20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Filling gaps in daily rainfall data: a statistical approach

M.M. Hasan[a], **B.F.W. Croke**[a,b]

[a] Integrated Catchment Assessment and Management (iCAM) Centre, and the National Centre for Groundwater Research and Training, The Fenner School of Environment and Society, The Australian National University, Canberra, Australia
[b] Department of Mathematics, The Australian National University, Canberra, Australia
Email: masud.hasan@anu.edu.au

Abstract: Daily rainfall data are one of the basic inputs in hydrological and ecological modeling and in assessing water quality. However, most data series are too short to perform reliable and meaningful analyses and possess significant number of missing records. The study focuses on developing a methodology to fill the gaps in daily rainfall series considering data of twenty rainfall stations from Brahmani Basin, Rachi, India. A probabilistic approach is adopted to generate data for filling on missing points.

The Poisson-gamma (PG) distributions were explored in the study as they possess useful properties to simultaneously model both the continuous (rainfall depth) and discrete (rainfall occurrence) components of daily rainfall. First, the PG distributions were fitted to the daily rainfall data of targeted stations and the parameters were estimated. The models were compared with the widely used inverse distance interpolation method. To compare the fit of the models, a dataset of size equal to the size of the observed dataset were generated. The means and percentages of days with no rainfall of observed and simulated datasets were very similar. However, PG distributions slightly overestimate the 95th percentile and underestimate the variance and 99th percentile. This indicates that the models do not capture well the extremely heavy rainfall events; hence, the PG distributions need to modify to capture better the extreme events. However, with respect to all statistics, the PG model performs better than the inverse distance interpolation method.

The methodology considers two basic assumptions.

- The rainfall data of missing period have similar statistical properties to the data from available periods. Fairly large amounts of data exist to generalize the parameters from the available periods to the points with no data. The assumption is also supported by the fact that, for the studied stations, the first and second halves of the available datasets possess similar statistical properties.
- Spatial correlations exist among rainfall occurrence and amounts of neighboring stations. The fact is reasonable as fairly negative relationship were observed between correlation of daily rainfall and distances among the studied stations.

Once the PG distributions were decided, samples were generated with the parameters of respective stations. The generated data for a station is completely random in nature and independent of the rainfall amounts of neighboring stations. To match the data, first the rainfall amount of the region is estimated as the weighted mean of rainfall amounts from four closest stations. Weights were taken as the inverse of the distances of the neighboring stations from the target station. Days were sorted from driest to wettest on the basis of the mean rainfall amounts of neighboring stations, and finally, the generated data were matched.

Instead of using two separate models for generating continuous data (rainfall depth) with exact zero (no rainfall), the proposed method use a single model to model both components of daily rainfall simultaneously. The method resolves the problem of overestimating non-zero rainfall amount that arises while using traditional interpolation methods. However, the method may not work well when the neighboring stations are not close to the target station.

Keywords: *Daily rainfall, Poisson Gamma (PG) distribution, interpolation*

1. INTRODUCTION

Daily rainfall data are one of the basic inputs in hydrological (e.g. streamflow, rainfall-runoff, recharge) and environmental (e.g. crop yield, drought risk) models and in assessing water quality. However, most daily rainfall data series are too short to perform reliable and meaningful analyses and possess significant number of missing records (Elshorbagy *et. al.*, 2000; Bennett *et. al.*, 2007; Kajornrit *et. al.*, 2012). Filling the gaps in daily rainfall data is therefore a crucial issue.

Spatial interpolation techniques are widely used methods for filling the gaps in daily rainfall series through estimating the unknown rainfall amount for a point from the known data from adjacent stations (Wei and McGuinness, 1973; Burrough and McDonnell 1998). Paulhus and Kohler (1952) explored two methods of interpolation, the normal-ratio and 3-station-average, to fill the missing values in monthly rainfall data. The Inverse Distance Weighting (IDW) methods estimate the rainfall amount of a location as a weighted average of the rainfall amount of adjacent stations and the weights are considered as a function of the distances (Teegavarapu and Chandramouli, 2005). Various IDW methods have been developed on the basis of the functional form of the distances. For instance, inverse of squares, higher powers or exponential of distances (Garcia *et al*., 2008; Ly et al., 2011; Chen and Liu, 2012) are used as weights. The correlation coefficients between data series are also explored to estimate the weights (Teegavarapu and Chandramouli, 2005; Ahrens, 2006). Spatial interpolation methods yield non-zero rainfall amounts for a station if even just one of the neighboring stations has rainfall, and hence, due to the poor sampling by rainfall gauge networks, tend to significantly overestimate the number of rainy days. Moreover, the errors in estimating the missing records due to the faulty measurement process of rainfall at neighbouring stations can't be ignored (Teegavarapu 2009).

Regression based methods are also used for estimating missing precipitation values (Presti *et. al.*, 2010**).** Regression models consider climate data, elevation, topography, proximity to coastal area etc. as explanatory variables to estimate missing rainfall series of a station (Daly *et. al.,* 1994). Some regressive techniques explored in filling missing daily rainfall series include: simple substitution, parametric regression, ranked regression, and the Theil method (Presti *et. al.*, 2010). In addition to spatial interpolation and regression methods, neural network algorithms are also explored for imputation of missing precipitation values (Malek *et al*., 2009). The neural network algorithms adapt the weighted interpolation technique from neighbouring stations. Regression based methods also underestimate the number of days with no rainfall (Simolo *et. al.,* 2010).

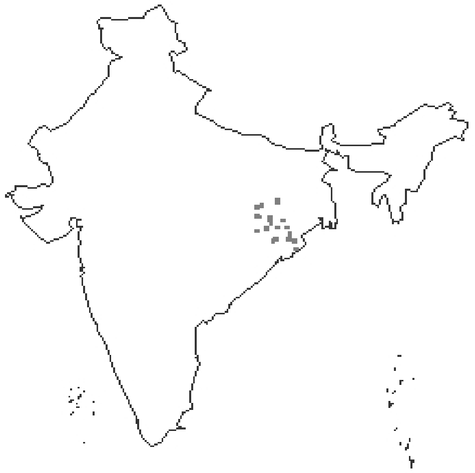

Figure 1. Location of studied stations

Simolo *et. al*., (2010) proposed a two steps probabilistic approach to fill the missing values in rainfall series. First, weighted average of rainfall amounts of neighbouring stations is obtained and the days with rainfall amount below a threshold is considered as dry. The threshold is estimated on the basis of the probability of no rainfall in the original data series. Then the rainfall amount of wet-classified days is estimated by multivariate regression considering the rainfall series form the surrounding stations as explanatory variables. Two parameter Gamma distributions were fitted to the original rainfall series. Finally, to correct for the bias induced by fit in multivariate regression, the generated values are forced to satisfy the daily probability distribution associated with the original series.

To avoid the use of an arbitrary threshold (as in Simolo *et al.*, 2010), this study focuses on filling the missing values of daily rainfall series with data generated from the appropriate probability distributions and parameters estimated from available rainfall data of respective stations. The generated data are reorganized on the basis of the wetness condition of the region measured as the weighted average of the rainfall from adjacent stations. The Poisson-Gamma (PG) distributions were adapted here to model occurrence and amount of daily rainfall simultaneously.

The method has the advantages that it explores the various statistics and the probability distribution of rainfall of the target station. The method also resolves the problem of overestimating non-zero rainfall amount that arises while using traditional interpolation methods.

2. DATA AND STUDY REGION

Daily rainfall data of twenty stations from Brahmani Basin, Rachi, India (Croke *et. al.*, 2011) were considered in this study (Figure 1). The basin is located between longitude 83.97°E and 86.60°E and latitude from 20.10°N to 23.42°N. Data ranges from 1st of January, 1969 to 31st of August 2004. However, none of the stations have data for the entire period, with about half of the studied stations having less than fifty percentage coverage due to gaps in the data (Figure 2).

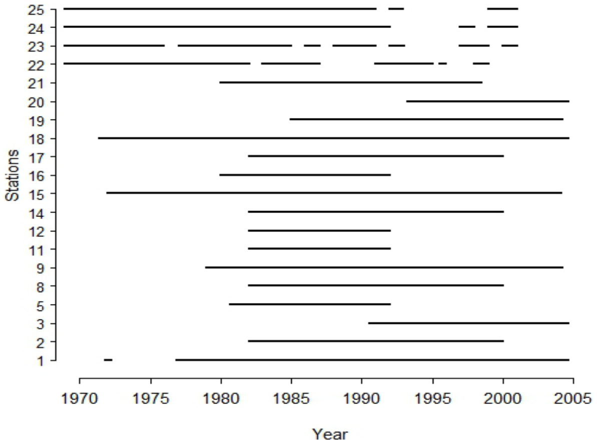

Figure 2. Data coverage for studied stations. Gaps in the line indicate no available data for the period.

3. METHODS

The aim of the study was to develop a probabilistic methodology to fill the gaps in daily rainfall data. Instead of simple, direct interpolation methods, statistical properties (distribution, average and dispersion) were used to generate data for missing points. One of the major challenges in generating daily rainfall data is that, the data are highly right skewed (long tail to the right) and continuous with lots of exact zeros (Table 1 and Figure 3). For better visualization, Figure 3 is constructed with only four stations; however the other stations have very similar properties in the context of skewness of the data.

For filling missing values in the rainfall series of a station, two basic assumptions were made:

- The missing rainfall values have similar statistical properties to the data from available periods. Fairly large amounts of data exist to generalize the parameters from the available periods to the points with no data. This assumption is also supported by Table 1**,** which indicates, almost everywhere, the first and second halves of the available datasets have similar properties.
- Spatial correlations exist for rainfall occurrence and amounts between neighboring stations. The fact is supported by Figure 4 which indicates fairly negative relationship between correlation of rainfall amounts and distances among the stations.

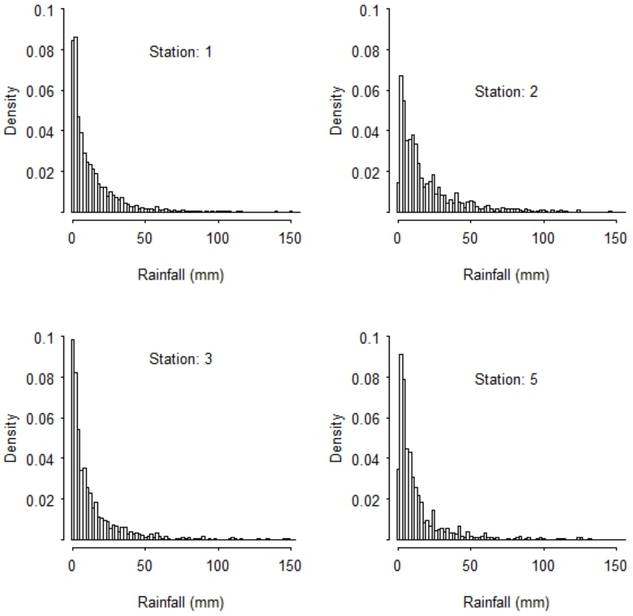

Figure 3. Distribution of non-zero daily rainfall of four stations

To fill the gaps in rainfall series, first, appropriate probability distribution of daily rainfall data of studied stations were decided. Daily rainfall data are highly positively skewed (have a long tail at the right) and have lots of exact zeros (days with no rainfall). Hasan and Dunn (2010, 2011) showed that Poisson-Gamma (PG) distributions fit well to the monthly rainfall data from Australian stations. The distributions prove to be adequate for simultaneously modelling both the continuous (rainfall depth) and discrete (rainfall occurrence) components of daily rainfall. The similar approach is adopted here for daily rainfall data.

The PG is a three parameter distribution and the parameters are mean μ, dispersion parameter ϕ, and index parameter p. The parameters (μ, ϕ and p) of PG distribution of each station were estimated from the available datasets.

Table 1. Various statistics of first and second halves of the data series

	Mean		CV		95th percentile		99th percentile		% days no rainfall	
	First	Last	First	Last	First	Last	First	Last	First	Last
Station 1	3.9	4.1	291.5	320.9	24.1	24.8	55.2	53.8	69.6	69.3
Station 2	3.8	3.1	323.9	356.7	25.0	22.0	63.0	56.3	81.2	83.1
Station 3	3.9	4.0	301.2	283.6	24.4	23.7	55.1	58.0	69.0	67.1
Station 5	3.2	2.9	348.4	365	18.9	15.9	57.4	48.4	78.6	77.4
Station 8	4.1	2.7	341.8	392.6	26.0	17.0	70.3	53.0	80.7	85.8
Station 9	4.3	4.2	312.5	328.5	26.0	25.5	62.1	62.0	67.6	71.2
Station 11	4.3	3.9	330.1	337.7	25.0	24.8	69.0	69.0	80.5	80.8
Station 12	3.9	2.4	317.7	332.4	25.0	14.0	66.5	39.5	81.3	80.5
Station 14	3.4	3.6	393.7	410.0	20.0	25.0	57.5	68.2	81.8	85.4
Station 15	3.4	3.8	322.1	330.0	20.5	23.6	55.5	59.0	71.5	70.8
Station 16	4.0	3.7	283.6	290.8	24.0	21.6	59.1	51.7	70.3	71.9
Station 17	2.5	2.3	356.7	358.4	17.0	15.0	42.0	44.2	84.2	84.3
Station 18	4.1	3.8	285.2	302.4	24.5	21.5	57.0	54.9	68.7	65.0
Station 19	4.0	3.7	316.9	324.4	24.0	22.6	62.0	64.2	71.4	72.5
Station 20	4.2	3.8	274.1	282.3	24.9	23.5	56.8	53.2	67.4	70.6
Station 21	4.0	3.8	326.0	339.3	23.4	23.0	64.3	63.2	78.3	76.6
Station 22	4.5	4.8	290.0	270.5	27.8	30.0	64.7	58.8	74.6	72.7
Station 23	3.5	3.7	332.0	352.8	21.4	23.1	60.9	65.0	76.7	79.9
Station 24	4.4	4.3	385.4	458.9	27.0	23.1	73.6	71.0	75.7	72.3
Station 25	4.4	4.3	387.5	358.2	24.0	25.2	66.1	65.9	74.8	72.2

For generating daily rainfall data for each station the parameters of the PG distribution were estimated using the available dataset of the station. Using these parameters, a set of random numbers were generated to fill the missing values of respective stations using the following procedure.

Let station 'A' has *N* missing observations. A dataset of size *N* is generated from PG distribution with parameters estimated from available dataset of station 'A'.

The generated data (in previous step) for a station is completely random in nature and independent of the rainfall amounts of neighboring stations. To match the data, first the rainfall amount of the region is estimated as the weighted mean of rainfall amounts from four closest stations. Weights were taken as the inverse of the distances of the neighboring stations from the target station. Considering six different interpolation methods, Teegavarapu *et. al.*, (2011) concluded that, inverse distance based on four nearest neighbors was the best for the transformation of data.

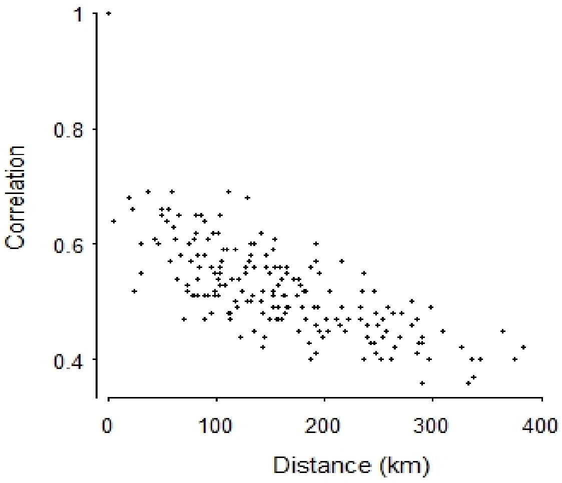

Figure 4. Scatterplot showing the relation between distance of the rainfall stations and their correlation

Days were sorted from driest to wettest on the basis of the mean rainfall amounts of neighboring stations. Sorted (smallest to largest) randomly generated data are then matched with the sorted days. Finally, days with the missing rainfall data were filled with the generated data

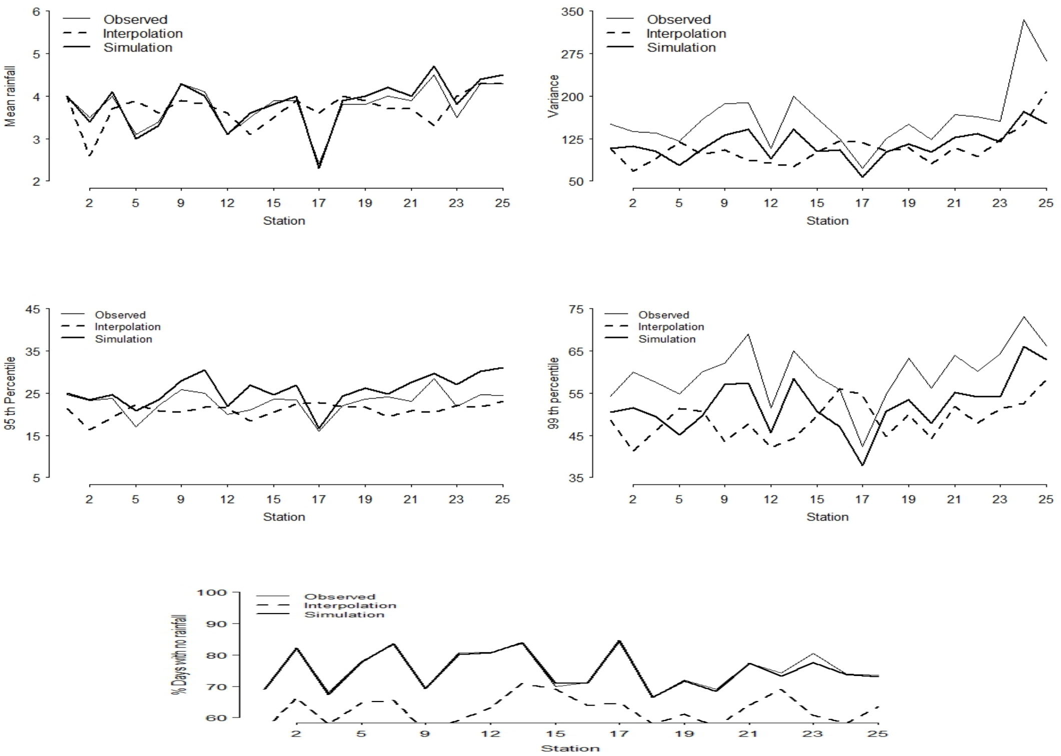

Figure 5. Line plots comparing mean, variance, 95th percentile, 99th percentile and percentage of days with no rainfall for modelled validation and simulated data.

4. RESULTS AND DISCUSSION

The various statistics, such as the mean, variance, 95th percentile, 99th percentile and percentage of days with no rainfall of observed, interpolated and generated datasets were presented by line plots in Figure 5. From the plots it is evident that the model generate data with very similar properties of observed data with respect to the mean and percentage of days with no rainfall. Whereas, for interpolated data the mean differ slightly and the probability of no rainfall differ significantly from observed dataset. The PG distributions underestimate the variability in daily rainfall data. The simulated data slightly overestimate the 95th percentile and underestimate the 99th percentile. However, the models capture well all the statistics than the interpolated datasets. The PG models do not capture well the extremely heavy rainfall events; hence, PG distributions need to be modified to capture better the extreme events at a daily time scale. Notably, the models did not use any predictor variable and the inclusion of predictor variables into the modeling framework may improve the performance of the simulations. The PG distributions belong to the EDM family and hence the predictor variables may be included into the generalised linear modelling framework.

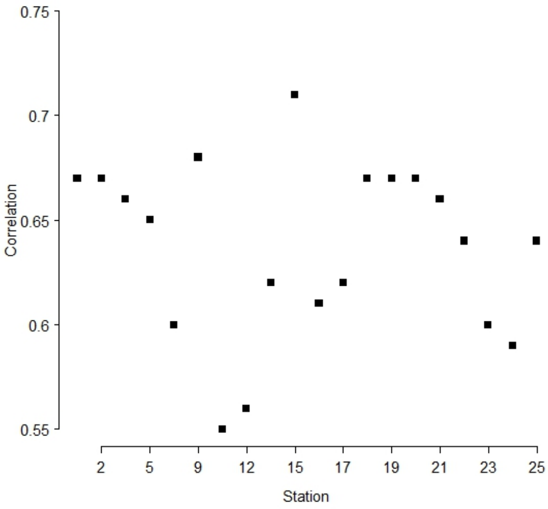

Figure 6. Dot plots representing the R-square values from generated and observed data for studied

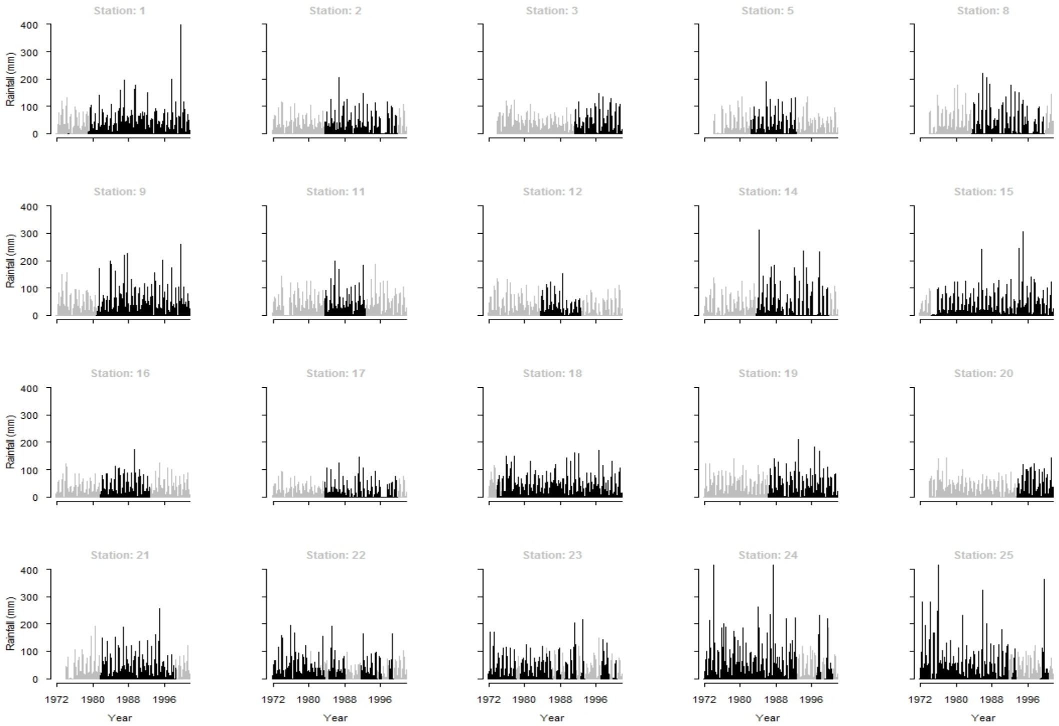

Figure 7. Plot showing the available and generated data for studied stations

To check how the method works in filling the gaps in the daily rainfall data, a timeframe is selected where the stations have available datasets. For the studied stations, data were generated for the period. The generated data were compared with the available datasets using spearman's correlation coefficient. The Spearman's correlation coefficients were used considering the non-normal nature of the data. However, the statistics are based on the ranked data and may be affected by lots of zeros in the data series. For the studied stations, the correlation coefficients range from 0.55 to 0.71 (Figure 6) indicating fairly good relationships between the simulated and observed datasets. Lack of fitting may be due to matching the data on the basis of the neighboring stations with poor correlation between rainfalls of neighboring stations at a daily time scale. The mean and maximum distances to the closest stations from the targeted station are 33.57 km and 65.29 km respectively. The interpolation is done with rainfall data from stations up to 154.58 km from the targeted station. A better interpolation and data match may be achieved with data from closer stations (if available). Finally, the observed available data with the generated infill missing data were plotted (Figure 7). The observed available data sets were presented by black bars and grey bars represent the filled data. With an overview of the plots we can conclude the method fill the missing values fairly well, however fails to produce some extremely large rainfall amounts.

5. CONCLUSIONS

This article discusses a hybrid method (a probabilistic method for data generation and interpolation method for matching the data points) for estimating missing data in daily rainfall series. Information regarding the statistical properties of historical rainfall amounts of targeted station and the rainfall amount of neighbouring stations were considered in the process. The missing rainfall days were filled with the data generated from the appropriate distribution and parameters of the targeted stations. The main focus here is to correct for common bias, such as the overestimation of the number of rainy days which affect traditional models. The method also minimises the errors in estimating the missing records due to the faulty measurement process of rainfall at neighbouring stations. The PG model generates data with very similar properties to the observed dataset with respect to the probability of no rainfall, mean rainfall amounts and 95th percentiles. However, the method underestimates the variability in the rainfall data and so can't capture well the extremely large rainfall events. Hence a modification in the estimation procedure of PG parameters may be required. Moreover, the data matching procedure may not work well while considering data points for interpolation those are far from the targeted rainfall station.

ACKNOWLEDGMENTS

This work was funded by the National Centre for Groundwater Research and Training, a collaborative initiative of the Australian Research Council and the National Water Commission. The data used in this study was obtained through a project funded by the Australian Centre for International Agricultural Research.

REFERENCES

Ahrens B. 2006. Distance in spatial interpolation of daily rain gauge data. *Hydrology and Earth System Sciences* 10: 197-208.

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

Burrough, P.A. and McDonnell, R.A. (1998) Principles of geographical information systems. Oxford University Press, Oxford.

Chen, F.W. and Liu, C.W. (2012) Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy & Water Environment,* 10(3): 209-222.

Croke, B.F.W., Islam, A., Ghosh, J. and Khan, M.A.(2011)Evaluation of approaches for estimation of rainfall and the unit hydrograph. *Hydrology Research*, 42(5): 372-385.

Daly, C.**,** Neilson, R.P., and Phillips, D.L. (1994) A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain. *Journal of Applied Meteorology*, 33(2): 140-158.

Elshorbagy, A.A., Panu, U.S. and Simonovic, S.P. (2000) Group-based estimation of missing hydrological data: I. Approach and general methodology. *Hydrological Sciences,* 45(6): 849-866.

Garcia M, Peters-Lidard CD, Goodrich DC. 2008. Spatial interpolation in a dense gauge network for monsoon storm events in the South western United States. *Water Resources Research* 44: W05S13, DOI: 10.1029/2006WR005788.

Hasan, M.M. and Dunn, P.K. (2010) A simple Poisson–gamma model for modelling rainfall occurrence and amount simultaneously. *Agricultural and Forest Meteorology,* 150(10): 1319-1330

Hasan, M.M. and Dunn, P.K. (2011) Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia. *International Journal of Climatology*, 31: 1389-1397. doi: 10.1002/joc.2162

Kajornrit, J., Wong, K.W. and Fung, C.C. (2012) A Comparative Analysis of Soft Computing Techniques Used to Estimate Missing Precipitation Records. Proceedings of the 19th ITS Biennial Conference 2012 Bangkok, Thailand.

Ly, S., Charles, C., and Degre, A. (2011). Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium. Hydrology and Earth System Sciences, 15: 2259-2274.

Malek, M.A., Harun, S., Shamsuddin, S.M., and Mohamad, I. (2009). Reconstruction of Missing Daily Rainfall Data Using Unsupervised Artificial Neural Network. *International Journal of Electrical and Computer Engineering*, 4: 340-345.

Paulhus, J.L.H. and Kohler, M.A. (1952) Interpolation of Missing Precipitation Records. *Monthly Weather Review,* 80(8): 129-133.

Presti, R.L, Barca, E. and Passarella, G. (2010). A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental Monitoring and Assessment,* 160:1-22.

Simolo, C., Brunetti, M., Maugeri, M., and Nanni, T. (2010) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *International Journal of Climatology,* 30: 1564-1576.

Teegavarapu, R. S. V. & Chandramouli, V. (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology, 312: 191-206.

Teegavarapu, R.S.V. (2009) Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. *Journal of Hydroinformatics*, 11(2): 133-146.

Teegavarapu, R.S.V., Meskele, T., and Pathak, C.S. (2011) Geo-spatial grid-based transformations of precipitation estimates using spatial interpolation methods. Computers & Geosciences, 40: 28-39.

Wei, T.C. and McGuinness, J.L. (1973) Reciprocal Distance Squared Method, A Computer Technique for Estimating Area Precipitation. *Technical Report ARS-Nc-8. US Agricultural Research Service*, North Central Region, Ohio.