Available online at www.sciencedirect.com

ScienceDirect

Aquatic Procedia 4 (2015) 1016 – 1022

Aquatic Procedia

www.elsevier.com/locate/procedia

INTERNATIONAL CONFERENCE ON WATER RESOURCES, COASTAL AND OCEAN ENGINEERING (ICWRCOE 2015)

Infilling of Rainfall Information using Genetic Programming

C. Sivapragasam[a], Nitin Muttil[b], M. Catherin Jeselia[c], S. Visweshwaran[a]*

[a]*Department of Civil Engineering, Kalasalingam University, Tamil Nadu-626126, India*
[b]*College of Engineering and Science, Victoria University, PO Box-14428, Australia*
[c]*Department of Civil Engineering, NITK Surathkal, Karnataka-575025, India*

Abstract

The study suggests the use of Genetic Programming (GP) based monthly model for infilling of missing rainfall records in the rainfall time series for 3 rain gauge stations in the Yarra River Basin in Australia from the available rainfall information from the nearby stations. This study compares simple linear model, polynomial model, logarithmic model and a complex model based on GP to infill the missing monthly rainfalls. The RMSE and CC values of the validation data indicate the potential of the suggested model. Further, it is also interesting to note that GP evolved mathematical models are able to predict the subtle inherent non-linearity in the apparently predominantly linear behavior of the process.

Peer-review under responsibility of organizing committee of ICWRCOE 2015

*Keywords:*infilling rainfall; mathematical model; genetic programming; rain gauge stations

1. Introduction

All the hydrological studies in a river basin depend primarily on how accurately the rainfall is recorded and its distribution over the basin is estimated both temporally and spatially. As it is often seen, gaps do occur in the rainfall time series due to various reasons. Of the two methods primarily used for estimating the missing rainfall viz., stochastic modeling of rainfall sequences and interpolation based methods (Villazon and Willems, 2010), the later has been seen many applications in the literature which are implemented starting from simpler techniques like

* Corresponding author. Tel.: +91-4563-289042; fax: +91-4563-289322.
E-mail address:sivapragasam25@gmail.com

Peer-review under responsibility of organizing committee of ICWRCOE 2015
doi:10.1016/j.aqpro.2015.02.128

Arithmetic Mean Method, Normal Ratio Method, Inverse Distance Method to more complex and sophisticated techniques like kriging and Artificial Neural Networks (ANN) etc.

For instance, Ilunga (2010) preferred ANN to infill the Missing Annual Total Rainfall data for the stations around Orange River in South Africa. By taking the Bleskop rainfall station as the control and the Luckhoff-Pol rainfall station as the target, missing values are being infilled by using Standard Back-Propagation (BP) techniques and Generalized BP techniques. The results based on the RMSE values indicate that the generalized BP technique performed slightly better than the standard BP technique when applied to annual rainfall data.

Coulibaly and Evora (2007) investigated six different types of ANNs namely the Multilayer Perceptron (MLP), the Time-lagged Feed Forward Network (TLFN), the Generalized Radial Basis Function (RBF) network, the Recurrent Neural Network (RNN), the Time Delay Recurrent Neural Network (TDRNN) and the Counter-propagation Fuzzy Neural Network (CFNN) along with different optimization methods for infilling missing daily total precipitation records and daily extreme temperature series. The results suggest that the MLP, the TLFN and the CFNN can provide the most accurate results for the missing precipitation values. Over all the MLP appears the most effective at infilling missing daily precipitation values.

De Silva et al. (2007) studied Arithmetic Mean, Normal Ratio and Inverse Distance method besides proposing a new technique Arial Precipitation Ratio method for selected rain gauge stations in agro-ecological zones of Sri Lanka for daily records. The results indicate that different methods are appropriate for different zones of location of rain gauges. Villazon and Willems (2010) considered application of linear and multiple linear regression models for infilling monthly missing rainfall in the Pirai River Basin, a tributary of the Amazon River.

Although, many approaches have been reported in infilling of missing rainfall information, it is generally believed that no single method can be considered universally best. The choice of a particular approach should account for both topographic and orographic effects of rainfall.

Teegavarapu and Chandramouli (2005) report that even computationally expensive techniques such as kriging may not necessarily result in significant improvement in estimation accuracy. Recognizing that methods like ANN can be computationally expensive may over fit and difficult to interpret, Bennett et al. (2007), preferred simple methods like nearest neighbour by distance, nearest neighbour by correlation, inverse distance methods etc. As such it behaves one to compare both simple methods and computationally expensive methods and make a judicious choice to select an appropriate method for a given basin.

In this study, an attempt has been made to infill missing monthly rainfall information in the agricultural region of the Yarra Catchment, Melbourne district, Victoria, Australia for 3 stations. Due to continuous missing records during the study period considered, the task of infilling is highly challenging. Moreover, the currently operative 43 rain gauge stations in the region are sparsely located, which makes the task further difficult. This study compares simple linear model, polynomial model, logarithmic model and a complex model based on Genetic Programming (GP) to infill the missing monthly rainfalls. GP is chosen in lieu of ANN due to its superiority, particularly in its ability to evolve mathematical models which can be compared with other models. GP has shown to have almost as good modeling accuracy or even more when compared to ANN in many recent studies.

Nomenclature

X_m	measured rainfall
X_s	predicted rainfall
$\bar{X}$	average rainfall
n	total number of training records
R_{An}	two letter code “A” followed by number representing the month for missing rainfall station
R_{Bn},R_{Cn}	two letter code “B”, “C” followed by number representing the month for nearby rainfall station
R_{GR}	Greenvale reservoir station
R_{Mi}	Mickelham station
R_{Si}	Silvan station
R_{Mo}	Monbulk station

2. Study Area and Data Preparation

The Yarra river catchment was selected as the case study catchment because it is a major source of water supply for Melbourne, which is the capital and most populous city in the state of Victoria, and the second most populous city in Australia. The Yarra River travels 245 kilometers and has a catchment area of 4,044 square kilometers. Out of the open 43 stations, in our study the Rainfall stations located in the Agricultural area (3 stations) has been taken for infilling since it is an extensive Flood Plain in Melbourne District. Moreover, for any decision making on irrigation scheduling or related activities the complete set of rainfall information can be used in the future.

In this study, the monthly rainfall time series for 3 stations in Agricultural area from January 1981 to December 2010 for 30 years has been considered. Fig. 1 shows the rainfall gauge stations located in the Yarra River Catchment with stations in agricultural area highlighted with the green colour.

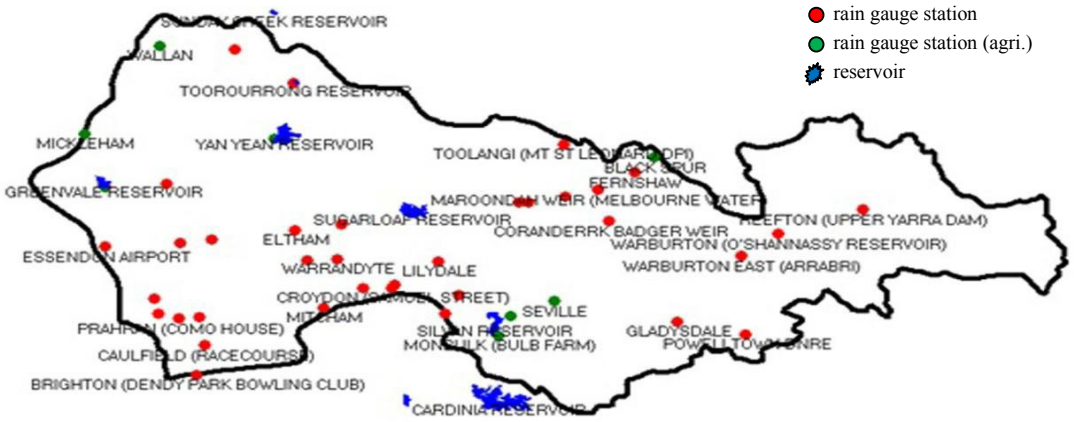

Fig. 1.Location of Rain Gauge stations in Yarra Catchment

The Table 1 summarizes the general locational details as well as rainfall statistics of the selected 3 rain gauge stations. It is seen that Monbulk station has recorded a highest rainfall of 259.3mm when compared to other stations. The Mean monthly average rainfall of the stations lies below 100 mm except Monbulk station which records a value of 101.7 mm. The Annual Average Rainfall of the 3 stations shows that, the rainfall values may exceed 1000 mm at particular stations and since it is an extensive flood plain, it must be considered as an important one in the considerations for Irrigation, Water supply etc.

Table 1.Description and Rainfall Statistics of selected Rain Gauge Stations

Station Name	Mickleham	Monbulk	Silvan
Station Number	86073	86359	86106
Latitude (in degrees)	-37.55	-37.86	-37.83
Longitude (in decimal degrees)	144.88	145.41	145.44
Height of station above MSL	270	235	259
Minimum Rainfall (in mm)	4.6	2	2.6
Maximum Rainfall (in mm)	175.4	259.3	258
Monthly average rainfall (in mm)	46.09	101.77	98.62
Average annual rainfall (in mm)	561.29	1294.78	1198.77
Percentage of missing rainfall	6.11	13.89	21.67
Number of months missing	22	50	78

3. Genetic Programming

Genetic Programming (GP) is very similar Genetic Algorithm (GA), being an evolutionary algorithm based on Darwinian theories of natural selection and survival of the fittest. However, GP operates on parse trees, rather than on bit strings as in a GA, to approximate the equation in a (symbolic form) that best describes how the output relates to the input variables.

The algorithm considers an initial population of randomly generated programs (equations), derived from the random combination of input variables, random numbers and functions, which include arithmetic operators (plus, minus, multiply, divide), mathematical functions (sin, cos, exp, log), logical/ comparison functions (OR/AND) etc., which has to be appropriately chosen based on some understanding of the process and the fitness (a measure of how well solve the problem) of the evolved programs are evaluated; individual programs that best fit the data are then selected from the initial population.

The main operators used in evolutionary algorithm such as GP are crossover and mutation. Besides these some control parameters that need to be set are, Population Size, Maximum Number of Generations and the function set. GP has been implemented using Disciplus Tool.

4. Performance Measure

The infilling performance is evaluated using two criteria viz., the Root-Mean-Square-Error (RMSE) and the Correlation Coefficient (CC).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(X_m\right)_i - \left(X_s\right)_i\right]^2} \tag{1}$$

$$CC = \frac{\sum_{i=1}^{n}\left[\left(X_m\right)_i - \overline{X}_m\right]\left[\left(X_s\right)_i - \overline{X}_s\right]}{\sqrt{\sum_{i=1}^{n}\left[\left(X_m\right)_i - \left(\overline{X}_m\right)_i\right]^2}\cdot\sqrt{\sum_{i=1}^{n}\left[\left(X_s\right)_i - \overline{X}_s\right]^2}} \tag{2}$$

5. Methodology

Genetic programming technique is applied for modeling rainfall infilling by developing individual model for each month. First of all, in the neighbourhood of the missing station, up to three stations are identified which is believed to affect the rainfall information in the missing station. The selection of stations is purely based on the proximity to the missing station. As far as possible, stations from same land use are adopted. The Input and the Output information are represented in the functional form as follows:

$$R_{An} = f\left(R_{Bn}, R_{Cn}\ldots\right) \tag{3}$$

The training, testing, and applied sets for GP are identified for each month of the target station appropriately. Infilling is done for 3 stations namely, Mickleham, Monbulk, and Silvan. Monthly models corresponding to each month of the year is constructed. Using the Disciplus tool, GP is implemented. GP can identify the importance of the given input variables in the modelling process. Consequently, the variables which perform poorly are removed in the subsequent trials to improve the modelling result.

6. Results and Discussions

The analysis of the results for estimation of missing rainfall is presented as below.

Mickleham station has only 6.11% data missing but these form continuous records. Greenvale Reservoir station rainfall is considered as the input and a total data set of 30 years monthly rainfall is used for training (18), testing (9) and validation (3). The CC and RMSE varied over a range of 0.94 – 0.96 and 1 mm – 7 mm for the months from January to December.

The Monbulk station has about 13% data missing. This station is infilled by considering the rainfall data of Silvan. The CC and RMSE varied over a range of 0.95 – 0.99 and 1 mm – 10 mm for the months from January to December. Silvan station is affected by 21.67% of missing data. It is surrounded by 2 stations namely Monbulk and Seville at a distance of 9.5 km and 8.9 km respectively which are used for infilling the missing records. The CC and RMSE varied over a range of 0.94 – 0.99 and 5 mm – 30 mm for the months from January to December.

The typical GP models for selected months are shown below. For example, for the station Mickleham, GP model for the month of March and August are:

March

$$R_{Mi3} = 14.89 + 0.93R_{GR3} \tag{4}$$

August

$$R_{Mi8} = \left(\frac{2.44R_{GR8}\left(\frac{0.14 + R_{GR8}^2}{0.95R_{GR8}} \right)^2}{R_{GR8}\left(\frac{1.04 + R_{GR8}^2}{0.95R_{GR8}} - 3\left(\frac{0.14 + R_{GR8}^2}{0.95R_{GR}} \right)^2 \right)} \right) + R_{GR8} \tag{5}$$

At attempt is made to fit a simple linear model for the month of August which looks like: $R_{Mi8} = 0.9398\ (R_{GR8})$ with a RMSE of 12 mm which is almost same as that obtained from GP model. However, the form of the model is highly complex in GP. A closer analysis indicates that the first term in above GP model for the month of August (which is a non-linear term) has much smaller value when compared to the second term. Thus, in essence, $R_{Mi8} \sim R_{GR8}$ which is what is obtained from linear model. However, the important difference to be noted is that the rainfall process captured by GP model indicates a minor non-linear effect superimposed over the predominant linear effect. GP is able to model this more precisely.

The GP models for Silvan station for the month of January and February are as below.

January

$$R_{Si1} = 0.92R_{Mo1} + 2.23 \tag{6}$$

February

$$R_{Si2} = \frac{\left(\frac{-25.25 + R_{Mo2}}{-24.33} \right)^2 + R_{Mo2}}{1.07} \tag{7}$$

April

$$R_{Si4} = \frac{0.15R_{Mo4}^3 + R_{Mo4}^2 - 0.5R_{Mo4} - 0.6}{R_{Mo4}^2 + 15.4R_{Mo4} - 3} \tag{8}$$

The model obtained for April is typically non-linear and very well depicts the process. The plot comparing missing rainfall and GP generated is shown in Figure 2 and 3.

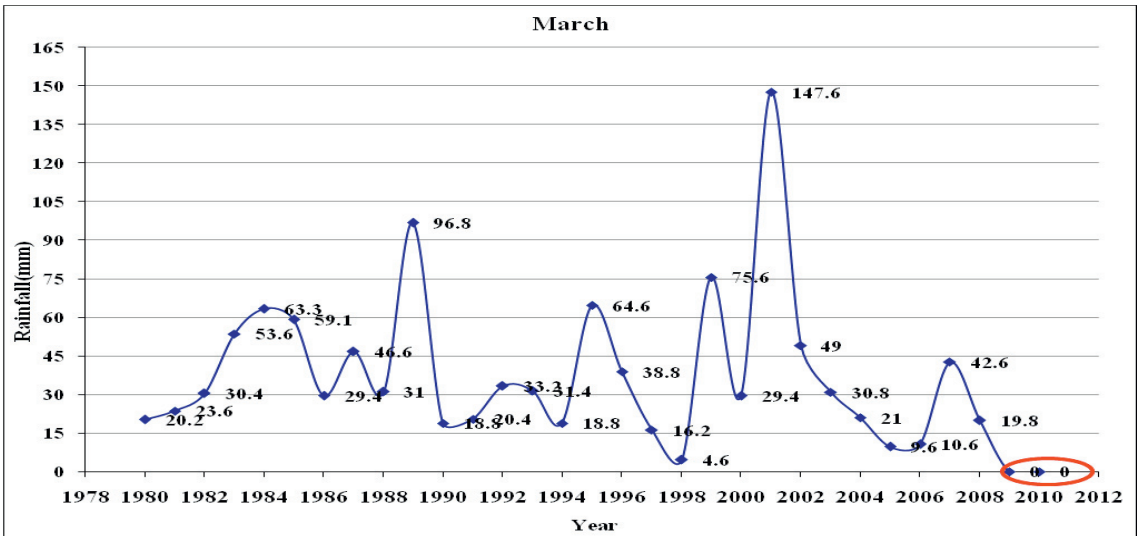

Fig. 2. Missing rainfall for Mickelham Raingauge station (March).

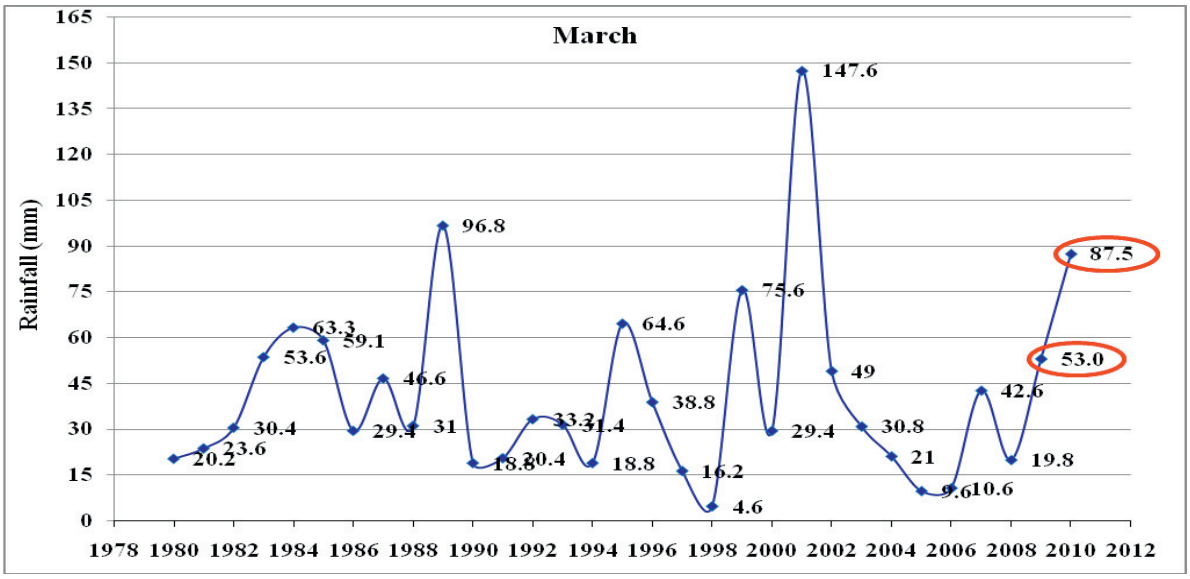

Fig. 3. GP generated rainfall for Mickelham Raingauge station (March).

7. Conclusion

Based on this study, the following conclusions are arrived:

1. Monthly models, even with limited rainfall information, seem to be more potential approach.
2. GP is able to detect the subtle non-linear effect superimposed over the linear behaviour.

References

Bennett, N,D., Newham, L,T,H., Croke, B,F,W., Jakeman, A,J., 2007. Patching and Disaccumulation of Rainfall Data for Hydrological Modelling and Simulation (MODSIM 2007), ed. Les Oxley & Don Kulasiri, Modelling and Simulation Society of Australia and New Zealand Inc., New Zealand, 2520-2526.

Coulibaly, P., Evora, N, D., 2007. Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology 341, 27-41.

De Silva, R.P., Dayawansa, N,D,K., Ratnasiri, M.D., 2007. A Comparison of methods in estimating Missing Rainfall Data. The Journal of Agricultural Sciences, Samaragamuwa University of Sri Lanka, pp 101-108. vol.3, no.2 pg 101-108.

Ilunga., 2010. Infilling annual rainfall data using feed forward back-propagation Artificial Neural Networks (ANN): Application of the standard and generalised back propagation techniques. Journal of The South African Institution Of Civil Engineering, Vol.52, No 1, 2010, Pages 2–10, Paper 663.

Mauricio F.Villazón., Patrick Willems., 2010. Filling gaps and Daily Disaccumulation of Precipitation Data for Rainfall-runoff model. BALWOI, Ohrid, Rep of Macedonia, 25- 29.

Teegavarapua, R., Chandramoulia, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology 312, 191-206.