantization and thus suppressing inter-modulation noise floor.

Table 2: *Factors taken into account with rating.*

Factor	Data Range	Rating
Rainfall(P1)	$\geq$1401 mm	5
	1301-1400 mm	4
	1201-1300 mm	3
	$\leq$ 1200 mm	2
Floodplain in the (P2)	Flooded	5
	Non-flooded	2
Height above sea level (P3)	0-1.080 m	5
	1.081-1.924 m	4
	1.925-2.404 m	3
	$\geq$2.405 m	2
Water density (P4)	0-2458.9 km^{-1}	5
	2459.0-8885.5 km^{-1}	4
	8885.6-36103.7 km^{-1}	3
	$\geq$36103.8 km^{-1}	2
Water blockage (P5)	$\geq$2.27 km^{-1}	5
	0.01-2.26 km^{-1}	4
	0 km^{-1}	3
Sub-basin area (P6)	$\geq$11671 km^2	5
	2575-11670 km^2	4
	$\leq$ 2574 km^2	3
Soil Drainage Capability (P7)	Bad	5
	Controllable	3
Land Uses (P8)	Residential & Bldgs.	5
	Agriculture	4
	Industrial	3
	Other	2
Monthly Rainfall (P9)	$\geq$236 mm	5
	221-235 mm	4
	191-220 mm	3
	$\leq$190 mm	2

3.4 Training and Flood Predicting Processes

Once the data had been rated and normalized, they were then reformatted into ARFF, so that they could be fed into either MLP or RBF configured ANN. The learning process was performed using WEKA [29].

During the training and predicting, the K-fold Cross Validation [30, 31] was employed as a metric to measure the merit of each configuration. With the scheme, the data were divided equally into K sets. For each round, one data set was selected for training the model, while the remaining K - 1 datasets were used for testing, from which the results were compared with the actual event, and hence the predicting error was calculated. In order to minimize data dependent biases, the process was repeated K times, each with training data iterating through a different dataset in turns. In this study, the value K was set to 10. The errors were averaged.

As for the ANN, the learning rate was set to 0.2; the momentum coefficient was set to 0.1; and the number of iterations was set to 1500. The reason for these values were so empirically specified was that, from our experiment these combinations yielded the best predicting results and thus were reported here. Since one of the main purposes of this study was to come up with the optimal model between MLP and RBF, they were therefore configured using the same topology, *i.e.*, 9 input layer nodes for P1 - P9 abovementioned factors, 3 hidden layers and 2 output layer classes, where 0 and 1 meant not being and being flooded, respectively.

4. RESULTS

The results predicted were compared against what actually happened in 2011, 2010 and 2006 in Pathumthani, whose data were gathered during our prior investigation [2].

In order to determine a proper topology of the neural, the number of three hidden layer nodes was firstly set to 10:10:10, 5:5:5, and 5:4:3. It was found as shown in the Table 3 that (the units were in percents), the hidden layers of 5 nodes yield the best overall accuracy and therefore will be opted.

Table 3: *Comparisons among Different Hidden Nodes.*

Hidden Layer Nodes	2011	2010	2006
10:10:10	74.9	74.9	80.8
5:5:5	**73.8**	**78.4**	**80.6**
5:4:3	73.2	73.2	80.8

The number of iterations for learning process was similarly determined by comparing results in those years with the predictions given by 500, 1500, 2000 and 5000 rounds trained neural. The results as reported in Table 4 indicate the 1500 learning iterations neural well balanced between accuracy and time complexity.

Table 4: *Comparisons between Different Iterations.*

Number of Iterations	2011	2010	2006
500	72	76.2	80.6
1500	**73.8**	**78.4**	**80.6**
2000	73	77.9	80.8
5000	72.8	78.2	80.6

Once the ANN topology was determined, we tested our hypothesis of extension from [4] that, if monthly rainfall (P9) could be included in the factors, augmenting the model with it would increase the predicting accuracy. As references, MLP and RBF model were trained with P1 - P8 data from 2011, 2010 and 2006. Except in 2006 where both models did equally well, the MLP could pre-dicted the flood better than the RBF, with the accuracies of 68.8, 77.8 and 81.1 percents, respectively compared to those by using RBF of 63.4, 73.9 and 81.1 percents (**Fig 10**). When including P9 into the learning of both models, the predictability increased as anticipated for 2011 and 2010, yet slightly dropped in 2006.

It had been learnt from [4] that, increase in accuracy alone may not however be the only justification of including this particular factor into the model, as there existed risk of over fitting. To clarify the decision, the accuracy predicted in each year was averaged over the models with 8 and 9 factors. The MLP did better than RBF with the accuracies of 71.3, 78.1 and 80.85 percents compared to 64.8, 74.45 and 81.05 percents (**Fig 11**).

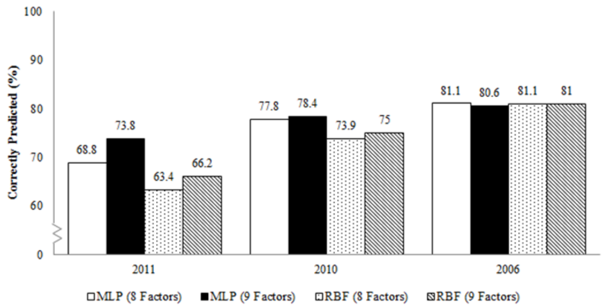

Fig.10: *Comparing MLP and RBF using 8 and 9 Factors.*

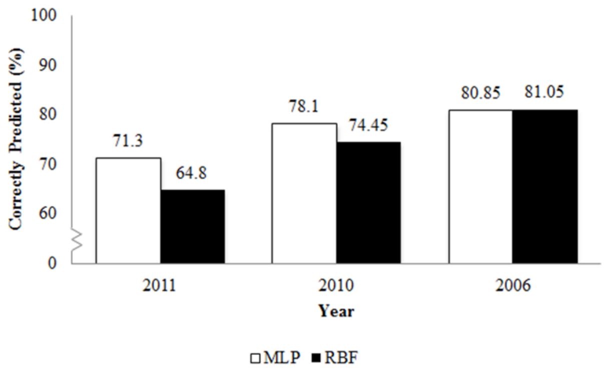

Fig.11: *Comparing MLP and RBF.*

Based on this average, we were then able to conclude that, by including the monthly rainfall (P9), the accuracy was better in 2011, while improvements in 2010 and 2006 were almost undetectable. Looking into the raw the data, the contrast of P9 across the city was higher in 2011 than in 2010 and 2006, which implies that P9 could be faithfully employed, without contaminating it with the background noise even if the rainfall did not greatly differ. Nonetheless, should the differences exist (as is the case in 2011) P9 could play a part in improving the accuracy.

It was also pointed out earlier [4] that the training da-ta should cover all plausible variations, but not overly so. This concern was proven by training the model with data from 2011 and testing it against those in 2010. The accuracy was 67.67 percents. Swapping the training and the testing data set dropped the accuracy to 42.33 percents (worse than an educated guess or even flipping a coin). This can be explained by looking into the raw data that the 2011 data had greater coverage than those in 2010. The rainfall, for instance, in 2011 was 160 - 250mm, compared to 160 - 210mm in 2010.

Figs 12 - 14 illustrate the predictability of the model with data acquired in 2011, 2010 and 2006, respectively. The blue and white areas are those affected and not affected by flood in respective year. The green dots are the *true positive*, where flood were correctly predicted or *true negative*, where they are not affected by flooded as expected. The red dots are regions that were flooded unexpectedly or *false negative*. Finally, the yellow dots are those areas where wrongly predicted as flooded or false positive. It is clear that, most of the areas are correctly predicted, with a few scattering errors.

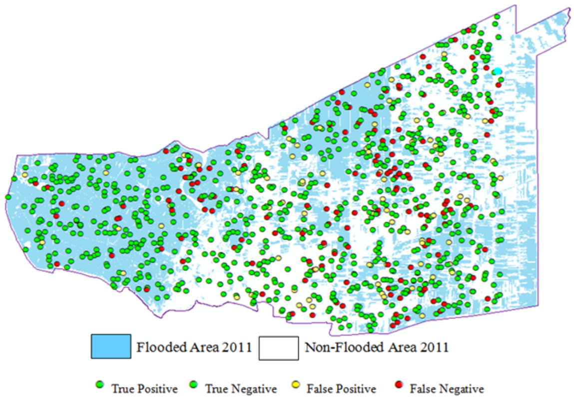

Fig.12: *Geographical Illustration of 2011 Predictability.*

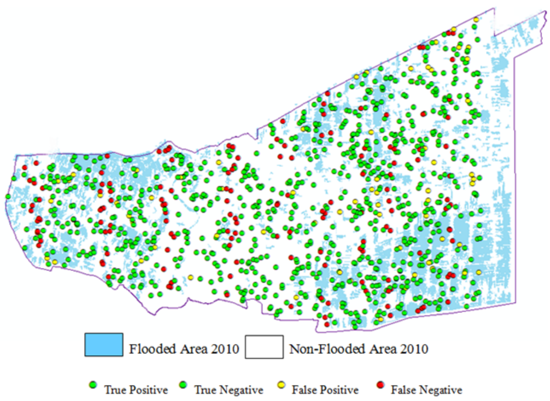

Fig.13: *Geographical Illustration of 2010 Predictability.*

The measured errors were categorically divided into 2 groups, which are *false negative* and *false positive*. According to the data acquired from 2011, 2010 and 2006. The former group is of higher proportion than the later, *i.e.*, 14.25, 15 and 18.5 percents of false positive compared to 8.3, 6.6 and 0.9 percents of false negative. In order to elucidate these findings, each of the factors and their weights were analyzed. Similar to the previous study [4], each factor was rated with low, medium or high flood risk according Table 2. It could then be deduced that, the *false negative* instance was predicted because it contained less low and

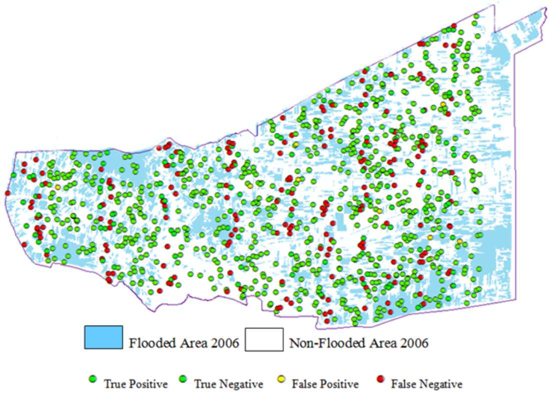

Fig.14: *Geographical Illustration of 2006 Predictability.*

medium risk factors than the high risk ones. Those lower risk factors are, based on 2011 data, of significant neural weights. In this year, for instance, the low - medium risk factors are annually and monthly rainfall, flood plain in the past, height and sub-basin are-as. Similarly, the *false positive* instance occurred because it has more data of high risk factors than those lower ones.

It was evident from the above findings that contributing factors played a very important role in flood predictability. In order to determine whether a given area would be flooded, a set of weights were assigned to these factors (attributes), forming a prediction function. In our experiment, the weights were specified by either expert or computed by the ANN. These weights are listed in Table 5. It should be noted that, P9 was only introduced in this study based on current hypothesis and not previously obtained in the survey [2], hence its absence in the expert column.

Table 5: *Weights assigned to each factor.*

Factor	**ANN Models**			
	Expert	*2011*	*2010*	*2006*
P1	8	-20.99	3.43	-0.05
P2	7	8.56	-13.01	-14.39
P3	6	-1.06	1.29	-6.96
P4	5	-17.79	-8.09	2.82
P5	4	4.77	8.64	-11.25
P6	3	0.64	0.74	0.72
P7	2	-20.35	-12.14	-2.61
P8	1	-3.34	7.56	3.65
P9	-	2.90	8.43	11.45

The analysis on the deterministic attributes (factors) that had been derived from the ANN differed from year to year depending on the training data. However on average, the most prominent (based on their magnitudes) ones are flood plain in the past (P2), soil drainage capability (P7) and water density

(P4). Although not ranked among the top 3 factors, water blockage (P5), annually and monthly rainfall (P1 and P9) also played a very important role in predicting the event. The manifestation also highlights the underlying major indicators of the flood, in this order:

1) Whether or not it has been flooded before
2) How efficiently it can dissipate the water
3) How much is the water added by rainfalls

Although the computerized models offered the predictions mostly similar to that of human recognition, the weights they assigned to the lower signified factors did not exactly coincide. These discrepancies, on one hand, strengthen the fact that artificial learning models could emulate human decision making in this context with varying degrees of success. On the other hand, they also high-light the fact that the weights specified by human experts are subjective and can only be applied in area bounded by their familiarity. They are also less reliable as they get lower, thanks to variability due to disagreement among individuals.

To further assess the generalization ability of the model, leave one out cross validation (LOOC) was used. The idea of this scheme is that all but one districts were included in the training data, leaving the removed district as the validation. The process was then repeated for the remaining districts and then the overall predicting errors were averaged across the total of 7 Pathumthani districts. However, it was hypothesized in [4] that, since the ANN are data dependent, training it with one set of data and then testing it against another with may be of rather different characteristics can be error prone and misleading.

Consequently, unlike [4], instead of simply feeding the input data into a single model, they were first into K clusters, using K-mean clustering [32], with K set to 2. Each cluster would be used as training dataset of separate ANN. Prior to predicting the unseen data; they would be classified first into one of these clusters, based on their averaged factors similarity and then assigned to the respective ANN. With this modification, it was ensured that a proper model was chosen to predict the data with similar characteristics. The results are illustrated in **Fig 15**. If unseen prediction was made by the model trained using data with similar characteristics, the accuracy are 69.53, 75.85 and 69.89 percents, in 2011, 2010 and 2006 respectively. Unless proper classification and hence model was assigned, the results was no better than flipping a coin.

The LOOC results corresponded well with the above hypothesis: learning based on similar characteristic fac-tors as the validating data results in better accuracy. In other word, the flood predicting model cannot be applied universally, unless both the training and testing attributes share the similar characteristics.

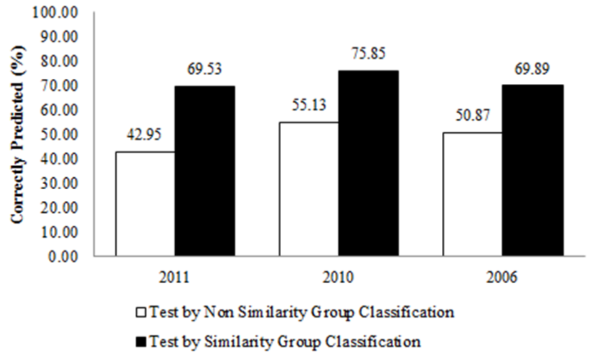

Fig.15: *LOOC using clustered and single ANNs.*

5. CONCLUSIONS

This study was carried out in the hope that it would serve as a prototype for applying the artificial learning system in public hazard (flood and landslide) monitoring and warning strategy, so as to prepare the public for the unexpected, well in advance, thus minimizing lost that could have incurred otherwise.

It was an extension of and improvement over its preliminary [4], in terms of fine tuning the model and model preparation. More specifically, 5:5:5 MLP was employed with its merits tested and a *proper* model was selected as the predictor based on their constituent data similarity.

One of the key findings worth mentioned here is that, according to the computerized ANN model, an area may be predicted if it would be flooded, depending highly on 3 indicators, *i.e.*, flooding history, water dissipation and rainfall. This interpretation offers not only the basis on further improving the flood prediction model, but also the methodological guideline for mining significant factors in different disaster scenario.

In order to achieve more accurate prediction, other realistic attributes such as debris/ weed along the water passage, land slope, local width of local water passages as well as the water influx should be quantified and taken into account when building a flood prediction model.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Public Works and Town and Country Planning, Pathumthani, for their sponsorship of the investigation.

References

[1] Asian Disaster Reduction Center. *Natural Disaster Data Book*, 2012.

[2] Suranaree University of Technology, *Prevention and Restoration of the Flood-Risked Areas in Pathumthani*, Final Report, the Department of

Public Works and Town and Country Planning, Pathumthani, 2012.

[3] S. Dhanarun and J. Amornsanguansin, "Application of Geographic Information System for Flood Risk Area Assessment in Angthong Province," *Journal of Environmental Management*, Vol. 6, No.2, pp. 19-34, 2010.

[4] S. Puttinaovarat, S. Khechonrak, P. Horkaew and K. Khaimook, "Assessing Deluge Predictability and Deterministic Attributes of Artificial Learning Systems," *Proceedings 5^{th} International Conference on Knowledge and Smart Technologies*, pp. 70-74, 2013.

[5] R. P. Deshmukh and A. A. Ghatol, "Short Term Flood Forecasting using Static Neural Networks a Comparative Study," *International Journal of Computer Science and Network Security*, Vol. 10, No.8, pp. 69-74, 2010.

[6] F. J. Chang, J. M. Liang and Y. C. Chen, "Flood forecasting using radial basis function neural net-work," *IEEE Transactions*, Vol. 31, No.4, pp. 530-535, 2001.

[7] R. Subramanian, "Implementation of Neural Networks in Flood Forecasting," *International Jounal of Scientific and Research Publications (IJSRP)*, Vol. 2, No.10, pp. 1-3, 2012.

[8] A. W. Jayawardena, et al, "Comparison of multilayer perceptron and radial basis function networks as tools for flood forecasting," *Proceedings of the conference held*, pp. 173-182, 1997.

[9] E. Mutlu, I. Chaubey, H. Hexmoor, and S. G. Ba-jwa, "Comparison of artificial neural network models for hydrologic predictions at multiple gauging stations in an agricultural watershed," *Hydrological Processes*, Vol. 22, No.26, pp. 5097-5106, 2008.

[10] A. Lahsasna, R. N. Ainon, and T. Y. Wah, "Credit Scoring Models Using Soft Computing Methods: A Survey," *The International Arab Journal of Information Technology*, Vol.7, No.2, pp. 115-123, 2010.

[11] D. West, "Neural network credit scoring models," *Computers & Operations Research*, Vol. 27, No.11, pp. 1131-1152, 2000.

[12] A. Wefky, F. Espinosa, A. Prieto, J. J. Garcia, and C. Barrios, "Comparison of neural classifiers for vehicles gear estimation," *Applied Soft Computing*, Vol. 11, No.4, pp. 3580-3599, 2011.

[13] G. Sahoo, "Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA," *International Journal of Information Technology and Computer Science (IJITCS)*, Vol. 4 No.7, pp. 43-49, 2012.

[14] D. Jetpipattanapong and R. Tanapat-tanadol, "A Comparative Impact Study of the Changing Number of Outputs in Artificial Neuron Network on Yom River Tide Forecasting, Phrae Province," *Journal of Environmental Management*, Vol. 6, No.2, pp. 35-53, 2010.

[15] Y. Chidthong and S. Supharatid, "Developing a hybrid multi model for peak flood forecasting," *Proceeding of 12^{th} National Convention on Civil Engineering*, pp. 403-409, 2007.

[16] T. Thongwan, A. Kangrang and S, Homwuttiwong, "An Estimation of Rainfall using Fuzzy Set-Genetic Algorithms Model," *American Journal of Engineering and Applied Sciences*, Vol. 4, No.1, pp. 77-81, 2011.

[17] A. K. Lohani, N. K. Goel and K. K. S. Bhatia, "Development of fuzzy logic based real time flood fore-casting system for river Narmada in central India," *Proceeding of Innovation Advances and Implementation of Flood Forecasting Technology*, pp. 1-10, 2005.

[18] T. Kohonen, "Self-Organizing Maps," *Proceedings of the IEEE*, pp. 1464-1480, 1990.

[19] F. J. Chang, L. C. Chang, and Y. S. Wang, "Enforced Self-organizing Map Neural Networks for River Flood Forecasting," *Hydrological Processes*, Vol. 21, No.6, pp. 741-749, 2007.

[20] E. Toth and A. Brath, "Flood forecasting using artificial neural networks in black-box and conceptual rainfall-runoff modelling," *Proceedings of 1^{st} Biennial Meeting of the iEMSs*, pp.166-171, 2002.

[21] L. Richard, "An introduction to computing with neural nets," *ASSP Magazine, IEEE*, Vol. 4, No.2, pp. 4-22, 1987.

[22] H. B. Hwarng and N. F. Hubele, "Back-propagation pattern recognizers for X control charts: methodology and performance,"*International Journal of Computers and Industrial Engineering*, Vol. 24, No.2, pp. 19-235, 1993.

[23] D. F. Lekkas, C. Onof, M. J. Lee, and E. A. Baltas, "Application of artificial neural networks for flood forecasting," *Global Nest: The International Journal*, Vol. 6, No.3, pp. 205-211, 2004.

[24] L. H. Feng and J. Lu, "The practical research on flood forecasting based on artificial neural networks," *Expert Systems with Applications journal*, Vol. 37, No.4, pp. 2974-2977, 2010.

[25] M. B. Kia, S. Pirasteh, B. Pradhan, A. R. Mahmud, W. N. A. Sulaiman and A. Moradi, "An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia," *International Journal of Environmental and Earth Sciences*, Vol. 67, No.1, pp. 251-264, 2012.

[26] P. Pramojanee, C. Tanavud, C. Yongchalermchai and C. Navanugraha, "An application of GIS for mapping of flood hazard and risk area in Nakorn Sri Thammarat Province, South of Thailand," *Proceedings of International Conference on GeoInformation for Sustainable Management*, pp. 17-21, 1997.

[27] C. Anavud, C. Yongchalermchai, A. Bennui and O. Densreeserekul, "Assessment of flood risk in Hat Yai municipality, Southern Thailand, using GIS," *Journal of Natural Disaster Science*, Vol. 26, No.1, pp. 1-14, 2004.

[28] S. Thudchai and K. Bhaktikul, "The Application of Hydraulic Model with Geographic Information System To Create Flood Risk Mapping In Mae Klong River," *Proceeding of 4^{th} INWEPF Steering Meeting and Symposium*, pp.1-9, 2007.

[29] R. R. Bouckaert, E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "WEKA—Experiences with a Java open-source project," *The Journal of Machine Learning Research*, Vol. 11, No.1, pp. 2533-2541, 2010.

[30] R. Kohavi, "A study of cross-validation and boot-strap for accuracy estimation and model selection," *Proceeding of International joint Conference on artificial intelligence*, pp. 1137-1145, 1995.

[31] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction," *Proceedings of 4^{th} ACM conference on Digital libraries*, pp. 254-255, 1999.

[32] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," *Proceedings of 1^{st} Workshop on High performance Data Mining*, pp. 556-560, 1998.

Supattra Puttinaovarat received her B.B.A.(Second Class Honors) in Business Com-puter and M.S. degree in Management of Information Tech-nology from Prince of Songkla University, Thailand in 2007 and 2010, respectively. She is currently a Ph.D. in Infor-mation Technology in the Institute of Social Technology at Su-ranaree University of Technology, Thailand. She is a lecturer at Faculty of Science and Industrial Technology, Prince of Songkla University, Suratthani Campus, Thailand. Her re-search interest includes Flood Modeling and Geographic In-formation System.

Paramate Horkaew received his B.Eng. (First Class Honors) in Telecommunication Engineering (1999) from King Mongkut's Institute of Technology, Ladkrabang, Thailand and Ph.D. in Computer Science from Imperial College London, University of London, London, United Kingdom (2004). He is now a lecturer at the School of Computer Engineering, Suranaree University of Technology, Thailand. His main research interests include Computational Anatomy, Digital Geometry Processing, Computer Vision and Graphics, and Evolution of Harmonic Maps on Riemannian Surfaces with Applications to Nonlinear PDE.

Kanit Khaimook received his Ph.D. in Educational Measurement and Evaluation (1991) and M.S. degree in Computer Science (1982) from Chulalongkorn University, Thailand. He is now a special lecturer at the School of Social Technology, Suranaree University of Technology, Thailand. His main research interests include Modeling and Geographic Information System.