Penerbit Akademia Baru

Journal of Advanced Research in Applied Mechanics
ISSN (online): 2289-7895 | Vol. 26, No. 1. Pages 1-9, 2016

Modelling Daily Streamflow using Genetic Algorithm

S. N. Mohammad[*,1,a], *I. Atan*[2,b], *A. M. Baki*[3,c], *A. R. Zulkifli*[4,d], *M. K. M. Yusof*[5,e], *D. C. Lat*[6,f] *and M. A. Hasbullah*[7,g]

[1,5,6,7]Faculty of Civil Engineering, Universiti Teknologi MARA, Cawangan Johor Kampus Pasir Gudang
[2,4]Faculty of Civil Engineering, Universiti Teknologi MARA, Shah Alam
[3]ENVIRAB Services, Shah Alam, Selangor, Malaysia
[a,*]*shahrul9688@johor.uitm.edu.my*, [b]*ismail5852@salam.uitm.edu.my*, [c]*aminbaki2@gmail.com*, [d]*assrul9552@salam.uitm.edu.my*, [e]*mohdkhatif@johor.uitm.edu.my*, [f]*dianacl@johor.uitm.edu.my*, [g]*mohdamran@johor.uitm.edu.my*

Abstract – *This paper presents the result of modelling daily streamflow for a catchment area in Malaysia by using rainfall and water level data. The data used in this paper are based on the data recorded at the Sungai Pahang basin obtained from Department of Irrigation and Drainage (DID) and this paper applies an approach of Genetic Algorithm model. Genetic programming software called Discipulus is used for modelling the daily streamflow. In this paper, the data of the rainfall and water level are extracted and filtered before the data are transformed into Training, Validation and Applied Data for optimization process of Discipulus. The daily streamflow model that generated from Discipulus is compared to actual streamflow data. The major aim of the research is to forecast daily streamflow based on rainfall and water level data, to evaluate the performance and reliability of the time series model for daily streamflow forecasting where evaluation of model performance is based on Genetic Algorithm modelling by using Discipulus. The result shows that genetic programming able to predict reliable model of daily streamflow for Sungai Pahang basin and it is recommended that more catchment areas to be modeled using genetic programming in order to achieve better result interpretation.*

Keywords: Discipulus, Modelling, Streamflow, Genetic algorithm, Genetic algorithm programming, Rainfall and water level data.

1.0 INTRODUCTION

Rainfall is the driving force behind storm water studies and design. Adequate knowledge of the rainfall-runoff process is important for optimal design of water storage and drainage networks, management of extreme events such as floods and droughts, and determination of the rate of pollution transport [1, 2]. An understanding of rainfall process and accurate rainfall data is necessary for preparing satisfactory drainage and storm water management systems. In Malaysia, rainfall data is collected by several departments and authorities including Malaysian Meteorological Department and Department of Irrigation and Drainage (DID).

Urbanization has brought remarkable impacts such as increase in peak runoff and decrease in time of concentration due to land covers change [3]. The principles of watershed hydrology cannot be directly applied to urban hydrology because flow in an urban setting concentrated into swales, open channels and storm sewers, all of which accelerate the flow [4]. Thus modern

methods of rainfall-runoff analysis are needed to quickly assess and monitor the degree of severity of these impacts. Considerable effort has been directed towards development of mathematical models that can be used to analyse the urban drainage system [5]. This paper aims to forecast the daily streamflow based on rainfall data and water level input and at the same time to evaluate the performance and reliability of the time series model for daily streamflow forecasting. The evaluation of model's performance is based on genetic algorithm programming.

In order to provide understanding of rainfall-runoff process, a model is calibrated to mimic real hydrological conditions in the catchments. This model might give the peak discharge from a small watershed as some function of the watershed area, a flood frequency curve, a comprehensive "deterministic" model capable of generating synthetic streamflow records or it might be a stochastic model for generating a time series of hydrologic data [6].

Modifying model parameters is essential to minimize prediction errors since surface runoff varies in response to the characteristics of the catchments [7, 8]. Therefore, in most application, a model must be guided into a useful tool by calibration and validation. A model should be able to reproduce the behaviour of a watershed and be precise in terms of reproducibility, which is strongly related to uncertainty or random errors [5].

In genetic algorithms, solutions to a problem are translated into coded strings. Genetic algorithm search the optimal solution based on Darwin's theory on the survival of the fittest [9]. It means the strings survive from one generation to another and trade part of their genetic material with other strings depending on their robustness as defined by the objective function [10]. Genetic algorithms are extensively used stochastic search methods initially developed by Holland [11, 12]. Genetic algorithms easily accept discontinuities present in the formulation of the problem. They are also less dependent on initial conditions as they consider several solutions simultaneously instead of only one, and this reduces the risk of ending the optimization process at a local optimum. They are also well suited for combinatorial problems and algorithm solving model [13, 14].

Genetic algorithms have been regularly employed on applications to water resources and hydrology. Recent applications in these domains have been performed by Samuel and Jha in 2003 [15] for groundwater modelling and Cheng in 2002 [16] for rainfall-runoff modelling. A particular advantage of genetic algorithms is that they are easy to implement, using very simple yet efficient rules for reproduction of coded strings from one generation to another, crossover for the sharing of information among the strings, and mutation to ensure the diversity of solutions [13].

2.0 METHODOLOGY

2.1 Study Area

The study is conducted at sungai Pahang basin. Sungai Pahang Basin has a catchment area of 29300km2 where the entrance to Sungai Pahang is located on an extensive, flat coastal plain with 30km wide. Sungai Jelai and Sungai Tembeling linked at Kuala Tembeling to form Sungai Pahang, which flows to south. The gradient of Sungai Pahang is even, averaging 1 to 6500 from Kuala Tembeling to the South China Sea. In this paper, daily rainfall and water level data from the gauging stations of Sungai Pahang Basin are selected. Figure 2 shows telemetry station in Pahang.

Penerbit Akademia Baru
Journal of Advanced Research in Applied Mechanics
ISSN (online): 2289-7895 | Vol. 26, No. 1. Pages 1-9, 2016

2.2 Data Collection

The hydrological data for the rainfall and water level data in the study were obtained from DID for the Sungai Pahang catchments. Daily rainfall and water level data of 20 years period were selected for the eight selected stations. However, analysis has been done to determine and filter only the accurate and complete data. Then, these data are input in Discipulus to be learnt by the software in order to develop a streamflow forecasting model. The optimization process of Discipulus used in the methodology for this paper is summarized in Figure 1.

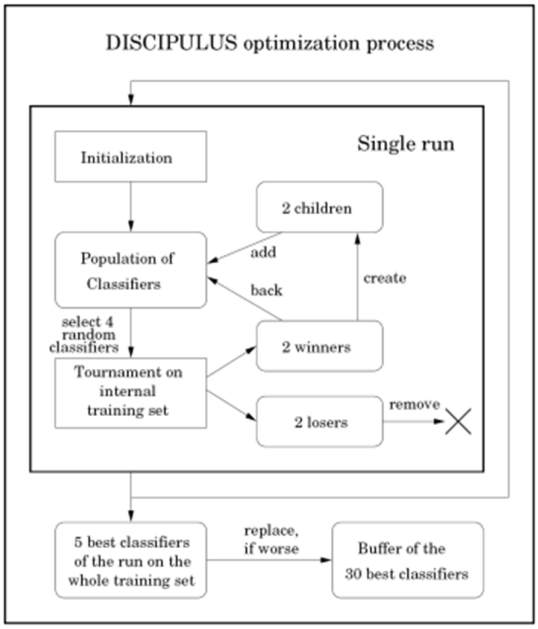

Figure 1: Optimization process of Discipulus

2.2 Presentation of Data

The samples of data obtained are presented in table form as in Table 1. The data for rainfall at four different stations (X1, X2, X3, X4) and water level at four different stations (Y1, Y2, Y3, Y4) is recorded daily for 20 years. Therefore, there are all together 7300 numbers of data but only 2048 numbers of data are fit to be used in this paper due to incomplete and inaccurate data. The data were split into Training, Validation and Applied Data sets before the data are input in the Discipulus. The data is analysed by Discipulus and genetic programming is applied to develop relationship between the future water level at station Y4 with the rainfall data (X1, X2, X3, X4) and water level data (Y1, Y2, Y3).

Journal of Advanced Research in Applied Mechanics
ISSN (online): 2289-7895 | Vol. 26, No. 1. Pages 1-9, 2016

Table 1: Samples of rainfall and water level data

X1	X2	X3	X4	Y1	Y2	Y3	Y4
Rainfall (mm)	Rainfall (mm)	Rainfall (mm)	Rainfall (mm)	Water level (m)	Water level (m)	Water level (m)	Water level (m)
2.8	0.0	8.5	0.0	31.61	43.40	67.37	13.00
0.0	0.0	12.5	0.0	31.85	43.59	67.53	12.97
6.6	6.5	10.0	0.0	31.65	43.91	66.93	13.03
0.0	0.0	30.0	0.0	31.55	44.47	66.85	13.16
0.0	0.0	6.5	0.0	31.86	44.50	66.87	13.66
0.0	5.5	6.5	0.0	32.13	43.84	67.01	14.12
0.0	3.0	30.0	0.0	32.38	43.60	67.18	14.03
47.0	1.0	5.0	0.0	32.40	43.62	67.58	13.68
18.0	4.0	15.0	0.0	32.42	43.75	67.47	13.55
0.0	7.5	35.0	0.0	32.60	44.11	67.50	13.56

3.0 RESULTS AND DISCUSSION

In order to simplify the daily streamflow model, the data for rainfall at four stations (X1, X2, X3, X4) and water level at four stations (Y1, Y2, Y3, Y4) is consider as the daily data. Therefore, there is only one set of data for each day. After these data is input in Discipulus, it creates comprehensive reports that show the Best Program and the Best Team Model output after the tabulated input data are ran. The target output is the actual reading of water level at station Y4. It is found that the genetic programming model is able to evolve an equation that trains well on the training and validation data and the performance on the applied data is also good.

The data need to be divided into three sets which are training, validation, and applied data. Training and validation data need to be input in Discipulus at the same time and Discipulus will come out with a daily streamflow model and also to verify the model. Then, applied data is used to check the performance and reliability of the model to predict the water level at station Y4.

3.1 Genetic Programming Forecast and Results

Figure 3, Figure 4, and Figure 5 shows the actual water level and forecasted water level at the catchment outlet (Y4) versus time. The time is referring to the day where the data of actual

Penerbit Akademia Baru

Journal of Advanced Research in Applied Mechanics
ISSN (online): 2289-7895 | Vol. 26, No. 1. Pages 1-9, 2016

water level is recorded. Actual water level is obtained from DID based on recorded data while forecasted water level is generated by Discipulus. The data is plotted to evaluate how well the Discipulus would be able to predict the water level at catchment outlet (Y4) by comparing with the actual water level. It is presented in Figure 3, Figure 4 and Figure 5 that the genetic programming for training model, validation model, and applied model consistently predicts the same trend of the water level at station Y4 compared with actual data. It shows that Genetic programming model trains well and able to duplicate the actual reading of the water level at station Y4 with a small tendency of error.

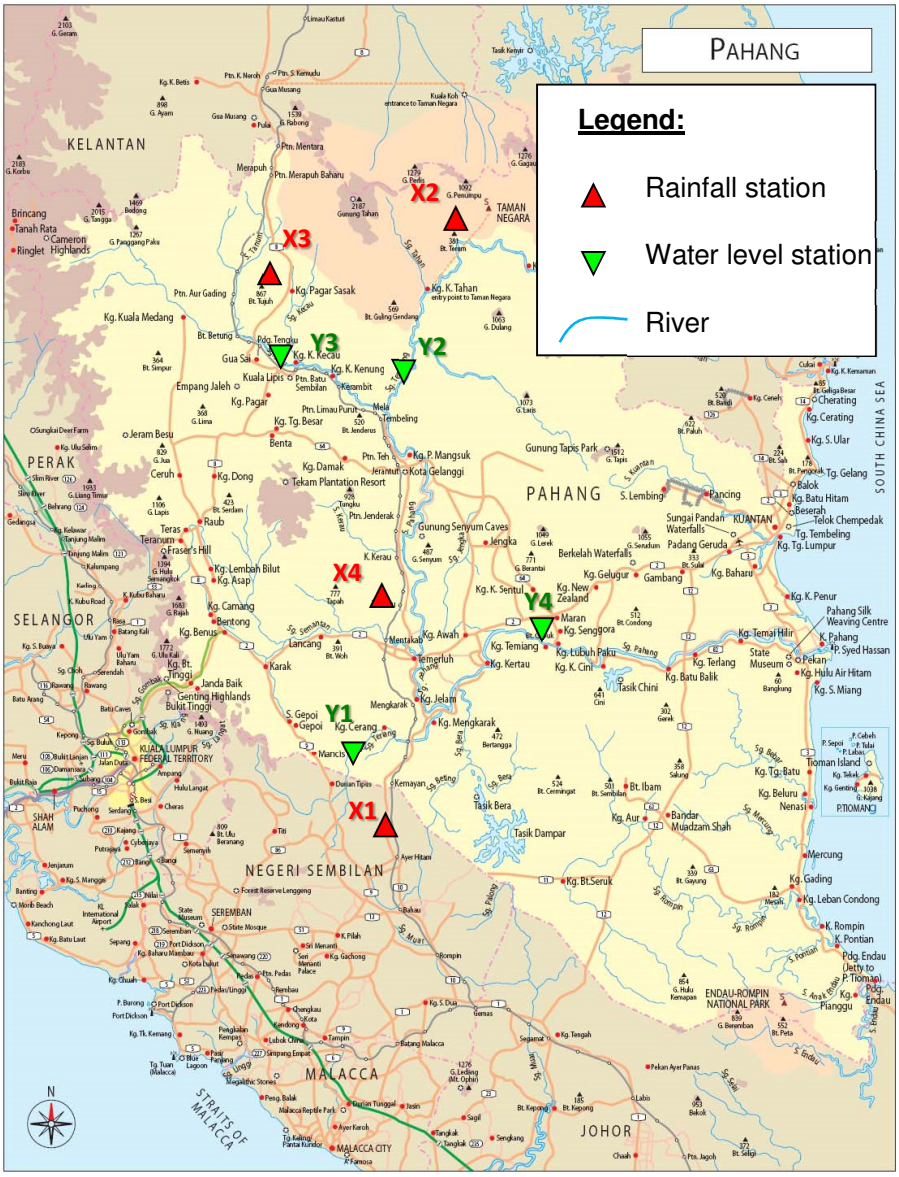

Figure 2: Observation stations for rainfall and water level

Journal of Advanced Research in Applied Mechanics
ISSN (online): 2289-7895 | Vol. 26, No. 1. Pages 1-9, 2016

3.2 Training Data

The training data contains the data that Discipulus used for learning. In other words, the fitness function or the hit-rate in Discipulus is calculated on the training file. The training data is a daily basis data that obtained from DID which consist of daily tabulated rainfall at station X1, X2, X3, and X4 and also daily water level at station Y1, Y2, Y3, and Y4. Once this data is uploaded, Disciplulus will evolve program based on daily tabulated rainfall at station X1, X2, X3, X4 and daily water level at station Y1, Y2, Y3 to predict the water level at station Y4. An evolved program is ranked as more fit during training session if it generates better prediction of water level at station Y4 when compared to the actual data of water level at station Y4 in the training file. The difference between Discipulus forecasted data and actual data of water level at station Y4 is visualised in Figure 3. It shows that Discipulus is able to imitate actual data of water level at station Y4.

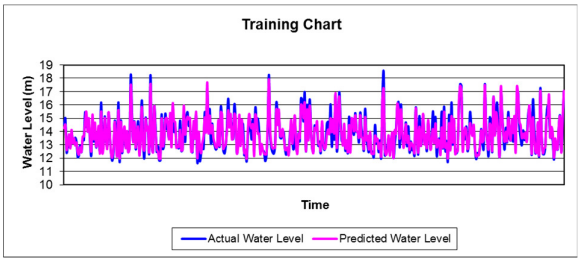

Figure 3: Training Data Chart

3.3 Validation Data

Training and validation data must be uploaded simultaneously before Discipulus can be ran. In the practise, Discipulus will train and validate the program at the same time. The validation data is used by Discipulus to pick the best programs from the population. The validation data should contain examples that are of the same type and structure as the training examples and that comprise a good representative set of samples from the learning domain. Validation data will ensure that the best program for the prediction of water level at station Y4 is chosen so that it can accurately forecast the actual water level. Figure 4 shows the tabulated data for the forecasted water level by Discipulus and actual water level at station Y4. Figure 4 indicates that Discipulus come out with the best program which able to predict the actual water level at station Y4 with minimum deviation.

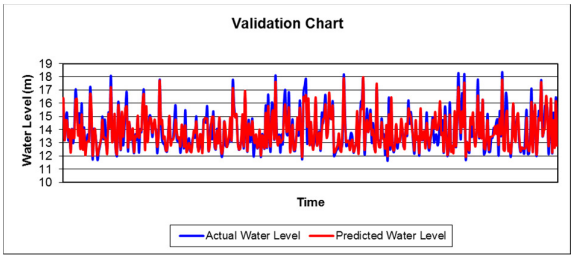

Figure 4: Validation Data Chart

Journal of Advanced Research in Applied Mechanics
ISSN (online): 2289-7895 | Vol. 26, No. 1. Pages 1-9, 2016

3.4 Applied Data

The whole point of Discipulus is to evolve programs that are useful on data that are unavailable when the program is trained and it can be done by using applied data. The applied data contain examples that are of the same type and structure as the training examples and that comprise a good representative set of samples from the learning domain. At this stage data of daily rainfall at station X1, X2, X3, and X4 and also daily water level at station Y1, Y2, Y3 is uploaded into Discipulus. Discipulus utilize the best program to predict the water level at station Y4. Then the forecasted water level at station Y4 is compared with the actual data of the water level at station Y4 and presented in Figure 5. It shows that Discipulus manage to imitate the actual data where the pattern of forecasted data and actual data of water level at station Y4 is similar although there is small discrepancy.

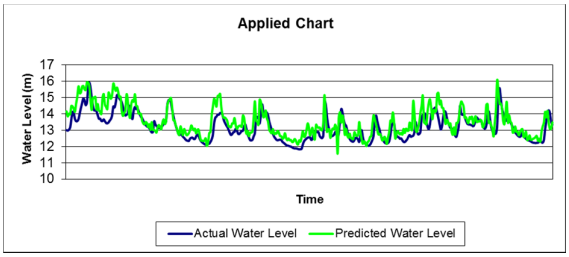

Figure 5: Applied Data Chart

Table 2: Goodness of fit measures for the catchments using MAPE

Data	MAPE (%)
Training	2.36
Validation	2.85
Applied	3.66

In this paper, Discipulus forecast the water level at station Y4 based on input data of daily rainfall at station X1, X2, X3, and X4 and also daily water level at station Y1, Y2, Y3. The comparison of data between forecasted water level and actual water level at station Y4 is presented in Figure 3, Figure 4, and Figure 5 for training data, validation data and applied data respectively. Then, the forecast accuracy for the data is measured by applying Mean Absolute Percentage Error (MAPE) as in (1). The Mean Absolute Percentage Error between forecasted data and actual data of water level at station Y4 is calculated for training data, validation data and applied data. The result in Table 2 shows that Mean Absolute Percentage Error of 2.36% for training, 2.85% for validation, and 3.66% for applied data. This shows that the program has used the training and validation data as a learning domain and produces the best applied data.

Journal of Advanced Research in Applied Mechanics
ISSN (online): 2289-7895 | Vol. 26, No. 1. Pages 1-9, 2016

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100\% \quad (1)$$

Where At, Ft, n are the actual data (actual data of water level at station Y4), forecasted data (forecasted data of water level at station Y4), and number of data respectively.

4.0 CONCLUSION AND RECOMMENDATION

In this paper, the suitability of using the data-driven genetic programming for modelling the rainfall-runoff process in catchments is studied. It is found that Disciplulus successfully forecast the water level for the Sungai Pahang Basin catchments based on daily rainfall data and daily water level data where the MAPE for training, validation, and applied data is 2.36%, 2.85% and 3.66% respectively. Thus, it proves that Discipulus is reliable in modelling the daily streamflow and it shows that simple and small genetic programming models can be evolved and come out with the best program which able to forecast the actual water level for the Sungai Pahang Basin catchments.

Based on the results, genetic algorithm programming can be extended to be applied in forecasting water level at river basin during flood to enhance disaster management in the country. For that purpose, a more accurate data of rainfall and water level from corresponding stations is essential so that genetic algorithm manage to forecast an accurate and useful data for that purpose. Based on the study of streamflow modelling using genetic algorithm programming, the followings conclusions are drawn:

- Genetic algorithm programming is easy to be applied and used by end-user to analyse any kind of data problems that requires forecasting.
- Genetic algorithm programming is a reliable tool in developing a model to forecast the daily streamflow based on rainfall and water level input.
- Rainfall and water level input can be used to study streamflow for forecasting water level of river basin with long reliable record.
- It is identified that genetic algorithm programming is able to provide an alternative method of streamflow generation. This can be applied to predict the water level of a specific catchment area for future studies.

The followings are the suggestions for the further studies in similar streamflow modelling:

- More catchment areas needed to be modeled using genetic algorithm approach in order to achieve better result interpretation. A large number of data stations may lead to a better result.
- Hourly raining and water level data should be used in similar studies to improve capability of solving more difficult problem.

REFERENCES

[1] M. Franchini, and G. Galeati. "Comparing several genetic algorithm schemes for the calibration of conceptual rainfall runoff models." Hydrological Sciences Journal, vol.42 (3), (1997): 357– 379.

[2] A. Ismail, L. Jahanshaloo and A. Fazeli. "Lagrangian grid LBM to predict solid particles' dynamics immersed in fluid in a cavity." Journal of Advanced Research in Fluid Mechanics and Thermal Sciences, vol. 3 (1), (2014): 17–26.

[3] C. E. Imrie, S. Durucan, and A. Korre. "River flow prediction using artificial neural network: Generalization beyond the calibration range." Journal of Hydrology, vol.233, (2000): 138–153.

[4] J. Smith, and R.N. Eli. "Neural network models of the rainfall runoff process." Journal of Water Resources Planning and Management ASCE, vol. 121, (1995): 499–508.

[5] N. R. Siriwardene, and B. J. C. Perera. "Selection of genetic algorithm operators for urban drainage model parameter optimization." Mathematical and Computer Modelling, vol. 44, (2006): 415–429.

[6] L. S. Kuchment, and A. N. Gelfan. "Dynamic stochastic models of rainfall and snowmelt runoff." Hydrology Science Journal, vol. 36, (1991): 153–169.

[7] L. S. Kuchment, and A. N. Gelfan. "Estimation of extreme flood characteristics using physically based models of runoff generation and stochastic meteorological inputs." Water International, vol. 27, (2002): 77–86.

[8] M. H. Kashani, M. A. Ghorbani, Y. Dinpashoh, and S. Shahmorad. "Integration of volterra model with artificial neural networks for rainfall-runoff simulation in forested catchment of northern Iran." Journal of Hydrology, vol. 540, (2016): 340–354.

[9] Y. H. Yun, W. T. Wang, B. C. Deng, G. B. Lai, X. B. Liu, D. B. Ren, Y. Z. Liang, W. Fan, and Q. S. Xu. "Using variable combination population analysis for variable selection in multivariate calibration." Analytica Chimica Acta, vol. 862, (2015): 14–23.

[10] Schoenauer, Marc. Handbook of evolutionary thinking in the sciences. Springer Netherlands, 2015.

[11] J. H. Holland. Adaptation in Natural and Artificial Systems, 2nd ed. MIT Press, 1992.

[12] J. Periaux, F. Gonzalez, and D. S. C. Lee. "Evolutionary methods." Evolutionary Optimization and Game Strategies for Advanced Multi-Disciplinary Design, vol. 75, (2015): 9–20.

[13] D. E. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning, 1st ed. Addison Wesley Longman Publishing Co, 1989.

[14] J. H. Mbaya, and N. Amin. "Modelling unsteady flow of gas and heat transfer in producing wells." Journal of Advanced Research in Fluid Mechanics and Thermal Sciences, vol. 6 (1), (2015): 19–33.

[15] M. P. Samuel, and M. K. Jha. "Estimation of aquifer parameters from pumping test data by genetic algorithm optimization technique." Journal of Irrigation and Drainage Engineering, vol. 129 (5), (2003): 348–359.

[16] C. T. Cheng, C. P. Ou, and K. W. Chau. "Combining a fuzzy optimal model with a genetic algorithm to solve multi objective rainfall runoff model calibration." Journal of Hydrology, vol. 268 (1–4), (2002): 72–86.