Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 24 (2016) 31 – 38

Procedia Technology

International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

Daily Rainfall Prediction using Generalized Linear Bivariate Model - A Case Study

Jany George [a]*, Letha J. [b], Jairaj P. G. [c]

[a] *Research Scholar, Dept. of Civil Engineering, College of Engineering Trivandrum, Thiruvananthapuram, 695016, India*
[b] *Professor in Civil Engineering, Vice Chancellor, Cochin University of Science and Technology, Kochi, 682022, India*
[c] *Professor in Civil Engineering, College of Engineering Trivandrum, Thiruvananthapuram, 695016, India*

Abstract

The present study focuses on the simulation of daily rainfall series based on atmospheric predictors and historical data using a bivariate Generalized Linear Model. Temperature and precipitation data along with a set of covariates were made use of in generating the simulations. Probability of occurrence of rainfall was predicted using logistic regression models. The amount of rainfall on a rainy day was modelled using a gamma distribution. The covariates in the model comprise of different categories such as site effects for spatial variation, year effects allowing long term trends, month effects for seasonality, day effects with temporal auto correlation and atmospheric predictors. Rainfall series were generated for both future and past periods at multi sites simultaneously using atmospheric predictors. The model developed was applied in a typical catchment in the state of Kerala in India. The model simulations were acceptable on the basis of the performance evaluated using statistical analysis. The model can be used as a weather generator to simulate the daily rainfall series for both past and future periods.

Keywords: Daily rainfall; Generalized Linear Models; atmospheric predictors; weather generator

1. Introduction

Water resources planning and management always rely on historic, current, and future states of precipitation on yearly monthly and daily timescales. Prediction of daily precipitation is a challenging task in water resource

* Corresponding author. Tel.: +91 9895733847
*E-mail address:*jany_george@yahoo.com

doi:10.1016/j.protcy.2016.05.006

management. Rainfall stands as one of the most unpredictable event even with the updated climatic models. The extent of uncertainty varies from model to model depending on the physical phenomenon of atmospheric conditions and the complexity associated with its mathematical modelling. Statistical models contribute to a great extent to reduce this uncertainty.

Even though General circulation Models(GCM) are regarded as the most credible tools to provide information of the atmospheric circulation systems on a global scale, the climatic variables like precipitation cannot be well modelled by the GCMs [1]. The common practice is to downscale the results from the GCMs either by dynamic downscaling through a nested high-resolution regional climate model (RCM) or through statistical downscaling [2]. Statistical downscaling can be classified into three types, weather generators, weather typing and transfer function [3]. Most of the works on downscaling were done on monthly timescales. Disintegration of the monthly values into daily time series is yet another challenging task.

Attempts have been made previously to predict the daily rainfall based on different methodologies. A multivariate downscaling model for precipitation and temperature scenarios based on atmospheric circulation indices [4]; simulation process coupled with atmospheric circulation patterns [5,6]; Markov process and gamma distribution [7] are some of the reported works in this area. As per the Intergovernmental Panel of Climate Change (IPCC) definition, a stochastic weather generator simulates time series of weather data for any location on the basis of the statistical characteristics of the historical weather at that particular location [8]. The first weather generator WGEN is introduced by Richardson [9] for generating the daily weather sequences. In a typical weather generator, the idea of Markov chain is used for precipitation occurrence with the transition matrix giving the probability of occurrence of rainfall. If the model predicts the day as rainy, the amount of rainfall on that rainy day is computed by gamma distribution with separate parameters of gamma distribution for each month of the year. The drawback of the Richardson-type weather generator is that it fails to explain adequately the length of dry and wet series [10]. Also these weather generators have a tendency to underestimate variability of seasonal means or totals. Another drawback is the underestimation of high return levels [11]. To improve the performance of weather generators, different approaches have been suggested which includes higher-order Markov chains, heavy-tailed intensity distributions, nonparametric modelling, approaches based on spell lengths, Generalized linear models (GLM) and Sub-daily weather generators. The Generalized linear models (GLM) comprise of classical structure weather generators within a flexible framework that permits many extensions to basic model structure [12].

In the present study, daily rainfall series were simulated using a weather generator based on Generalized Linear model in the catchment of Idukky reservoir in Kerala. The simulations were generated for the missing period from 1981 to 1995 using the atmospheric predictors derived from the National Centre for Environmental Prediction/ National Centre for Atmospheric Research (NCEP/NCAR) re-analysis data and also for a future period of 2013 to 2025 using the predictors derived from the Coupled Global Climate Model CGCM3 for IPCC SRES A2 scenario. The methodology adopted, model application and results of the study are discussed in the subsequent sessions.

2. Methodology

Daily rainfall is simulated using Generalized Linear Models for occurrence and intensity of rainfall and for temperature in three stages. A GLM, for a *nx1* vector of random variables $Y = (Y_1, Y2 \ldots, Y_n)$, is a model for the probability distribution generating Y [13]. Each of the Y s depends on p covariates, whose values are arranged in a *nxp* matrix X. The distribution of Y has the mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$, which is related to X as in Eqn. (1).

$$\boldsymbol{g(\mu_i) = X_i\beta = \eta_i} \tag{1}$$

The daily rainfall Y_i is assumed to be generated in the form of Eqn. (1) from the same family of distribution with mean μ_i with $g(\mu_i)$ as the link function and β a *px1* vector of coefficients. The members of η are called linear predictors. The distribution of each Y_i belongs to the exponential family, the family of all distributions with density function in the form of Eqn. (2) for some parameters ψ and Φ and functions *a*(:), *b*(:) and *c*(:) [14].

$$f(y;\psi,\phi) = \exp\left[\frac{y(\psi - b\psi)}{a(\phi)} + c(y,\phi)\right] \tag{2}$$

The parameter *ψ* determines the mean of the distribution and the parameter *Φ* determines the variance. In a GLM, the prime importance is in the development of the relationship between the mean and the covariates. The parameter *ψ* is a function of the covariates and of the coefficient vector *β*. The dispersion parameter *Φ* is usually assumed constant in a GLM [14]. The GLM is composed of the three elements, a choice of distribution with mean μ_i, a linear predictor $\eta=X_i\beta$, a linear combination of unknown parameters *β*, and a link function $g(\mu_i)$. Expected value of Y_i is μ_i which is related to X_i through the relationship in Eqn. (1). In this GLM, a logistic regression model is used to predict the probability of a day to be dry or wet [15]. If p_i is the probability of rainfall on the ith day, and X_i is the corresponding predictor vector or covariate vector of atmospheric circulation pattern and *β* is the coefficient vector, then the probability of rainfall occurrence is given by Eqn. (3).

$$ln\frac{p_i}{(1-p_i)} = X_i\beta \tag{3}$$

The amount of rainfall on a rainy day is modelled with the help of gamma distribution. Gamma distribution is regarded to be most appropriate to model for the daily rainfall amount generation [16]. Let μ_i denote the intensity of rain fall at a site on the ith day. If z_i denote the predictor vector based on atmospheric circulations, the mean rainfall at the site on the ith wet day is given by Eqn. (4).

$$ln(\mu_i) = Z_i\alpha \tag{4}$$

In a multivariate model, the *nx1* predictor vector changes to *nxm* matrix of *Y*. The columns represent the observations of a different variable such that Y_{ij} represents the value of the jth variable for the ith case in the dataset. If the distributions of the individual variables conditioned on covariates are all Gaussian, a regression model for all of the variables can be developed simultaneously. Multivariate modelling can be described using the standard factorization of the joint distribution. If $Y_i = (Y_{i1}, Y_{i2},.., Y_{im})$ is the collection of all variables for the ith case in the dataset, and let X_i be the associated vector of covariates. Then the joint density of Y_i can be factorized as in Eqn. (5).

$$f(Y_i|X_i) = f_1(Y_{i1}|X_i) \times f_2(Y_{i2}|Y_1, X_i) \times \dots \times\ f_m(Y_{im}|Y_1, Y_2, \dots, Y_m, X_i) \tag{5}$$

where *f*(*a*|*b*) denotes the density of *a* conditional upon the values of *b* [14]. The GLM is developed for the first variable, conditioned upon the covariates, then for the second variable conditioned on the covariates along with the first variable until all variables have been modelled. In order to link the precipitation and temperature to produce a bivariate model that describes the dependence between the two variables, one of the variables is assigned as a covariate in the model for the other. In the bivariate model, R_t and T_t represent precipitation and temperature for all sites on day t, then the joint probability density function of R_t and T_t can be represented as in Eqn. (6).

$$f_{R,T}(r,t) = f_T(t) \times f_{R|T}(r|T = t) \tag{6}$$

The parameter estimation of the model is performed through the method of Maximum Likelihood estimation. If a data vector $y = (y_1, y_2, \dots, y_n)$, drawn from some family of distributions having a parameter vector *θ*, this parameter vector estimated by the method of maximum likelihood by choosing the value of *θ* which allocates highest probability to the observations *y*. If *f*(*y; θ*) represents the joint density of *y*, the Likelihood for *θ* given *y* is given by

$$L(\theta|y) = f(y;\theta) \tag{7}$$

If the observations are independent, then their joint density and likelihood functions are given by Eqn. (8) and (9).

$$f(y;\theta) = \prod f_i(y;\theta) \tag{8}$$

$$\ln(f(y;\theta)) = \sum_{i=1}^{n} f_i\ (y;\theta) \tag{9}$$

Atmospheric predictors used for driving the models should represent the circulation patterns. The circulation indices are derived from atmospheric variables sea level pressure, u-wind and v-wind components, air temperature and relative humidity provided by the NCEP/NCAR re-analysis data [17] for both model building and validation period. Dependence between the sites and also the interaction between the predictors are taken into account in developing the models. The covariates are from different categories namely site effects, year effects, month effects, day effects and external effects representing spatial variation, long term trends, seasonality, day to day temporal auto correlation and external atmospheric predictors respectively [12].

3. Model Application

3.1. Location and Hydrology

The methodology was applied to the catchment of Idukky reservoir in Kerala which is having a storage capacity of 2000Mm^3 and situated at an elevation of 701 m from the mean sea level. It is located in the Westernghats mountain region that separates the state of Kerala from Tamilnadu. The reservoir at Idukky is formed by the construction of three dams, the Idukky arch dam across the Periyar River, the Cheruthoni dam on the Cheruthoni River, and the Kulamavu dam on the river Kilivallithodu. The three dams together form a reservoir of surface area of about 60 km^2. The river Periyar drains water from a catchment area of 526 km^2 to the reservoir at Idukky. Catchment area of the river Cheruthoni 123 km^2, both rivers together contribute from a total catchment area of 649 km^2 to the Idukky reservoir. The geographical extent of the catchment ranges from 9^0 30' N to 9^0 55' N and 76^0 50' E to 77^0 30'E. The catchment receives heavy rainfall, an average annual rainfall of 3500 mm, more than 90% of it is contributed by the South West (SW) and North East (NE) monsoons from June to November. The Western Ghats Mountains located on the eastern boundary of Kerala enables an orographic lifting of the monsoon winds which results in heavy rainfall during the SW monsoons over the western slopes and good rain over the other parts. NE monsoons also contribute to the annual rainfall, mainly in the southern parts of the Kerala [18]. The meteorology of Kerala is profoundly influenced by its orographical features. From the low-lands on the western sea coast, the land profile rises towards the east to the mid-lands and further towards the high-lands of the Western Ghats. The mountain ranges having elevation up to 2 km forms a natural wall separating Kerala from the adjoining state [19].

3.2. Data

Daily Rainfall and temperature data from six different stations around Idukky reservoir is collected from Kerala State Electricity Board research division for a period of 1981 to 2013. The stations are Idukky Vazhathoppu, Kulamavu, Ayyappankovil, Udumbanchola and Kattappana. The data contains missing values up to 10% except for Idukky Vazhathoppu station for which there were no missing values. Atmospheric variables namely mean sea level pressure, maximum air temperature, u-wind and v-wind components, wind speed, dew point temperature are derived from NCEP/NCAR re-analysis data. Re-analysis data has been obtained from a high resolution atmospheric model with data assembled from surface observation stations, upper-air stations, and satellite-observing platforms. Outputs obtained using these fields represent those that could be expected from an ideal GCM [20]. Monthly average values of the climatic variables on a grid of 2.5^{0}x2.5^0 for the period 1981-2013 is used for developing the model. The validated model is used to simulate the missing rainfall series at five stations and at Idukky station using the monthly atmospheric predictors derived from the NCEP/NCAR re-analysis data. The future rainfall series at the six stations are simulated with the atmospheric predictors derived from the global climatic model CGCM3 SRES A2 scenario.

3.3. Model Development

The present work makes use of the RGLIMCLIM package installed in R programming environment for multi variate, multi site weather generators using GLMs for simulating the daily rainfall series. The aim of the work is to reproduce some subset of the time series considering the marginal aspects such as mean and variance, temporal aspects such as trends, seasonality and auto correlation, spatial aspects such as regional variation and inter site relationships specifying inter-variable relationships [12]. The station information such as station names, location, latitude and longitude are composed into a file format which is read together with the definition files. The rain fall and temperature with station names compiled in specific formats and the external file for atmospheric predictors are the inputs for model fitting at each stage.

Starting from the basic simple model, a valid definition file is created in each step by adding each component of covariates for each of the occurrence, intensity and temperature models. For building a bivariate model for rainfall and temperature, a normal heteroscedastic model is used for daily temperatures, while a logistic regression model is used to model the rainfall occurrence and gamma model for intensity of rainfall. The observed data set is fitted to this definition file with an objective of maximizing the log likelihood. The diagnostic plots and the summary statistics are analyzed at each stage whether the additional covariates are significant or not. The models were calibrated for a period from 1995 to 2013 and validated for a period of 1981 to 1995 with 10 simulations done for each station. The developed models were used in the simulation mode for producing the simulated rainfall series for the period from 1981 to 1995 for all the six stations. The only input in the simulation procedure is the monthly values of the atmospheric predictors. The models were validated with the observed data at Idukky station. The future rainfall series from 2013 to 2025 is simulated with the validated model with monthly atmospheric predictors derived from CGCM3 SRES A2 scenario output.

4. Results and Discussions

The occurrence, intensity and temperature models were developed by progressive addition of covariates in the definition files. At each stage different types of diagnostic checks are carried out to assess the performance of the model. The p- value and Z statistics at each stage with each additional component helps to check whether the additional component is redundant or not. The occurrence frequency of the rainy days against the forecasted frequency of the final occurrence model when the probabilities are grouped in to probability deciles is shown in Table 1. The table shows that the occurrence and forecasted frequencies agree over a wide range of probability deciles for the calibrated model.

Summary statistics is provided which gives the residual means and residual standard deviations by month and year for the developed models. The model is considered to be correct if the mean Pearson residuals belong to a distribution with mean zero and standard deviation nearly one. The mean and standard deviations of the residuals by year and by month for the final Occurrence model and Intensity model are shown in Fig. 1(a) and (b). The dotted lines in the plots show the limiting range within which the values shall lie. These plots are analyzed to identify any temporal effects that have not been captured by the models. Most of the values for the mean residuals lie within the limit except for the month of May for the Occurrence Model. For the Intensity model, the annual residual means deviate only in the years 2003, 2005 and 2007. The deviations in the years 2005 and 2007 in the intensity model are

Table 1. Observed and Expected Frequencies of the Occurrence Model.

Probability	Observed proportion	Expected proportion
0.0-0.1	0.043	0.047
0.1-0.2	0.164	0.153
0.2-0.3	0.254	0.248
0.3-0.4	0.363	0.345
0.4-0.5	0.441	0.451
0.5-0.6	0.552	0.551
0.6-0.7	0.630	0.652
0.7-0.8	0.737	0.753
0.8-0.9	0.869	0.861
0.9-1.0	0.925	0.925

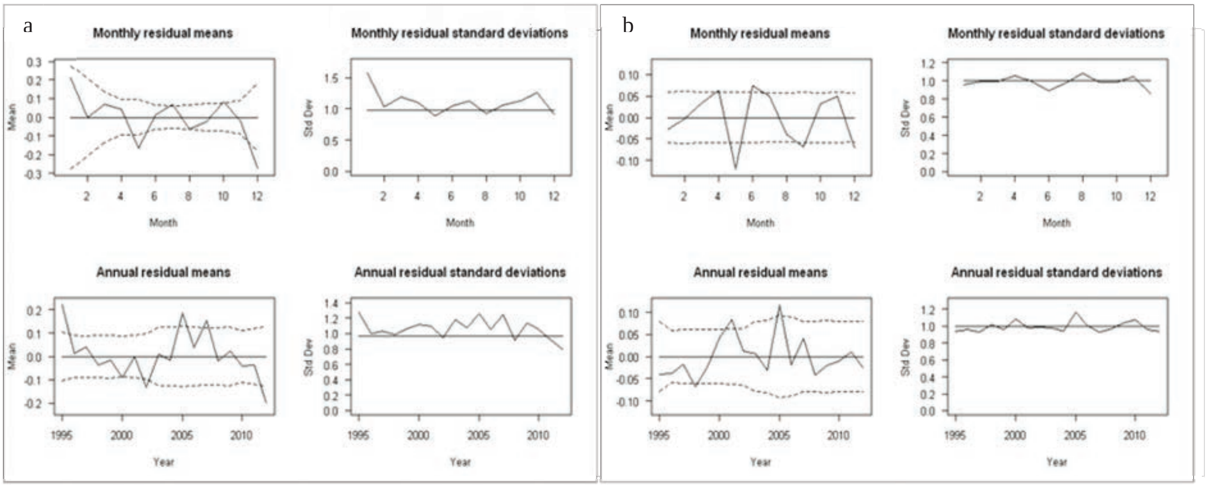
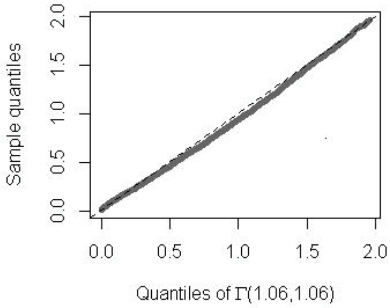


Fig. 1. Monthly and annual residual means and residual standard deviations of (a) Occurrence Model; (b) Intensity Model.

due to the heavy rains occurred during that time.

The monthly residual means of the Intensity model crosses the limits only in April and November. The deviations from the limits may be because of the inconsistency in the data at different stations with average annual values ranging from 3500mm to 1800mm. The seasonal structure can be observed in the monthly means in both the models which are shown by positive values. The model performance can be improved by adding some other components of the covariates which has significant influence on rainfall. The results of the temperature model are not discussed in this paper as daily temperature simulation is beyond the scope of this paper.

The Q-Q plot of the standardized residuals from the Intensity model is shown in Fig. 2. Monthly mean of the simulations generated using the developed model is plotted with observed monthly means as shown in Fig. 3. The Q-Q plot is for checking the gamma distribution assumption for every case in the data set. The standardised residual for an observation for the gamma distribution is the ratio of the observed value to the modelled mean. Q-Q plot is analyzed for comparing distributions of the modelled results with that of the observed data by plotting their quantile values against each other. If the two distributions are similar, their Q-Q plot will lie on the same line. The fitted model seems to be correct as the standardized residuals of the observations are from the same gamma distribution with equal shape and scale parameter which is obtained as 1.06 for the Intensity model. The data values shown as grey points are in agreement with the theoretical relationship which is shown in black line.



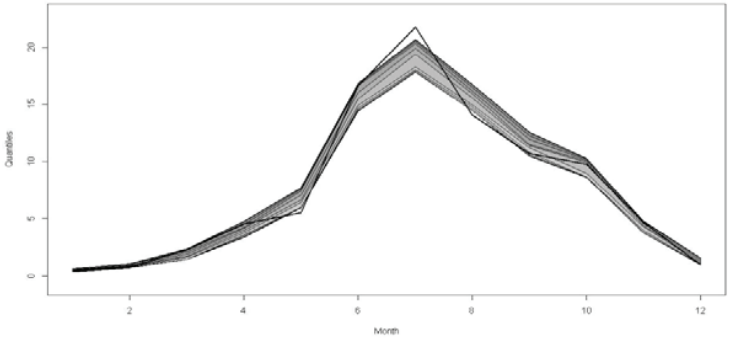


Fig. 2. Q-Q plot of standardized residuals for Intensity model. Fig. 3. Simulated and observed monthly means 1981-1995.

For Model validations, 10 simulations were generated for a period of 1981 to 1995. For this period, the daily data from 5 stations were unavailable. But for one station Idukky, the daily data was available. The simulated daily values of Idukky station was extracted from the output files. The simulation performance is assessed by calculating summary statistics for each simulation, and comparing the distribution of these summary statistics with the corresponding statistics of the observed value. The quantiles of the mean of the observed rainfall was plotted with the quantiles of the mean of the simulations for each month of the year. The figure shows that the monthly mean of the observed lies within the simulated range of the corresponding means for most of the months. The observed rainfall at Idukky station plotted with the simulated rainfall for a period from 1981 to 1995 is shown in Fig. 4. It can be seen from the plot that the model captures the seasonality perfectly. But 20% of the peak values of the historical rainfall are not fully captured by the model. This may be due to the inconsistency in the data set. The model can further be modified by including more stations around the catchment. Stations with high residual error may be omitted and the model can be refitted.

The validated models are used to produce the daily rainfall series for a future period. The simulations of daily rainfall produced for a period 2013 to 2025 using the atmospheric predictors derived from CGCM 3 SRES A2 scenario is shown in Fig. 5. These simulations are from the same gamma distribution with equal shape and scale parameters and it captures the seasonality perfectly. The GCMs provide predictions of atmospheric variables based on different scenarios specified by IPCC according to the emission of green house gases. Different simulations can be generated using the model with predictions from GCMs for different scenarios and these simulations can be reliably used as inputs to other hydrological models for impact assessment.

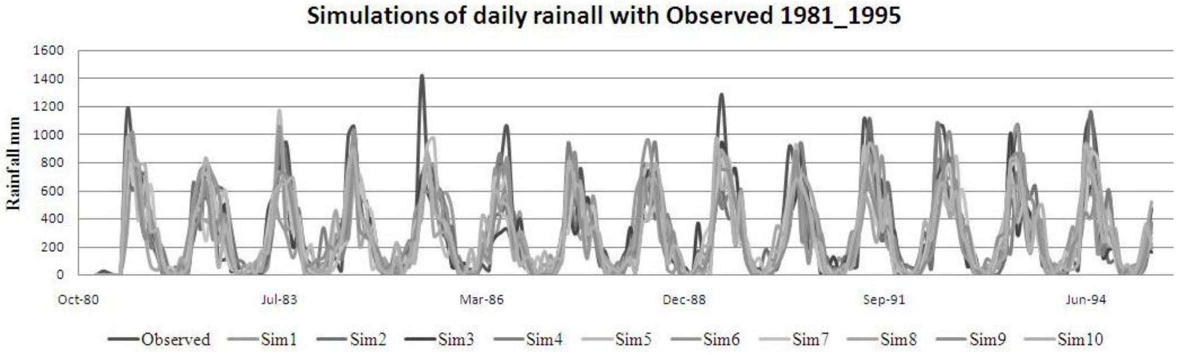


Fig. 4. Simulations of daily rainfall with observed from 1981_1995 at Idukky station.

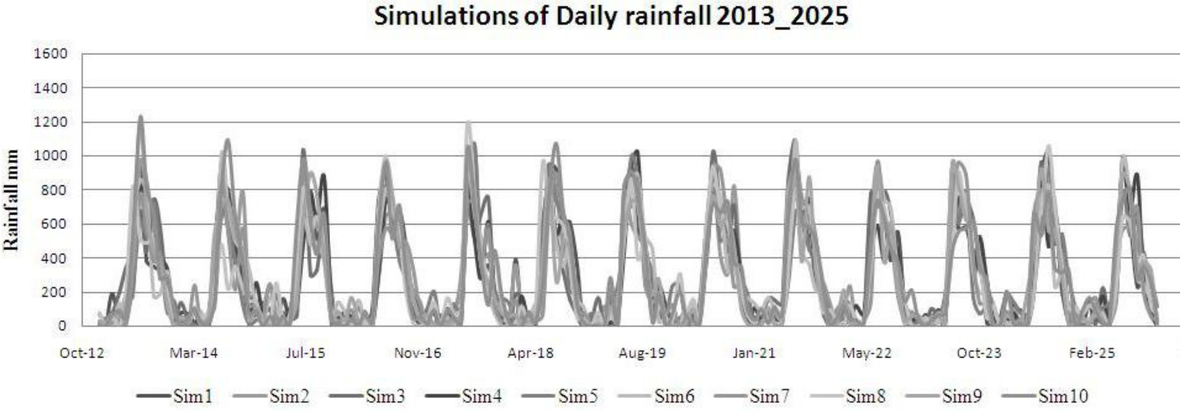


Fig. 5. Simulations of future rainfall 2013_2025 at Idukky station.

5. Summary

Generalized linear models for occurrence and intensity of daily rainfall and temperature were developed with historical data from 1995 to 2013 and atmospheric predictors derived from the NCEP/NCAR Re analysis data. Daily rainfall and temperature data from six rainfall stations in the catchment of Idukky reservoir were used for model development. Atmospheric variables namely mean sea level pressure, u-wind and v-wind components, wind speed, maximum air temperature and dew point temperature are used as external covariates for deriving the model. The model is validated for a period from 1981 to 1995. Rainfall series are generated for multi sites simultaneously. The performance of the model is assessed as acceptable by computing the statistics of the simulated series and comparing the distributions of these statistics with the statistics of the historical series. The developed model was successfully used for simulating the daily rainfall series for a future period from 2013 to 2025 in the Idukky catchment using the predictions of atmospheric circulation patterns derived from the global climatic model CGCM3 for IPCC SRES A2 scenario. These simulations can be used as inputs in impact assessment models or other hydrological models.

6. Conclusions

- The simulated daily rainfall series for the Idukky catchment follows the same distribution as that of the observed series as indicated by the performance evaluation based on statistical analysis.
- The Generalized Linear Bivariate model developed can be used as a weather generator to simulate the future rainfall series in a catchment using atmospheric predictors derived from the GCMs.
- The daily series of rainfall can be generated with several variables having different distributions at different locations simultaneously and even for ungauged stations or stations with substantial amount of missing data.
- The methodology adopted is simple and easy to implement.

References

[1] Bardossy A. Downscaling from GCM to local climate through stochastic linkages. J Environ Manage 1997;49:7-17.

[2] Benestad ER, Deliang Chen, Inger Hanssen-Bauer. Empirical statistical downscaling. London: World Scientific; 2008.

[3] Wilby RL, Charles SP, Zorita E, Timbal B, Whetton P, Mearns LO. The guidelines for use of climate scenarios developed from statistical downscaling methods. (http://ipccddc.cru.uea.ac.uk/guidelines/StatDown_Guide.pdf) 2004.

[4] Panagoulia D, Bardossy A, Lourmas G. Multivariate stochastic downscaling models for generating precipitation and temperature scenarios of climate change based on atmospheric circulation. J Global NEST 2008;10(2):263-272.

[5] Bardossy A, Plate EJ. Space Time Model for daily rainfall using atmospheric circulation patterns. Water Resour Res 1992;28(5):1247-59.

[6] Stehlik J, Bardossy A. Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation. J Hydrol 2002;256:120-141.

[7] Barkotula MAB. Stochastic generation of occurrence and amount of daily rainfall. Pak J Stat Oper Res 2010;6 (1):61-73.

[8] Stochastic weather generators. IPCC data, www.ipcc-data.org/guidelines/pages/weather_generators.html.

[9] Richardson CW. Stochastic simulation of daily precipitation, temperature and solar radiation. Water Resour Res 1981;17:182–190.

[10] Katz R, Parlange M. Over dispersion phenomenon of stochastic modelling of precipitation. J Climate 1998;11:591–601.

[11] Katz R, Parlange M, Naveau P. Statistics of extremes in hydrology. Adv Water Resour 2002;25:1287–1304.

[12] Chandler RE. Stochastic weather generators. Third VALUE training workshop, Trieste, Italy; 2014.

[13] Fahrmeir L, Tutz G. Multivariate statistical modelling based on generalized linear models. New York: Springer; 1994.

[14] Chandler RE. On the use of generalized linear models for interpreting climate variability. Environmetrics 2005; 16(7): 699-715.

[15] Chandler RE, Wheater HS. Analysis of rainfall variability using generalized linear models - a case study from the West of Ireland. Water Resour Res 2002;38(10).

[16] Wilks DS. Adapting stochastic weather generation algorithms for climate change studies. Clim Chang 1992;22:67-84.

[17] Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. The NCEP/NCAR reanalysis 40-year project. Bulletin of the American Meteorology Society 1996;77:437-471.

[18] Pradeepkumar PK. Physiographic features and changes in rain fall pattern of Kerala. Res Report, Physical Oceanography and Meteorology division, CUSAT; 1994.

[19] Simon A, Mohankumar K. Spatial variability and rainfall characteristics of Kerala. J Earth Sys Sci 2004;113(2):211-221.

[20] Ghosh S, Mujumdar PP. Future rainfall scenario over Orissa with GCM projections by statistical downscaling, Curr Sci 2006;90(3):396–404.