A clonal selection algorithm model for daily rainfall data prediction

N. S. Noor Rodi, M. A. Malek, Amelia Ritahani Ismail, Sie Chun Ting and Chao-Wei Tang

ABSTRACT

This study applies the clonal selection algorithm (CSA) in an artificial immune system (AIS) as an alternative method to predicting future rainfall data. The stochastic and the artificial neural network techniques are commonly used in hydrology. However, in this study a novel technique for forecasting rainfall was established. Results from this study have proven that the theory of biological immune systems could be technically applied to time series data. Biological immune systems are nonlinear and chaotic in nature similar to the daily rainfall data. This study discovered that the proposed CSA was able to predict the daily rainfall data with an accuracy of 90% during the model training stage. In the testing stage, the results showed that an accuracy between the actual and the generated data was within the range of 75 to 92%. Thus, the CSA approach shows a new method in rainfall data prediction.

Key words | artificial immune system, clonal selection algorithm, daily rainfall, prediction

N. S. Noor Rodi
M. A. Malek
Sie Chun Ting (corresponding author)
Department of Civil Engineering,
Universiti Tenaga Nasional,
IKRAM-UNITEN Road, 43000 Kajang, Selangor,
Malaysia
E-mail: *sie_chun@hotmail.com*

Amelia Ritahani Ismail
Department of Computer Science, Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia,
P.O. Box 10, 50728 Kuala Lumpur,
Malaysia

Chao-Wei Tang
Department of Civil Engineering and Geomatics,
Cheng Shiu University, Kaohsiung City,
Taiwan

INTRODUCTION

In this study, the clonal selection algorithm (CSA) in an artificial immune system (AIS) is utilized in the development of a daily rainfall prediction model. AIS is one of the branches of computing inspired by biological systems. The AIS is used to describe a wide-range of systems, taking inspiration from different aspects of immunology (Timmis & de Castro 2002). Like other bio-inspired paradigms, the AIS attempts to define the properties of the biological systems upon which they are based. Many philosophies and perceptions have been extracted from biological immune systems to develop a new algorithm. The new algorithm can be applied in the real world of engineering and scientific problems. In biology, the main role of an immune system is to defend human bodies from the attacks of external microorganisms or to launch a response to the invading pathogens (de Castro & Zuben 2002).

There are several applications of AIS to computer science and engineering problems, such as optimization, classification and intrusion detection. The AIS has three main algorithms, CSA (Burnet 1987), Immune Network Algorithm (Farmer *et al.* 1986) and Negative Selection Algorithm (Forest *et al.* 1994). Figure 1 shows the relationships in the human biological immune system.

CLONAL SELECTION ALGORITHM

CSA is inspired by the clonal selection principle that only those cells capable of recognizing an antigenic stimulus will proliferate and differentiate into selected cells (Timmis & de Castro 2002). The clonal selection theory in a human body system comprises the immunological processes of the cloning selection and affinity maturation. The main operators in a CSA are selection, cloning and mutation. The clonal selection principle describes how an adaptive immunity functions when a foreign pathogen attacks the organism (Timmis & de Castro 2002). The basic idea of a CSA involves matching the antibodies (Abs) with antigen. The generated Abs will create clones of themselves. This process is called proliferation (Yashwant & Amir 2011). The B-cells and T-cells are major cells in the immune system antibodies. The CSA objective is to develop an antibody pool that represents the solutions, while the antigens represent the elements of an evaluation. In short, the CSA essentially centers on a repeated cycle of match, clone, mutate and replace, and numerous parameters can be tuned, including the cloning rate, the initial number of antibodies, and the mutation rate for the clones (Greensmith *et al.* 2010).

doi: 10.2166/wst.2014.420

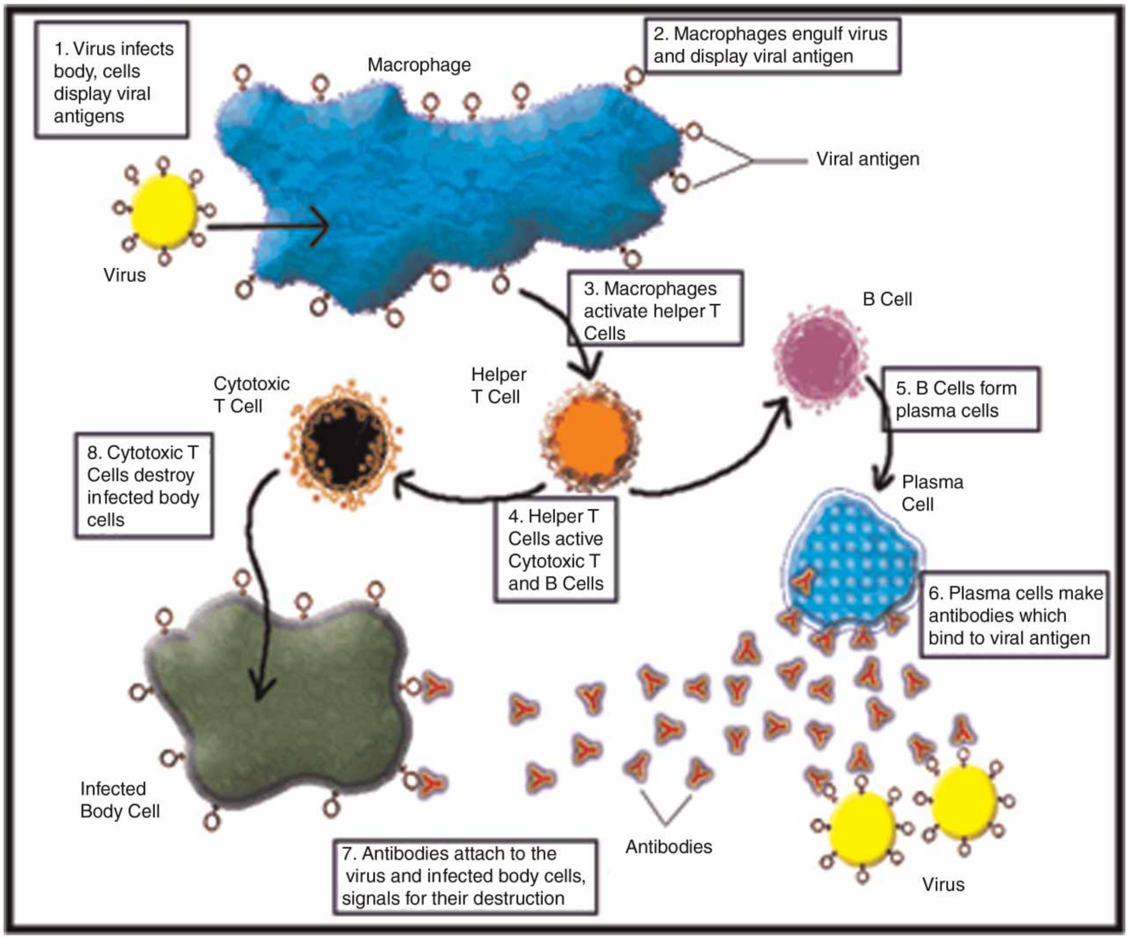
1. Virus infects body, cells display viral antigens
Macrophage
2. Macrophages engulf virus and display viral antigen
Viral antigen
Virus
3. Macrophages activate helper T Cells
B Cell
Cytotoxic T Cell
Helper T Cell
5. B Cells form plasma cells
8. Cytotoxic T Cells destroy infected body cells
Plasma Cell
4. Helper T Cells active Cytotoxic T and B Cells
6. Plasma cells make antibodies which bind to viral antigen
Infected Body Cell
Antibodies
7. Antibodies attach to the virus and infected body cells, signals for their destruction
Virus

Figure 1 | Human biological immune system.

RAINFALL MODELLING

The rainfall model has become an increasingly indispensable tool in the planning of water resources projects. It can be used to assess floods, aid in reservoir operation, and predict transport in water born contamination. Accuracy in rainfall prediction can bring a great achievement in water resources planning. Accurate rainfall analysis is pertinent for applications such as severe rainfall and flash flood forecasting (Georgakakos & Hudlow 1984; Seo & Smith 1992), long term water resources planning and management (Newsome & Collier 1991), crops-yield forecasting, and studies of probable design storms and floods (Hardaker & Collier 1995). However, detailed spatial and temporal information on rainfall intensity distribution is required for rainfall prediction. Therefore, model prediction is important for water resources engineering management.

There are many methods or models that have been used for rainfall prediction. Each model has its own achievement and accuracy. Methods for rainfall prediction can be divided into two categories, which are the statistical and the artificial intelligence. Several techniques have been utilized to generate rainfall data, such as regression (Ranhao *et al.* 2011), Gray Markov (Liu *et al.* 2011) and flow duration matching (Hughes & Smakhtin 1996). The biological immune system has good learning ability in new things, and it can be applied to complicated model matching and organization of network structure to support the memory of things it meets (Zhao & Davis 2011). In this study, a proposed CSA is employed for rainfall modelling.

Model development

The development of the rainfall prediction model utilizing the CSA approach is summarized in Figure 2. The steps in

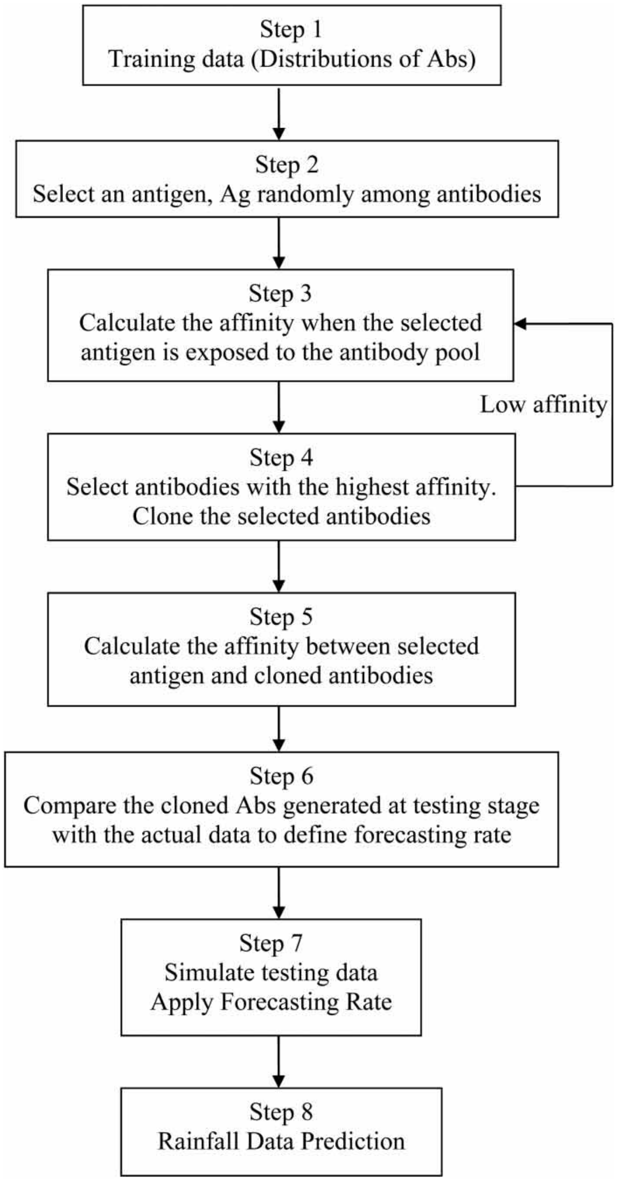
Step 1
Training data (Distributions of Abs)
Step 2
Select an antigen, Ag randomly among antibodies
Step 3
Calculate the affinity when the selected antigen is exposed to the antibody pool
Low affinity
Step 4
Select antibodies with the highest affinity. Clone the selected antibodies
Step 5
Calculate the affinity between selected antigen and cloned antibodies
Step 6
Compare the cloned Abs generated at testing stage with the actual data to define forecasting rate
Step 7
Simulate testing data
Apply Forecasting Rate
Step 8
Rainfall Data Prediction

Figure 2 | Flowchart of the CSA.

the CSA begin with the training of the input data, followed by the testing and prediction processes. The training process consists of four steps (Ting *et al.* 2013), as explained below. This is followed by the testing process, and then finally the prediction process.

Step 1: The daily rainfall data are used as an input. Theoretically this is where all the Abs are distributed. The Abs used during the training stage are the daily rainfall data from January 1990 to December 1999, excluding the missing data. In a CSA, the historical rainfall data represents the distribution of the Abs in a clonal selection theory. Each antibody in N forms a string of real numbers represented in a sequence used as the Abs

$$N = [\mathrm{Ab}_1, \mathrm{Ab}_2, \ \ldots \mathrm{Ab}_n] \tag{1}$$

Step 2: This step shows a group of antigens (Ag) attacking the Abs. In this study a random variable of antigen, Ag in the immune system is chosen from daily rainfall data on a random basis. When the Ag and Abs meet, the first measurement of an affinity value is calculated between the Ag and the Abs, which is performed in Step 3

$$\mathrm{Ag} = [\mathrm{Ag}_1, \mathrm{Ag}_2, \ \ldots \mathrm{Ag}_n] \tag{2}$$

Step 3: The affinity is measured using the Euclidean distance, D whereby the two attribute strings into a nonnegative real number that corresponds to their affinity or degree of match, $S^{\mathrm{L}} \times S^{\mathrm{L}} \rightarrow R^{+}$ as below

$$D = \sqrt{\sum_{i=1}^{L} (\mathrm{Ab}_i - \mathrm{Ag}_i)^2} \tag{3}$$

where Ab is the real numbers of antibodies; Ag is the random variable of antigen.

The mutation process is performed in Step 3. The number of mutations are named as number of detectors. In this study, the number of detectors used was 100, which had been proved (Noor Rodi *et al.* 2012).

Step 4: The cloning process is being performed by selecting the Abs with higher affinity. Here, the cells which are capable of recognizing an antigenic stimulus will proliferate and differentiate into the cells, being selected against those that do not. These selected Abs are then cloned. When the Abs are found to have a low affinity value, Step 3 is repeated. These iterations are named mutations. The testing process of the proposed prediction model begins from Step 4 until 7.

Step 5: The second measurement of the affinity values between the selected Ag and the cloned Abs (from Step 4) is calculated. The selected Ag is obtained from 10 cross-validations performed in the training process. This is to ensure that the accuracy of the model verification is being performed. In this study, 10 cross validations are used through trial and error. In order to obtain the selected Ag in Step 5, the forecasting rate (FR) is obtained from the calculated training process using

$$FR = \frac{\sum \text{Actual value} - \text{Simulated value}}{\text{No. of data}} \quad (4)$$

Step 6: The cloned Abs generated during the testing process are compared with the actual data.

Step 7: The percentage of accuracy of the proposed model is calculated. This is the end of the testing process.

Step 8: Lastly, the prediction process is performed using the final cloned Abs.

RESULTS AND DISCUSSION

In this study, the structure of an AIS Algorithm was designed using a self-written coding via MATLAB 7.12 (R2011a). The prediction process is performed using the final cloned Abs. A summary of the duration period was performed for the training, testing and prediction modelling as shown in Table 1.

Table 1 | Summary of duration period for CSA simulation

Rainfall station (ID)	Training period (year)	Testing period (year)	Prediction
Station 1 JPS Ampang (3117070)	1990–1999	2000–2011	2012–2022
Station 2 Genting Sempah (3317004)	1990–1999	2000–2011	2012–2022
Station 3 Klang Town (3014080)	2002–2006	2007–2011	2012–2016

Model validation

Historical rainfall data collected within the state of Selangor, Malaysia, were used to validate the proposed prediction model. The catchment area analysed is 7,950 km^2 and is located in the humid tropical zone. From the month of April to May, the state of Selangor is influenced by heavy rainy seasons called the south-westerly monsoon and once again from the month of October to December which is called the north-easterly monsoon. The source of the rainfall data is obtained from by the Department of Irrigation and Drainage Malaysia. The data from the three rain gauge stations located along Klang River are selected as input data. The three rain gauge stations selected are representing upstream (Station 1 – JPS Ampang), mid- (Station 2 – Genting Sempah) and downstream (Station 3 – Klang Town) of the Klang River.

In the model validation stage, the developed prediction model was validated through a graphical presentation of the actual and predicted data using the similarity percentage as the equation below:

$$\% \text{ of similarity} = 100 - \text{sum}(\text{accuracy}/10) \quad (5)$$

Comparisons between the actual and predicted rainfall data for all the three rainfall stations were analyzed in the testing stage as explained below.

(a) Station 1 (JPS Ampang)

Figure 3 shows a graphical presentation of the actual and predicted daily rainfall data in year 2000. From the simulation, the similarity percentage obtained was 86.53%. The predicted data were obtained using the historical data

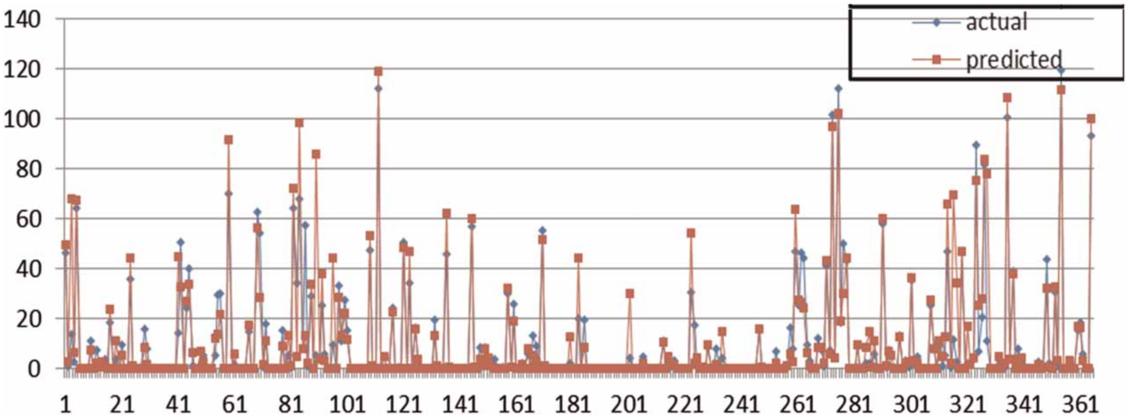

Figure 3 | Comparison between the actual and predicted rainfall data for Station 1 in year 2000.

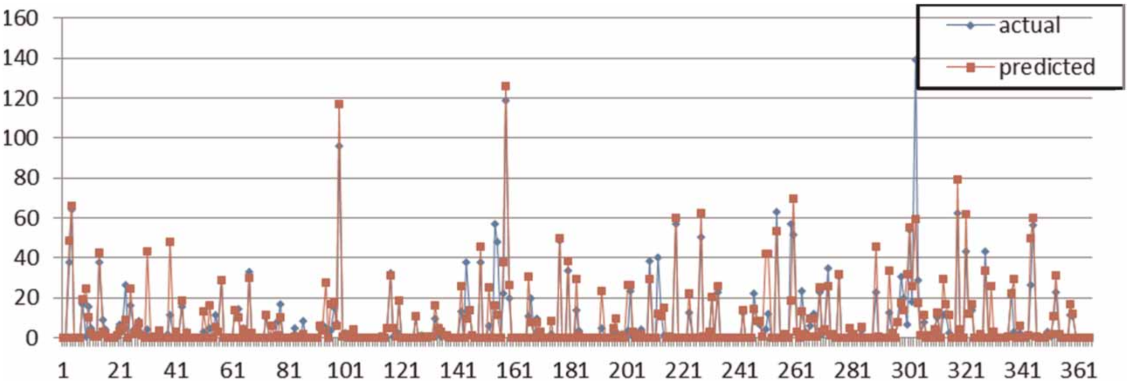

Figure 4 | Comparison between the actual and predicted rainfall data for Station 2 in year 2000.

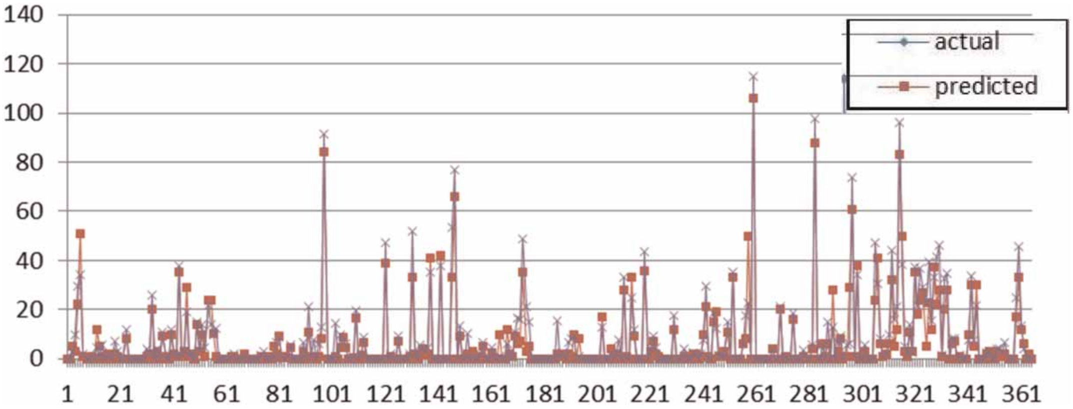

Figure 5 | Comparison between the actual and predicted rainfall data for Station 3 at year 2007.

from the pattern values from the years 1990 to 1999, which were calculated during the training stage.

(b) Station 2 (Genting Sempah)

The comparison between the actual and predicted daily rainfall data for Station 2 in year 2000 is shown in Figure 4. From the simulation, the similarity percentage obtained between the actual and predicted data was 92.24%.

(c) Station 3 (Klang Town)

Figure 5 shows the comparison between the actual and predicted rainfall data for Station 3 in year 2007. From the simulation, it was found that the similarity percentage between the actual and predicted rainfall data was 82.66%.

Table 2 | Summary of similarity percentage at testing stage

Rainfall station	Similarity percentage (%)	Historical data
Station 1 – JPS Ampang (3117070)	86.53	10 years
Station 2 – Genting Sempah (3317004)	92.24	10 years
Station 3 – Klang Town (3014080)	82.66	5 years

Table 2 shows a summary of similarity percentage between the actual and predicted rainfall data for all the three rainfall stations analyzed in the testing stage. The proposed model underwent a cross-validation process to further verify the model's improvement and capability as stated in Equation (4). From the results, Station 2, had the highest similarity percentage – 92.24%. During the training stage, the number of historical data for Station 1 was lower than Station 2 due to the missing data. Station 3 had the lowest percentage of similarity – 82.66%. The predicted data in this station were obtained from the 5 years of historical data available as compared to Station 1 and Station 2, which had 10 years of available historical data. It can be concluded that the greater the amount of historical data used, the greater the accuracy achieved. The CSA can be used as an alternative method for predicting future rainfall data, as compared to the traditional statistical, stochastic and artificial neural network techniques commonly used in hydrology.

CONCLUSIONS

This study has successfully developed a time series prediction model for the daily rainfall data using an AIS approach. It is proven that the theory of biological immune systems can be applied in this study. A large population of the historical rainfall data used can be represented by the population of antibodies in the human body system. The attacking antigens in the human body system are presented by the pattern of historical input data.

In the training stage of this study, it was found that the proposed CSA can predict with an accuracy of more than 90% at different numbers of detector sets. In the testing stage, the results obtained show that the percentages of accuracy between the actual and generated data were within the range of 75 and 92%.

During the testing stage, it was found that the Genting Sempah rainfall station had the highest percentage of similarity between the actual and generated data. From this observation, the numbers of historical data used during the training stage among these three rainfall stations were analysed differently. Even though the JPS Ampang and Genting Sempah rainfall stations were both using 10 years of historical data, the amount of missing data for the JPS Ampang rainfall station was higher than the Genting Sempah rainfall station. This has made the number of data available for the training stage less for the JPS Ampang rainfall station. Therefore, it can be concluded that the larger the amount of historical data available, the greater the accuracy of future data that can be made using the CSA approach.

ACKNOWLEDGEMENTS

The authors of this paper wish to acknowledge Prof. Dr Siti Mariyam bt. Shamsuddin of Faculty Computer Science, Universiti Teknologi Malaysia, Johor for her kind contributions and advice in preparation of this study. This study was supported by the Ministry of Higher Education Malaysia under grant no. FRGS/1-2012/(TK03)/(UNITEN)/(02/36). The authors wish to thank the Malaysia Drainage and Irrigation Department for access to data.

REFERENCES

Burnet, F. M. 1987 Clonal selection and after. In: *Theoretical Immunology* (G. I. Bell, A. S. Perelson & G. H. Pimbley Jr, eds). Marcel Dekker Inc., New York, pp. 63–85.

de Castro, L. & Zuben, F. J. 2002 Learning and optimization using clonal selection principle. *IEEE Transactions on Evolutionary Computation* **6** (3), 239–251.

Farmer, J. D., Packard, N. H. & Perelson, A. S. 1986 The immune system, adaptation, and machine learning. *Physica* **2**, 187–204.

Forest, S., Perelson, A. S., Allen, L. & Cherukuri, R. 1994 Self-nonself discrimination in computer. In: *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy*, Los Alamitos, CA.

Georgakakos, K. P. & Hudlow, M. D. 1984 Quantitative precipitation forecast techniques for use in hydrologic forecasting. *Bulletin American Meteorological Society* **65** (11), 1186–1200.

Greensmith, J., Whitbrook, A. & Aickelin, U. 2010 *Handbook of Metaheuristics*. 2nd edn, Chapter 14 (Artificial Immune Systems). Kluwer Academic Publishers, Norwell, MA, pp. 421–448.

Hardaker, P. J. & Collier, C. G. 1995 Radar and storm model-based estimation of probable maximum precipitation in tropics. In: *Proceedings of the 3rd International Symposium on Hydrology Applications of Weather Radar*, São Paulo, Brazil.

Hughes, D. A. & Smakhtin, V. U. 1996 Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. *Hydrological Sciences Journal* **41** (6), 851–871.

Liu, C., Tian, Y-m. & Wang, X-h. 2011 Study of rainfall prediction model based on GM(1,1)-Markov chain. In: *Proceedings of the International Symposium on Water Resource and Environmental Protection (ISWREP)*, 1, 744–747.

Newsome, D. H. & Collier, C. G. 1991 Possible hydrological applications of weather radar in Western Europe. In: *Hydrological Applications of Weather Radar* (I. D. Cluckie & C. G. Collier, eds). Ellis Horwood, Harlow, pp. 623–635.

Noor Rodi, N. S., Ismail, A. R. & Malek, M. A. 2012 Daily rainfall prediction using clonal selection algorithm. In: *Proceedings the International Conference Water Resources (ICWR2012)*, November 5–6, Langkawi.

Ranhao, S., Liding, C. & Bojie, F. 2011 Predicting monthly precipitation with multivariate regression methods using geographic and topographic information. *Physical Geography* **32** (3), 269–285.

Seo, D. J. & Smith, J. A. 1992 Radar-based short term rainfall prediction. *Journal of Hydrology* **131**, 341–367.

Timmis, J. & de Castro, L. N. 2002 *Artificial Immune System: A New Computational Intelligence Approach.* Springer Verlag, London.

Ting, S. C., Ismail, A. R. & Malek, M. A. 2013 Development of effluent removal prediction model efficiency in septic sludge treatment plant through clonal selection algorithm. *Journal of Environmental Management* **129**, 260–265.

Yashwant, P. S. & Amir, S. H. B. 2011 Modified clonal selection algorithm based classifiers. In: *Proceedings of the Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, IEEE Computer Society, Washington, DC, USA.

Zhao, W. & Davis, C. E. 2011 A modified artificial immune system based pattern recognition approach-an application to clinic diagnostics. *Artificial Intelligence in Medicine* **52** (1), 1–9.

First received 3 July 2014; accepted in revised form 6 October 2014. Available online 18 October 2014