HYDROLOGICAL PROCESSES
Hydrol. Process. **25**, 2573–2583 (2011)
Published online 10 May 2011 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.8132

An improved genetic programming to SSM/I estimation typhoon precipitation over ocean

Li Chen,[1]* Keh-Chia Yeh,[2] Hsiao-Ping Wei[3] and Gin-Rong Liu[4]

[1] *Department of Civil Engineering, Chung Hua University, Hsinchu 30012, Taiwan, ROC*
[2] *Department of Civil Engineering, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan, ROC*
[3] *National Science and Technology Center for Disaster Reduction, Taipei 32001, Taiwan, ROC*
[4] *Center for Space and Remote Sensing Research, National Central University, Tao-Yuan 32001, Taiwan, ROC*

Abstract:

This article proposes an improved multi-run genetic programming (GP) and applies it to estimate the typhoon rainfall over ocean using multi-variable meteorological satellite data. GP is a well-known evolutionary programming and data mining method used to automatically discover the complex relationships among nonlinear systems. The main advantage of GP is to optimize appropriate types of function and their associated coefficients simultaneously. However, the searching efficiency of traditional GP can be decreased by the complex structure of parse tree to represent the multiple input variables. This study processed an improvement to enhance escape ability from local optimums during the optimization procedure. We continuously run GP several times by replacing the terminal nodes at the next run with the best solution at the current run. The current method improves GP, obtaining a highly nonlinear meaningful equation to estimate the rainfall. In the case study, this improved GP (IGP) described above combined with special sensor microwave imager (SSM/I) seven channels was employed. These results are then verified with the data from four offshore rainfall stations located on islands around Taiwan. The results show that the IGP generates sophisticated and accurate multi-variable equation through two runs. The performance of IGP outperforms the traditional multiple linear regression, back-propagated network (BPN) and three empirical equations. Because the extremely high values of precipitation rate are quite few and the number of zero values (no rain) is very large, the underestimations of heavy rainfall are obvious. A simple genetic algorithm was therefore used to search for the optimal threshold value of SSM/I channels, detecting the data of no rain. The IGP with two runs, used to construct an appropriate mathematical function to estimate the precipitation, can obtain more favourable results from estimating extremely high values. Copyright © 2011 John Wiley & Sons, Ltd.

KEY WORDS genetic programming; meteorological satellite; SSM/I; evolutionary programming; data mining; back-propagated network (BPN)

Received 17 September 2010; Accepted 8 April 2011

INTRODUCTION

Taiwan is located at the centre of the western Pacific Rim and is particularly vulnerable to threat by typhoons. On average, there are 4·9 typhoons passing through Taiwan annually. Approximately 79% of these typhoons occur in the period from July to September (Wei *et al.*, 2006). Heavy rainfalls resulted from the typhoons cost human lives and financial damages in Taiwan every year. This is especially the case for stagnant typhoons over Taiwan area, which brings large-scale disasters in forms of floods and debris flows. For example, a recent ferocious typhoon Morakot passed through and ravaged Taiwan in August 2009. During its passage, more than 2215·5 mm of rainfall was recorded in southern Taiwan in 48 h (http://www.cwb.gov.tw/). Therefore, one of the most important topics in disaster prevention in whole world would be the accurate estimation of rainfall rate. Rainfall estimation using meteorological satellite data plays an especially important role in this topic.

Satellite rainfall retrieval can provide rainfall estimates more frequently and over a wider area than conventional raingauge measurements and can assist in detecting heavy rain events caused by typhoon (Li *et al.*, 2004). The special sensor microwave imager (SSM/I) has gained widespread applications in the ocean and atmosphere system in the past decades. It has been in operation since June 1987 (Hollinger, 1991), has the unique ability to penetrate through the cirrus clouds and senses the emitted and scattered radiation by raindrops and precipitation sized ice particles, respectively. Passive microwave retrieval can be grouped into two categories. The first is emission based, where liquid precipitation causes brightness temperature increases over a radiometrically cold (usually ocean) background. The second is scattering based, where precipitation, especially that above freezing level, causes brightness temperature decreases over a radiometrically warm (usually land) background. The emission method is the best known, based primarily on the work by Wilheit *et al.* (1997) at 19·25-GHz oceanic precipitation retrievals (Spencer

*Correspondence to: Li Chen, Department of Civil Engineering, Chung Hua University, Hsinchu 30012, Taiwan, ROC.
E-mail: lichen@chu.edu.tw

et al., 1988). Many algorithms applying the SSM/I data have been developed in estimating the precipitation (Liu *et al.*, 2008). The primary goal of the algorithms was to produce a reasonable relationship between the SSM/I's brightness temperature and rainfall.

Numerous algorithms applying the SSM/I data have been developed in estimating the precipitation. The formation of the various retrievals was based on the rain gauge data or cloud-resolving model simulations. Wilheit (1994) and Petty (1995) established various viable rainfall estimation algorithms. The second WetNet Precipitation Intercomparison Project (PIP-2) and PIP-3 were employed in comparing each SSM/I precipitation algorithm (Smith *et al.*, 1998; Adler *et al.*, 2001). They discovered that some of the algorithms performed satisfactorily, where their average bias was less than those of the rainfall approximation of weather radars. In these particular rainfall estimation algorithms, some utilized a single channel, such as the 19·35-GHz band (Chiu *et al.*, 1990), while others made use of all the SSM/I channels (Ferraro, 1997). The algorithms' primary goal was to produce a reasonable relationship between the SSM/I's brightness temperature and observed rainfall. In general, Ferraro's (1997) method is considered the most common approach in depicting the Tb and rainfall rate relationship, but it is still unclear whether it is a suitable tool for a typhoon's rainfall estimation. According to Huang (2000), among the various algorithms, Chiu *et al.*'s (1990) method owned a higher accuracy in the comparison of the rain gauge data in the northwest Pacific Ocean.

An alternative method in recent years, artificial neural networks (ANNs) are implemented with standardized black-box packages, so they are easier to use; and the performances of ANNs are usually better than those of traditional statistical methods. Moreover, they are highly nonlinear and can capture complex interactions among input/output variables in a system without any prior knowledge about the nature of these interactions. It is well known that a multilayer feedforward neural network, having at least one hidden layer, can approximate most nonlinear function relating inputs to outputs. The main advantage of ANNs is that one does not have to explicitly assume a model form, which is a prerequisite in the parametric approach. In the field of rainfall forecasting, many studies have been performed using an ANN approach with different remote sensing data such as satellite data (Hsu *et al.*, 1997; Sorooshian *et al.*, 2000; Kuligowski and Barros, 2001; Grimes *et al.*, 2003) and radar data (Mimikou and Baltas, 1996; Bellerby *et al.*, 2000; Trafalis *et al.*, 2002). All these studies have reported an improvement in performance using ANNs (Chiang *et al.*, 2007). For microwave, Staelin *et al.* (1999) used an ANN to estimate precipitation from 183-GHz humidity channels of the advanced microwave sounding unit (AMSU) on satellite NOAA-15. Hsu *et al.* (1999) demonstrated the utility of using ANN to generate functions linking infrared and visible image characteristics (including image texture characteristics) to precipitation. Krasnopolsky *et al.* (1999) developed an ANN to retrieve sea surface temperature, water vapour, liquid water, and wind speed over the ocean simultaneously. Xia (2001) used a similar method to retrieve sea surface temperature, wind speed, relative wind direction, and water vapour from SSM/I data. Recently, Chen and Staelin (2003) applied ANNs to estimate the precipitation rate based on brightness temperature data from the atmospheric infrared sounder (AIRS)/advanced microwave sounding unit (AMSU)/Humidity Sounder for Brazil sensors on board the AQUA satellite. However, these 'black box' models are unable to generate explicit formulas that can explain the essence of the precipitation mechanism.

Evolutionary computation techniques, which are based on a powerful principle of evolution: survival of the fittest, are very efficient optimization methods. Among these methods, genetic algorithm (GA) is one of the most popular search algorithms. But there are some kinds of difficulties of GA, such as fixed-length encoding and premature convergence. On the other hand, researchers have successfully used evolutionary algorithms for automatically generating programs or equations connecting the inputs and outputs. The genetic programming (GP) operates a population of the chromosome (a string of input variables, constants and mathematical operators) expressed as dynamic tree, which is more flexible than fixed-length data structure of GA. A great number of previous studies applied GP to their fields. However, it seems that not many efforts have been made to the applications to hydrologic estimation and water resources engineering (Omolbani *et al.*, 2010). Babovic (1996) introduced the GP paradigm in the area of water engineering first soon after Koza, who first proposed GP in 1992. Cousin and Savic (1997), Savic *et al.* (1999), Drecourt (1999), Whigham and Crapper (1999, 2001), and Babovic and Keijzer (2002) applied GP to rainfall–runoff modelling. Dorado *et al.* (2003) studied on prediction and modelling of the rainfall–runoff transformation of a typical urban basin using ANNs and GP.

Chen (2003) used a GP to evaluate the water quality in a reservoir through remote sensed imageries. The results show that the presented method can obtain satisfied accuracies for estimation. However, the estimation of precipitation through SSM/I microwave frequency channels is more difficult. Chen (2003) pointed out 'The traditional GP like general GA usually suffers from the problem of premature convergence, which cannot acquire a satisfactory solution'. This problem is addressed in the current study by enhancing the ability of the algorithm to escape from a local optimal solution. Therefore, in the following section, we will firstly begin with an introduction of the GP algorithm and a discussion concerning the improvement. Then, a case study of the precipitation at sea surface model is demonstrated in the later section. The results of this improved GP (IGP) are compared with those of traditional regression, several empirical equations, back-propagated network (BPN) and three empirical equations. Finally, we present the conclusions and some closing remarks.

GENETIC PROGRAMMING

GP is a conceptual model for system identification problems and it can acquire much information on the detail of insight relationships between input and output data. In the case study presented in this article, all the seven input variables and corresponding output variable are continuous. Therefore, it is suitable to use the GP to estimate the output values directly. Unlike the regression techniques, GP automates the trial and error process of system identification and can be used to build a model structure that best fits training data. GP works by emulating natural evolution to generate a model structure that maximizes (or minimizes) objective function involving an appropriate measure of the level of agreement between the model and system responses (Koza, 1992).

This model allows us to gain additional information on how the system performs, i.e. gives an insight into the relationship between input and output data. GP builds on methods derived from the GA (Goldberg, 1989). GP expresses the hierarchical computer programs as parse trees, rather than as the binary strings usually used by GA. This algorithm begins with the random generation of the dynamic parse tree of each individual for the initial population. Then, the chromosomes are expressed as mathematical equations and the fitness of each equation is evaluated by the errors between actual and estimating data. The individuals are then selected according to fitness to reproduce the offspring. The process is repeated for a certain number of generations or until the criterion for termination has been achieved. GP has a distribution-free advantage, i.e. no prior knowledge is needed about the statistical distribution of the data (Kishore *et al.*, 2000).

There are five major preparatory steps in using GP for a particular problem. These five steps involve determining (Chen, 2003)

1. The set of terminals consisting the variables and constants (a real number determined automatically by GP) of the program.
2. The set of primitive functions consists of the basic mathematical functions and other more complex user-defined functions.
3. The crossover, mutation operators and selection operator are the most important part of GP.
4. The parameters for controlling the run including the population size, crossover rate, mutation rate, etc.
5. The criterion for terminating a run generally is set by a predefined number of generations, the amount of variation of individuals between different generations or a target value of fitness.

Representation schemes

GP uses parse trees instead of lines of code to represent programs. Thus, for example, the simple algebraic expression $zy(y + 0{\cdot}639z)$ would be represented as the tree in Figure 1. The 'root node' is the first element of the tree, the 'interior nodes' (non-terminal nodes) are the functions and the 'leaf nodes' (terminal nodes) are the constants and/or the variables. The coding of a chromosome is a hierarchical structure, which consists of two layers. The first layer is to determine the type of terminal, non-terminal nodes. The second layer gives the actual value of variable number or the constant value.

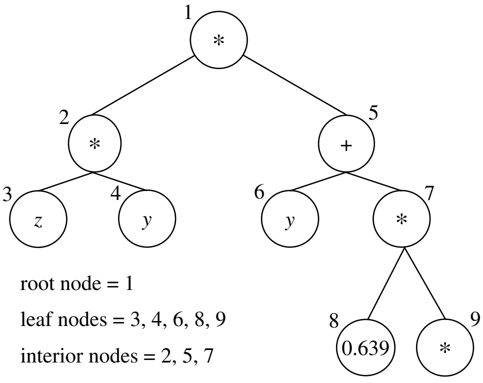

Figure 1. Parse tree representation of an expression

The three genetic operators of GP are described as follows:

(1) Reproduction (selection)

Reproduction is a process in which individual trees are set according to their fitness function values. The reproduction operator may be implemented in algorithmic form in a number of ways, such as proportional, rank, and tournament selection. A macro-evolutionary algorithm is used as a selection scheme in this article which was described as Chen (2003) to maintain diversity.

(2) Crossover (recombination)

After reproduction, the algorithm uses a crossover operator that exchanges arbitrary sub-trees between two individuals with probability Pc. The crossover operator used in GP must ensure that programs obey the syntax of the representation scheme. So it creates new offspring that consists of genetic material taken from the parents. Figure 2 shows how this operator works.

(3) Mutation

The mutation of GP simply consists of randomly exchanging a node in the tree with another node or a sub-tree.

Fitness function

The correlation coefficient (CC) between estimated and actual values is adopted as the fitness function of GP. Through several experiments, it is observed that this fitness function can accelerate the speed of search procedure compared with using the root mean squared error (RMSE) directly. It is able to achieve both 'high linear correlation' and 'small RMSE' simultaneous in

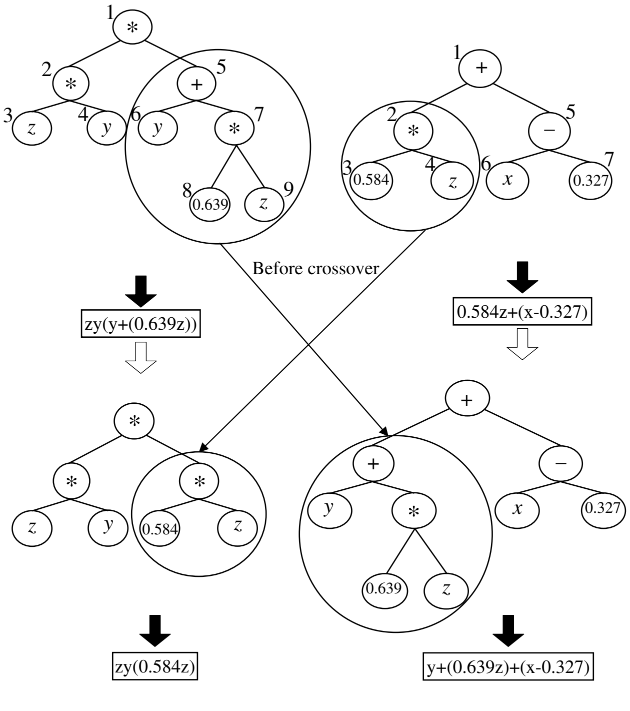

Figure 2. Crossover scheme of two parse trees

most cases, so we chose the former as the objective function. This study therefore employed single linear regression analysis to decrease the RMSE of estimation:

$$y = \alpha + \beta \cdot f \tag{1}$$

where f is the output value of data estimated by the GP, y is the actual output value in the dataset, and α and β are the regression coefficients.

According to the single linear regression analysis, the two regression coefficient can be estimated as follows:

$$\alpha = \overline{y} - \beta \cdot \overline{f} \tag{2}$$

$$\beta = \frac{\sum_{i=1}^{n} (f_i - \overline{f}) \times (y_i - \overline{y})}{\sum_{i=1}^{n} (f_i - \overline{f})^2} \tag{3}$$

where $\overline{y}$ is the mean of the actual output values in the dataset, $\overline{f}$ the mean of the output values estimated by GP, y_i the actual output value of the ith data in the dataset, and f_i the output value of the ith data estimated by GP.

Improved GP

One of the main drawbacks in the conventional GP is likely to be trapped in a region that does not contain the global optimum. This problem, called premature convergence, has been recognized as a serious failure mode for almost all the optimization models. This study improves GP to maintain the best result in the current run and then places it on the terminal nodes of the parse tree to run the GP again. The best result automatically includes the original one in the next run, so the equation is more sophisticated and accurate. To avoid a very complicated form of equation, this learning procedure continues several times until achieving a satisfactory result.

RAINGAUGE AND SATELLITE DATA

The data adopted in this article were collected by the SSM/I under the US Defense Meteorological Satellites Program (DMSP) satellites, along with offshore island raingauge data. SSM/I are sun-synchronous satellites, which orbit the earth at a height of 833 km and are oriented with an inclination of 98·8° (Hollinger *et al.*, 1990). The SSM/I are seven-channel passive microwave radiometers scanning in the dual-polarized (vertical and horizontal) channels at 19·35, 37·0, and 85·5 GHz and vertical-polarized channel at 22·2 GHz (Hollinger, 1989, 1991). This study utilizes the microwave data (Tb_{19}, Tb_{22}, Tb_{37}, and Tb_{85}) observed by DMSP F-13, F-14, and F-15 satellites mentioned above with respective seven channels.

Rainfall rates estimated by the SSM/I were compared with the hourly raingauge data of Taiwan's offshore islands, including Peng-Jia-Yu, Don-Gji-Dao, Lan-Yu, and Green Island, as shown in Figure 3. The area of interest locates around 115°–135 °E longitude and 10°–30°N latitude, which covers all the possible typhoon trajectories that could influence Taiwan. The typhoon rainfall data were used during 2000 to 2004. In this study, a total of 34 typhoon events with 1396 data were collected during this time period, as shown in Table I. The brightness temperature measurements from satellite-borne microwave radiometers and the rainfall measurements from the four offshore raingauges are coupled. Since the rain rate from ground rain station is just one point measurement compared with the area measurement from satellite, it is well known that the non-uniform beam filling is a major error source when a comparison between satellite data and ground rain station is made. The general beam filling error correction schemes include both homogeneous radioactive transfer calculation and based on cloud simulations with field of view (FOV)-average rain and brightness temperatures databases (Wei *et al.*, 2011).

All data were arbitrarily grouped in two sets, called the training (calibration) set and the testing (validation) set, which have roughly the same statistical properties (mean and variance). When the training process had been completed, the constructed model was used to estimate the output values for the data in the testing set (which the process had never seen during the training stage). Therefore, the use of these SSM/I data in the learning by IGP depends on splitting the 1396 records into two groups:

1. the first group is used for training the model and is called the training set including 900 data (22 events) and

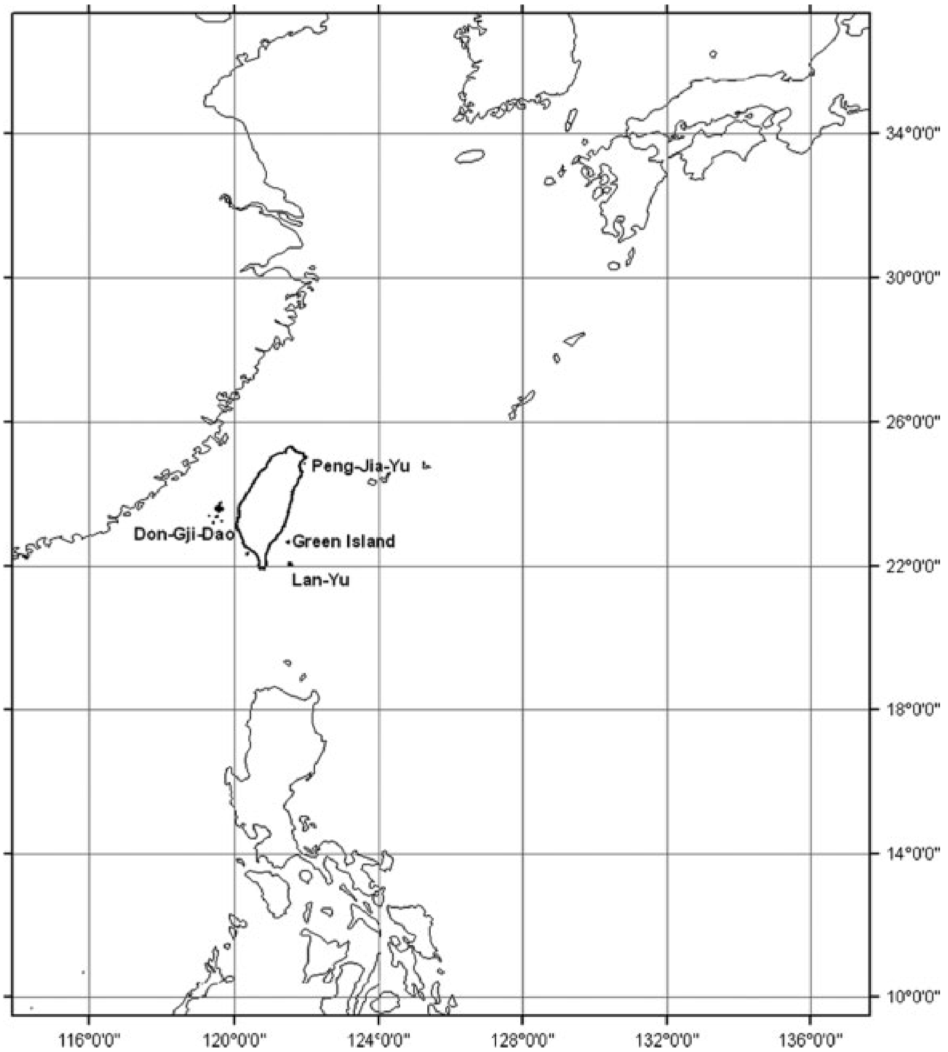

Figure 3. Location of the four islands around Taiwan

2. the second group is used to measure the performance of the model and is called the testing set including 496 data (12 events).

RESULTS

Improved GP

The IGP was applied to precipitation estimation. The input variables (terminals) involved seven channels of SSM/I, Tb_{19v}, Tb_{19h}, Tb_{22v}, Tb_{37v}, Tb_{37h}, Tb_{85v}, and Tb_{85h}, and operators (internal nodes), which apply several sets of mathematical operators such as $\{+, -, \times, /,$ LN, EXP, POWER$\}$, were used in the experiments. In this study, the GP model was conducted twice. With the population size (individuals or parse trees) equal to 400 and through 8000 generations, the final optimal equations obtained from the IGP are shown as Equations (4) and (5). Results of these two parse trees of the precipitation models at sea surface are shown in Figures 4 and 5, and the corresponding mathematical formulas are listed as follows.

The result of the first run:

$$RR'_{IGP} = -1{\cdot}9884 - \frac{0{\cdot}0022 \times Tb_{37h} \times Tb_{85h}}{Tb_{19h} - Tb_{19v}} \tag{4}$$

The result of the second run:

$$RR''_{IGP} = 0{\cdot}0233 + 0{\cdot}0035 \times \left[RR'_{IGP} \times Tb_{22v} + 0{\cdot}0654 + \frac{99{\cdot}76}{Tb_{37h} - Tb_{19v}}\right] \tag{5}$$

where RR'_{IGP} and RR''_{IGP} represent the retrieved rainfall rate (mm/h); $Tb_{()}$ represents the footprint brightness temperature; the unit in the brackets is the absolute temperature in Kelvin; and subscripts v and h represent the vertical and horizontal polarized channels, respectively.

This study employed single linear regression analysis to decrease the RMSE of estimation using Equation (1): $\alpha = -1{\cdot}9884$, $\beta = -0{\cdot}0022$, and $f = \frac{Tb_{37h} \times Tb_{85h}}{Tb_{19h} - Tb_{19v}}$ for Equation (4).

$\alpha = 0{\cdot}0233$, $\beta = 0{\cdot}0035$, and $f = RR'_{IGP} \times Tb_{22v} + 0{\cdot}0654 + \frac{99{\cdot}76}{Tb_{37h} - Tb_{19v}}$ for Equation (5).

Table I. Typhoon data collected in this study

Year	Typhoon	Duration	Number	Mean	Variance
2000	KAI-TAK	07/06∼07/10	40	0·58	1·98
	BILIS	08/21∼08/23	32	1·62	29·41
	PRAPIROON	08/27∼08/30	68	0·03	0·02
	YAGI	10/23∼10/26	10	0·0	0·0
	XANGSANE	10/30∼11/01	28	2·39	21·86
	BEBINCA	11/06∼11/07	10	0·75	2·75
2001	TRAMI	07/10∼07/11	54	2·13	42·5
	YUTU	07/23∼07/24	42	0·14	0·17
	TORAJI	07/28∼07/31	66	0·08	0·11
	NARI	09/08∼09/19	214	1·02	17·82
	LEKIMA	09/23∼09/28	58	2·54	17·66
	HAIYAN	10/15∼10/16	42	0·02	0·0
2002	RAMMASUN	07/02∼07/04	30	0·0	0·0
	NAKRI	07/09∼07/10	16	0·8	1·13
	SINLAKU	09/04∼09/08	60	017	0·30
2003	KUJIRA	04/21∼04/24	82	0·30	1·31
	NANGKA	06/01∼06/03	56	0·15	0·44
	SOUDELOR	06/16∼06/18	50	0·29	0·84
	IMBUDO	07/21∼07/23	32	0·5	3·13
	MORAKOT	08/02∼08/04	52	1·14	5·16
	VAMCO	08/19∼08/20	26	2·49	14·6
	KROVANH	08/22∼08/23	42	0·15	0·2
	DUJUAN	08/31∼09/02	36	1·53	11·6
	MELOR	11/02∼11/03	38	0·36	0·79
2004	CONSON	06/07∼06/09	44	0·59	2·37
	MINDULLE	06/28∼07/03	44	0·19	0·46
	KOMPASU	07/14∼07/15	2	0	0
	RANANIM	08/10∼08/13	17	0	0
	AERE	08/23∼08/26	36	1·52	21·83
	HAIMA	09/11∼09/13	19	0·5	2·17
	MEARI	09/26∼09/27	17	0	0
	NOCK-TEN	10/23∼10/26	22	0·07	0·10
	NANMADOL	12/03∼12/04	11	1·66	7·30

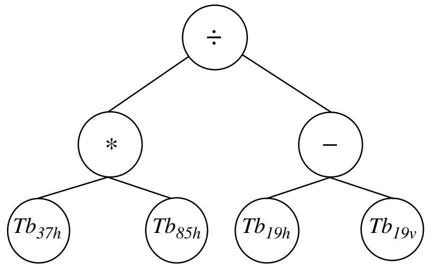

Figure 4. The parse tree of Equation (4)

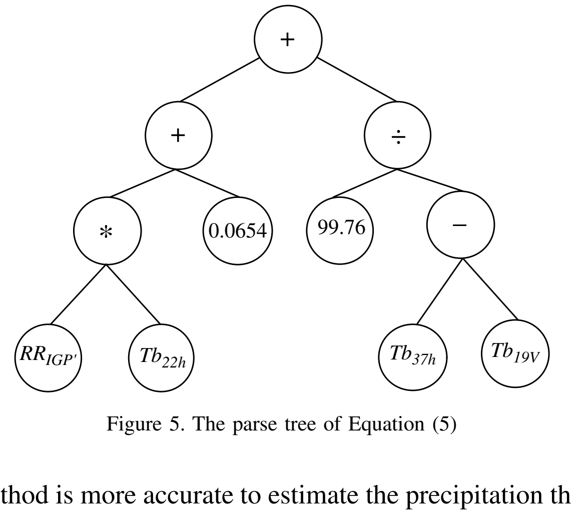

Figure 5. The parse tree of Equation (5)

Two constants −1·9884 and 0·0654 shown in Equation (5) were chosen automatically via IGP to optimize the defined objective function.

Four input variables are presented at the first run: $\mathrm{Tb}_{19\mathrm{h}}$, $\mathrm{Tb}_{19\mathrm{v}}$, $\mathrm{Tb}_{37\mathrm{h}}$, and $\mathrm{Tb}_{85\mathrm{h}}$ as shown in Equation (4). Then, $\mathrm{Tb}_{22\mathrm{v}}$ is included in Equation (5) at the second run of IGP. Table II shows the RMSEs of these two runs at both training and testing stages. Findings show that the RMSE equals 1·92 and 2·06 at the training and testing stage, respectively, for the first run. These represent the results of conventional GP without improvement. At the second run, the RMSE decreases to 1·77 and 1·85 for the training and testing set, respectively. The values of RMSEs decrease for both stages, indicating improved estimation accuracies. It can be concluded that the IGP method is more accurate to estimate the precipitation than the traditional one by 8%. The forms of these equations will become increasingly complicated run by run. When it reaches the third run, the RMSE decreases to 1·75 for the training set. This procedure stops at the second run, because the RMSE equal to 1·77 is considered convergent. Besides, no additional new input variable is included during the third run. Therefore, the procedure terminates at the second run, which is reasonable.

Figures 6 and 7 show the scatter diagrams of estimated values *versus* actual values of precipitation for the first

Table II. The RMSEs of all models

Model	Training	Testing
RR'_{IGP} (first run)	1·92	2·06
RR''_{IGP} (second run)	1·77	1·85
Chiu *et al.* (1990)	2·49	2·46
Ferraro *et al.* (1994)	2·23	2·16
Ferraro (1997)	2·71	2·47
RR_{RA} (regression)	2·11	2·10
BPN	1·85	2·08
$RR'_{SGA+IGP}$ (first run)	1·75	1·91
$RR''_{SGA+IGP}$ (second run)	1·56*	1·77*
SGA+RA	2·02	2·04
SGA+BPN	1·65	1·86

Note: the symbol
* represents the best result of these methods.

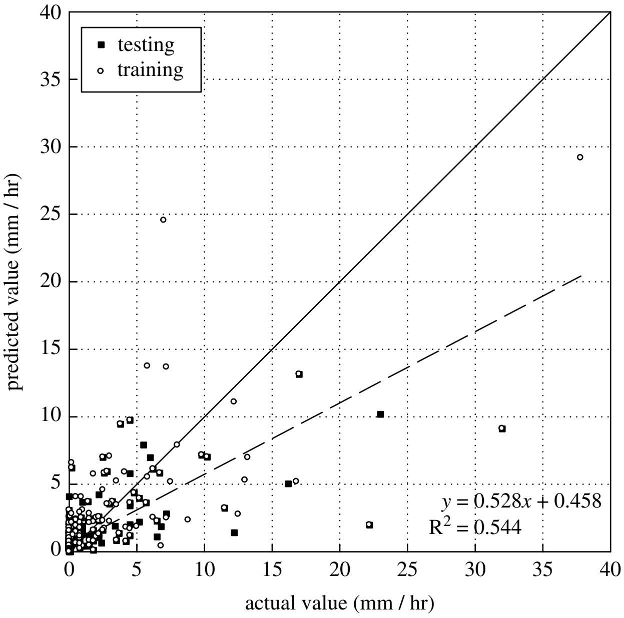

Figure 6. The scatter plot for the first run (RR'_{IGP})

and second run, respectively. One can tell that the average of estimated extreme high values of the second run shown in Figure 7 is closer to the ideal line than that of the first run shown in Figure 6. The accuracies of extreme high precipitation estimations are very important to disaster prevention during the typhoon periods. Therefore, the RMSE of IGP with second run is lower than that of traditional GP with first run indicating that the IGP improves the capabilities to prevent the serious disasters caused by typhoons.

Comparison with empirical equations

The results were compared with those obtained from the rainfall formulas developed by Chiu *et al.* (1990), Ferraro *et al.* (1994), and Ferraro (1997). These equations are shown as below.

(a) Chiu *et al.* (1990)

$$RR_{Chiu} = 5{\cdot}26 \times \log\left(\frac{102}{274 - Tb_{19h}}\right) \quad (7)$$

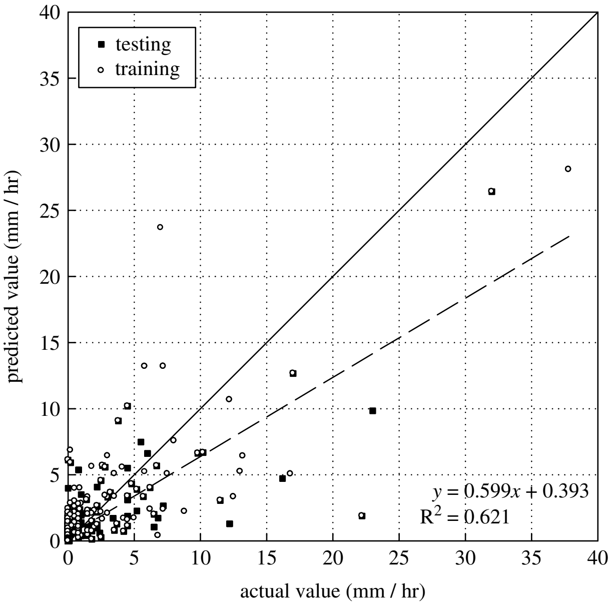

Figure 7. The scatter plot for the second run (RR''_{IGP})

Rain is detected if $Tb_{19h} < 274$.
where RR_{Chiu} represents the retrieved rainfall rate (mm/h).

(b) Ferraro *et al.* (1994)

$$RR_{Ferraro} = (Tb_{19h} + Tb_{19v} + Tb_{37h} - Tb_{22v} - Tb_{37v} - Tb_{85h} + 170{\cdot}2)/18{\cdot}3 \quad (8)$$

Rain is detected if $Tb_{19v} - Tb_{19h} < 60$.
where RR_{Chiu} represents the retrieved rainfall rate (mm/h).

(c) Ferraro (1997)

$$Q_{37} = -1{\cdot}15[\ln(290 - Tb_{37v}) - 2{\cdot}99 - 0{\cdot}32\ln(290 - Tb_{22v})] \quad (9)$$

$$RR_{Ferraro} = 0{\cdot}007107(100Q_{37})^{1{\cdot}7359} \quad (10)$$

where Q_{37} represents the liquid water estimates based on the 22- and 37-GHz channels, respectively; v represents the vertical and horizontal polarized channel; and $RR_{Ferraro}$ represents the retrieved rainfall rate (mm/h); rain is detected if $Q_{37} > 0{\cdot}20$. The minimum retrieved rainfall rate is 0·30 mm/h, while any value greater than 35 mm/h is set to 35 mm/h.

The RMSEs of these three empirical methods are also shown in Table II. It shows that the testing results of Ferraro *et al.* (1994) have the lowest RMSE (2·16) among these three empirical equations, but much higher than the IGP, regressive analysis (RA), and BPN.

Comparison with multiple linear RA

In the conventional statistical modelling process, RA is a popular tool for building a model. Because the proper form of these functions is unknown, this study considers only the simplest linear type. The coefficients in this regression equation were determined through the basic least squares method:

$$\begin{aligned}RR_{RA} &= -6{\cdot}9593 + 0{\cdot}2788 \times Tb_{19v} + 0{\cdot}0994 \\ &\times Tb_{19h} - 0{\cdot}2498 \times Tb_{22v} \\ &+ 0{\cdot}0952 \times Tb_{37v} - 0{\cdot}1389 \times Tb_{37h} - 0{\cdot}0547 \\ &\times Tb_{85v} + 0{\cdot}0438 \times Tb_{85h} \qquad (11)\end{aligned}$$

where RR_{RA} represents the retrieved rainfall rate (mm/h).

The RMSEs of RA are shown in Table II. It shows that the RMSEs equal 2·11 and 2·10 at the training and testing stages, respectively, which are worse than IGP and BPN.

Comparison with back-propagation network

The ANN with back-propagation algorithm, called BPN, might be one of the most widely used models for estimation. In scaling, the range of each variable is assigned between 0·1 and 0·9. A general sigmoid function is adopted as the activation function in this case study. The same data were selected for use in the training and testing stages to compare the performance of IGP with that of BPN. In this study, the settings of the hidden layers and hidden nodes are determined after a number of trials. The procedure used two hidden layers with two and four nodes in each layer for training single and four variable models, respectively. The training procedures were terminated after 1000 iterations for both models. The RMSE equals 1·85 and 2·08 for the training and testing set, respectively, which is better than the first run of IGP but worse than the second run of IGP. The scatter diagram of the estimated values and the actual values of training data and testing data are shown in Figure 8. It shows that BPN underestimates at the high end, but it overestimates at the mid-range. In addition, the linear CCs of Figures 6–8 are 0·74, 0·79, and 0·75, respectively. It also indicates that the performances of IGP with the second run are the best compared with those of IGP with the first run and BPN. Because there are very few peak rainfalls in a typhoon event with limited observed high rainfall data to model and train the IGP, all models tended to fit low values, which explains why the predicted heavy rainfalls were underestimated.

Detecting rainfall thresholds using simple GA

The variation of rainfall intensity for each typhoon event is extremely wide, as shown in Table I. A histogram of precipitation rates for the four rain gauges is shown in Figure 9, showing that the extremely high values of precipitation rate are quite few. In contrast, the number of zero values (no rain) is greater than one thousand. Therefore, a GA was used to search for the optimal threshold values of the bands, detecting the data of zero values (no rain). The IGP was then used to construct an appropriate mathematical function to estimate the precipitation using SSM/I.

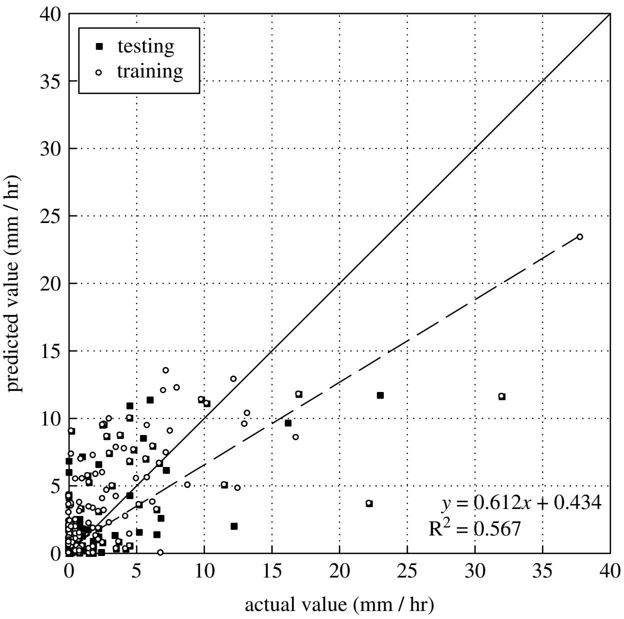

Figure 8. The scatter plot for BPN

A simple GA (SGA; Goldberg, 1989) was used to search for the thresholds of the seven channels to detect rainfall. The optimal thresholds were determined to be 185·7 for Tb_{19h}. In other words, when Tb_{19h} was less than 185·7 (°K), it detected 'No Rain'. The classification results of the SGA were demonstrated as a confusion matrix, as shown in Table III. The overall predicting accuracy equals 89·9%. Table III also indicates producer's accuracies are 92·7% (for 'No Rain') and 79·3% (for 'Rain') and user's accuracies are 94·4% (for 'No Rain') and 74·1% (for 'Rain'). There are two types of errors: omission error = 7·3% (actual value is rain, but estimating as no rain) and commission error = 5·5% (actual value is no rain, but estimating as rain). The commission error can be modified in the next step via the IGP. Although the omission error cannot be further corrected via the IGP, the error of rainfall prediction is relatively small.

The same procedures to train the IGP using 188 data in which SGA was detected as 'Rain', the obtained equations for two runs are shown as Equations (12) and (13). Results of these two parse trees of the precipitation models at sea surface are shown in Figures 10 and 11:

$$RR'_{SGA+IGP} = 9{\cdot}73 + 8{\cdot}39 \times \left(\frac{1}{270 - Tb_{19h}} - \frac{270 - Tb_{19h}}{74{\cdot}88}\right) \qquad (12)$$

$$RR''_{SGA+IGP} = 0{\cdot}122 + 0{\cdot}0694 \times \left[RR'_{SGA+IGP} \times \left(\frac{Tb_{19v}}{19{\cdot}187}\right) + \frac{LN(Tb_{19v})}{Tb_{37h} - Tb_{19v}}\right] \qquad (13)$$

$\alpha = 9{\cdot}73$, $\beta = 8{\cdot}39$, and $f = \frac{1}{270 - Tb_{19h}} - \frac{270 - Tb_{19h}}{74{\cdot}88}$ for Equation (12).

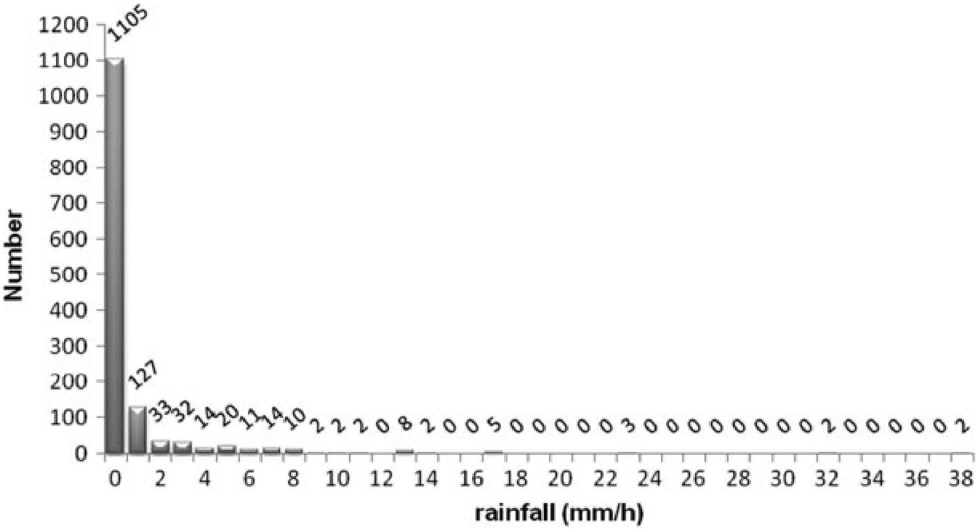

Figure 9. Histogram of rain rates derived from the four raingauges

Table III. The confusion matrix of SGA

Actual Value SGA	No Rain	Rain	Row Total	User's Accuracy%
No Rain	660	39	699	94.4
Rain	52	149	201	74.1
Column Total	712	188	900	
Producer's Accuracy%	92·7	79·3		89.9

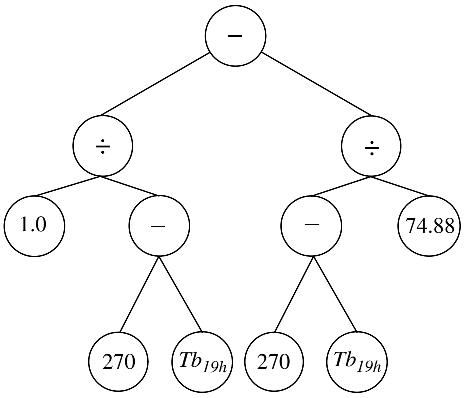

Figure 10. The parse tree of Equation (12)

$\alpha = 0{\cdot}122$, $\beta = 0{\cdot}0694$, and $f = \mathrm{RR}'_{\mathrm{SGA+IGP}} \times \left(\frac{\mathrm{Tb}_{19\mathrm{v}}}{19{\cdot}187}\right) + \frac{\mathrm{LN}(\mathrm{Tb}_{19\mathrm{v}})}{\mathrm{Tb}_{37\mathrm{h}} - \mathrm{Tb}_{19\mathrm{v}}}$ for Equation (13).

Only $\mathrm{Tb}_{19\mathrm{h}}$ was chosen as the most significant input variable to estimate rainfall using SSM/I shown in Equation (12). Then after the second run, $\mathrm{Tb}_{19\mathrm{v}}$ and $\mathrm{Tb}_{37\mathrm{h}}$ were included in Equation (13). The total number of input variables used to estimate rainfall decreased from an original five to three because of the SGA. The total training RMSEs of 900 data for Equations (12) and (13) equal 1·75 and 1·56, respectively, revealing that detecting rainfall thresholds using the SGA is beneficial in modelling for rainfall prediction using the IGP.

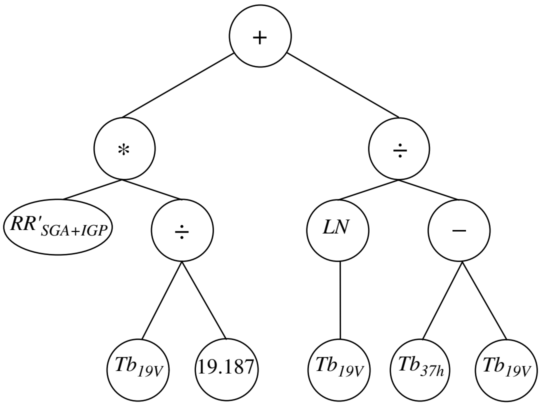

Figure 11. The parse tree of Equation (13)

The SGA combined with RA (SGA + RA) is shown as Equation (14):

$$\begin{aligned}\mathrm{RR}_{\mathrm{SGA+RA}} &= 4{\cdot}426 + 0{\cdot}1573 \times \mathrm{Tb}_{19\mathrm{v}} + 0{\cdot}2424 \\ &\quad \times \mathrm{Tb}_{19\mathrm{h}} - 0{\cdot}2348 \times \mathrm{Tb}_{22\mathrm{v}} \\ &\quad + 0{\cdot}2245 \times \mathrm{Tb}_{37\mathrm{v}} - 0{\cdot}2946 \times \mathrm{Tb}_{37\mathrm{h}} - 0{\cdot}5512 \\ &\quad \times \mathrm{Tb}_{85\mathrm{v}} + 0{\cdot}5002 \times \mathrm{Tb}_{85\mathrm{h}} \qquad (14)\end{aligned}$$

The same neural parameters of SGA combined with BPN (SGA + BPN) were set as the BPN without SGA. All the results of models with SGA are also summarized in Table II. The most crucial result was that the RMSEs of SGA + IGP for the second run were more favourable than those of the other models.

The scatter diagrams of SGA + IGP for the first and second runs are depicted and compared in Figures 12 and 13, respectively. Figure 12 shows that the low and middle values are more accurate using SGA + IGP for the first run than those shown in Figures 6 and 7 (which are without the SGA), but the high values are obviously underestimated. The predicted values of SGA + IGP for the second run are much closer to the ideal 45° line even for the extremely high values shown in Figure 13,

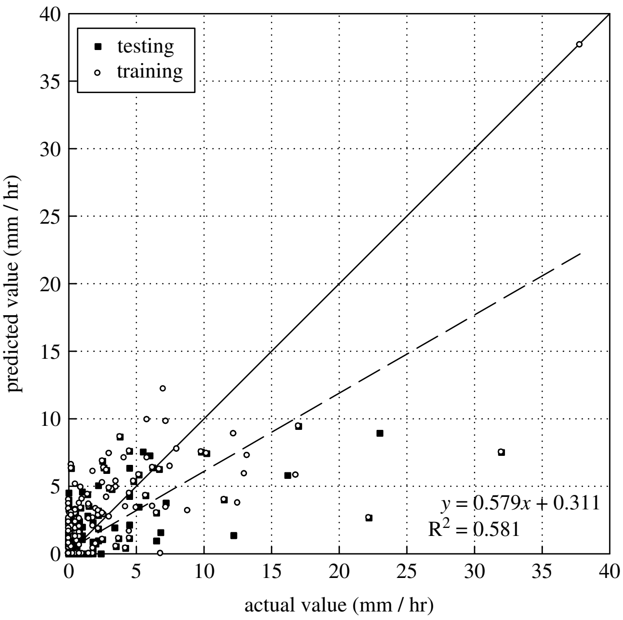

Figure 12. The scatter plot for the first run (SGA + IGP)

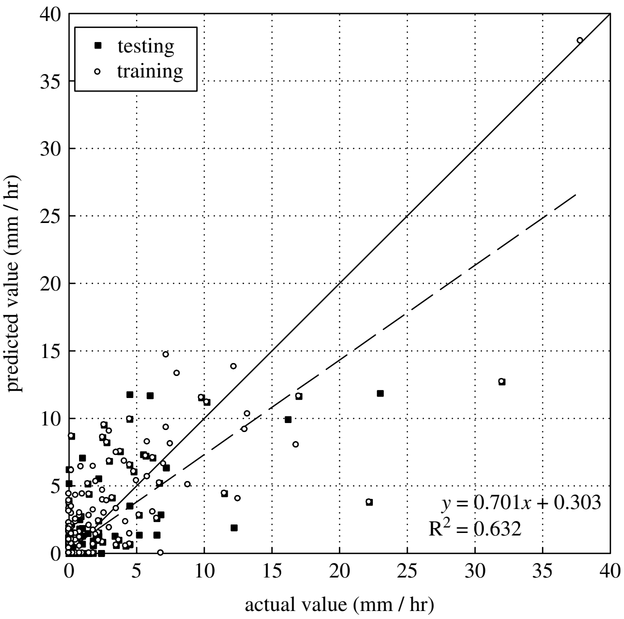

Figure 14. The scatter plot for SGA + BPN

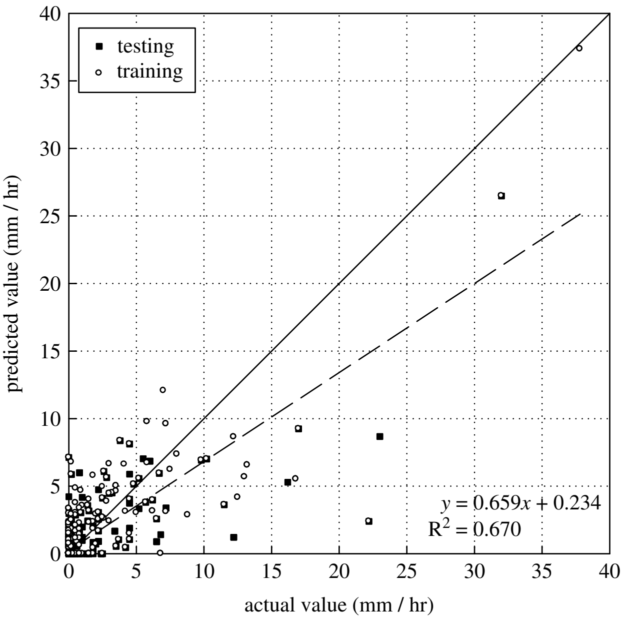

Figure 13. The scatter plot for the second run (SGA + IGP)

revealing that the model obtained by SGA + IGP for the second run more accurately predicts rainfall. The scatter diagram of the SGA + BPN is shown in Figure 14, and the estimating accuracy of SGA + BPN is between those of the SGA + IGP for the first and the second runs. The CCs of Figures 12–14 are 0·76, 0·82, and 0·79, respectively.

CONCLUSIONS

This article demonstrates the possibilities of adopting an IGP to estimate the precipitation at sea surface by meteorological satellite data. This model is more convenient and efficient to use for numerical expression to review the effects of each channel of SSM/I on the precipitation during typhoon periods. The hourly raingauge data in Taiwan's offshore islands along with the rainfall rates estimated by the SSM/I were collected from 34 typhoons that occurred during 2000 to 2004. The IGP deals easily with highly nonlinear problems via running the traditional GP two times to achieve higher accuracy in the case study. Five channels of SSM/I including Tb_{19h}, Tb_{19v}, Tb_{37h}, Tb_{85h}, and Tb_{22v} were used to estimate the precipitation. The results demonstrate that the IGP presented in this article is an appropriate system identified model compared with the traditional statistical regression. In addition, the performance of the second run of IGP was also found with lower RMSEs than those of the BPN and three empirical equations developed by Chiu *et al.* (1990), Ferraro *et al.* (1994), and Ferraro (1997). In addition, a SGA combined with the IGP improves the estimating accuracies for all models. Among those, SGA + IGP with two runs outperforms all other models. The optimal thresholds were determined to be 185·7 (°K) for Tb_{19h}, and only three channels of SSM/I including Tb_{19h}, Tb_{19v}, and Tb_{37h} were chosen after two runs of SGA + IGP.

REFERENCES

Adler RF, Kidd C, Petty G, Morissey M, Goodman HM. 2001. Intercomparison of global precipitation products: the third precipitation intercomparison project (PIP-3). *Bulletin of the American Meteorological Society* **82**: 1377–1396.

Babovic V. 1996. *Emergence, Evolution, Intelligence: Hydroinformatics*. Balkema Publishers: Rotterdam.

Babovic V, Keijzer M. 2002. Declarative and preferential bias in GP-based scientific discovery. *Genetic Programming and Evolvable Machines* **3**(1): 41–79.

Bellerby T, Todd M, Kniveton D, Kidd C. 2000. Rainfall estimation from a combination of TRMM precipitation radar and GOES multispectral satellite imagery through the use of an artificial neural network. *Journal of Applied Meteorology* **39**: 2115–2128.

Chen FW, Staelin DH. 2003. AIRS/AMSU/HSB precipitation estimates. *IEEE Transactions on Geoscience and Remote Sensing* **41**: 410–417. DOI: 10.1109/TGRS.2002.808322.

Chen L. 2003. A study of applying genetic programming to reservoir trophic state evaluation using remote sensor data. *International Journal of Remote Sensing* **24**: 2265–2275. DOI: 10.1080/01431160210154966.

Chiang YM, Chang FJ, Jou Ben J-D, Lin PF. 2007. Dynamic ANN for precipitation estimation and forecasting from radar observations. *Journal of Hydrology* **334**: 250–261.

Chiu LS, North GR, Short DA, McConnell A. 1990. Rain estimation from satellites: effect of finite field of view. *Journal of Geophysical Research* **95**: 2177–2185.

Cousin N, Savic DA. 1997. *A rainfall-runoff model using genetic programming*. *Centre for Systems and Control Engineering, Report* No. 97/03, School of Engineering, University of Exeter: Exeter, United Kingdom; 70.

Dorado J, Rabunal JR, Pazos A, Rivero D, Santos A, Puertas J. 2003. Prediction and modelling of the rainfall-runoff transformation of a typical urban basin using ANN and GP. *Applied Artificial Intelligence* **17**: 329–343.

Drecourt JP. 1999. Application of neural networks and genetic programming to rainfall-runoff modeling. D2K Technical Report 0699-1-1, Danish Hydraulic Institute, Denmark.

Ferraro RR. 1997. Special sensor microwave imager derived global rainfall estimates for climatological applications. *Journal of Geophysical Research* **102**: 16715–16735.

Ferraro RR, Grody NC, Mards GF Marks. 1994. Effects of surface conditions on rain identification using the SSM/I. *Remote Sensing Reviews* **11**: 195–209.

Goldberg DE. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley: Reading, MA.

Grimes DIF, Coppola E, Verdecchia M, Visconti G. 2003. A neural network approach to real-time rainfall estimation for Africa using satellite data. *Journal of Hydrometeorology* **4**: 1119–1133.

Hollinger JP. 1989. DMSP special sensor microwave/imager calibration/validation, Final Report, vol. I. Space Sensing Branch, Naval Research Laboratory, Washington, DC.

Hollinger JP. 1991. *DMSP special sensor microwave/imager calibration/validation, Final Report, vol. II.* Space Sensing Branch, Naval Research Laboratory, Washington, DC.

Hollinger JP, Peirce JL, Poe GA. 1990. SSM/I instrument evaluation. *IEEE Transactions on Geoscience and Remote Sensing* **28**: 781–790.

Hsu K, Gupta H, Gao X, Sorooshian S. 1999. Estimation of physical variables from multichannel remotely sensed imagery using a neural network: application to rainfall estimation. *Water Resources Research* **35**: 1605–1618. DOI: 10.1029/1999WR900032.

Hsu KL, Gao X, Sorooshian S, Gupta HV. 1997. Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology* **36**: 1176–1190.

Huang SW. 2000. Analysis of west Pacific typhoon characteristics by using SSM/I satellite data. Master degree thesis of National Central University, Taiwan, 95 pages (available from the library of National Central University, Taiwan, in Chinese).

Kishore JK, Patnaik LM, Mani V, Agrawal VK. 2000. Application of genetic programming for multicategory pattern classification. *IEEE Transactions on Evolutionary Computation* **4**: 242–257.

Koza JR. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press: Cambridge, MA.

Krasnopolsky VM, Gemmill WH, Breaker LC. 1999. A multi-parameter empirical ocean algorithm for SSM/I retrievals. *Canadian Journal of Remote Sensing* **25**: 486–503.

Kuligowski RJ, Barros AP. 2001. Combining IR-microwave satellite retrieval of temperature and dewpoint profiles using artificial neural networks. *Journal of Applied Meteorology* **40**: 2051–2067.

Li CC, Chen WJ, Chen YC, Hu JC, Tsai MD. 2004. Typhoon rain retrievals using TMI, AMSU-A, and island gauge data sets. In *13th Conference on Satellite Meteorology and Oceanography*.

Liu GR, Chao CC, Ho CY. 2008. Applying satellite-estimated storm rotation speed to improve typhoon rainfall potential technique. *Weather and Forecasting* **23**: 259–269. DOI: 10.1175/2007WAF2006101.1.

Mimikou MA, Baltas EA. 1996. Flood forecasting based on radar rainfall measurements. *Journal of Water Resources Planning and Management* **122**: 151–156.

Omolbani MR, Lee TS, Amir AD. 2010. Review of Genetic Programming in Water Resource Engineering. *Australian Journal of Basic and Applied Sciences* **4**(11): 5663–5667.

Petty GW. 1995. The status of satellite-based rainfall estimation over land. *Remote Sensing of Environment* **51**: 125–137.

Savic DA, Walters GA, Davidson JW. 1999. A genetic programming approach to rainfall-runoff modeling. *Water Resources Management* **13**: 219–231.

Smith EA, Lamm JE, Adler R, Alishouse J, Aonashi K, Barrett E, Bauer P, Berg W, Chang A, Ferraro R, Ferriday J, Goodman S, Grody N, Kidd C, Kniveton D, Kummerow C, Liu G, Marzano F, Magnai A, Olson W, Petty G, Shibata A, Spencer R, Wentz F, Wilheit T, Zipser E. 1998. Results of WetNet PIP-2 project. *Journal of the Atmospheric Sciences* **55**: 1483–1536.

Sorooshian S, Hsu KL, Gao X, Gupta HV, Imam B, Braithwaite D. 2000. Evaluation of PERSIANN system satellite based estimates of tropical rain. *Bulletin of the American Meteorological Society* **81**: 2035–2046.

Spencer RW, Goodman HM, Hood RE. 1988. Precipitation retrieval over land and ocean with the SSM/I: identifications and characteristics of the scattering signal. *Journal of Atmospheric and Oceanic Technology* **6**: 254–273.

Staelin DH, Chen FW, Fuentes A. 1999. Precipitation measurements using 183-GHz AMSU satellite observations. In *Proceedings of the 1999 IEEE International Geoscience and Remote Sensing Symposium*, vol. 4. Hamburg, Germany; 2069–2071.

Trafalis TB, Richman MB, White A, Santosa B. 2002. Data mining techniques for improved WSR-88D rainfall estimation. *Computers & Industrial Engineering* **43**: 775–786.

Wei C, Hung WC, Cheng KS. 2006. A multi-spectral spatial convolution approach of rainfall forecasting using weather satellite imagery. *Advances in Space Research* **37**: 747–753. DOI: 10.1016/j.asr.

Wei HP, Yeh KC, Liu GR, Chao CC. 2011. Combining satellite data for estimation of rainfall at watershed scale. *International Journal of Remote Sensing* DOI: 10.1080/01431161.2010.517227.

Whigham PA, Crapper PF. 1999. Time series modeling using genetic programming: an application to rainfall-runoff models. In *Advances in Genetic Programming*, Spector L, Langdon WB, Una-May O'Reilly, Angeline PJ (eds). MIT Press: Cambridge, MA; 89–104.

Whigham PA, Crapper PF. 2001. Modeling rainfall-runoff using genetic programming. *Mathematical and Computer Modelling* **33**: 707–721.

Wilheit TT. 1994. Algorithm for the retrieval of rainfall from passive microwave measurements. *Remote Sensing Reviews* **11**: 163–194.

Wilheit TT, Chang AT, Rao SV, Rodger EB, Theon JS. 1997. A satellite technique for quantitatively mapping rainfall rates over the ocean. *Journal of Applied Meteorology* **16**: 551–560.

Xia Y. 2001. Sea-surface temperature and sea-surface windspeed retrievals from spaceborne radiometer measurements. Ph.D. dissertation, University of Massachusetts-Amherst; 105.