KKU Engineering Journal

http://www.en.kku.ac.th/enjournal/th/

Rainfall forecast in northeast of thailand using modified k-nearest neighbor

Uruya Weesakul,*[1] Nkrintra Singhratta[2] and Narongrit Luangdilok[1]

[1]Department of Civil Engineering,Faculty of Engineering, Thammasat University, Pathumthani, Thailand.

[2] Civil engineer, Department of Public Works and Town & Country Planning, Bangkok, Thailand.

Received January 2014

Accepted May 2014

Abstract

Since damage from natural disasters have increased due to anomalous global climate, scientists and engineers are interested in studying incorporation of the occurence of natural disasters. Thailand faces with flood in the wet season and drought in the dry season every year. The Northeast of Thailand is a region where found damages from disasters especially. This study developed a statistical model for forecasting rainfall in the Chi River Basin using large-scale atmospheric variables (LAV) as the independent variables to the modified k-nearest neighbor model. The significant LAV were identified over both Indian and Pacific Oceans. The model performance was evaluated using box plot of 3-month rainfall to present how well the model can capture the historical data and likelihood skill score (LLH). From both model evaluation, approximately 62% of historical rainfall data was captured forecasting model. LLH of rainfall ensembles in the Chi River Basin are quite good and better LLH can be found post 2000, especially June-August and July-September rainfall.

Keywords : Large-scale atmospheric variables, Modified k-nearest neighbor

*Corresponding author. Tel.: +6-626-133-1202; fax: +6-622-241-376

Email address: wuruya@engr.tu.ac.th.

1. Introduction

According to the occurrence of worst flood in Thailand in 2011, it caused the damage in agriculture and rainfall related activities, which subsequently influenced country economy and development. Moreover, Thai governments are growing concerned to cope with flood in wet season and drought in dry season. To prevent those natural disaster, rainfall forecasting is vital to manage water resources and crop planning. In the Northeast of Thailand, there is not always enough supply water for agriculture in normal year, yet there is sometime overabundance of water. According to studying drought category using log-linear models based on SPI (Standard Precipitation Index) in the Northeast of Thailand [1], there are still many high drought area at 99% confidence level. In addition, Nagon Wattanakij used vegetation indices of multi-temporal satellite data to detect drought in the Northeast of Thailand, drought still is found although it is high amount of rainfall in those years [2]. As long as Thai departments involving water resource management could predict accurately amount of rainfall at lead time 3-6 months, they managed to plan when to reserve or release water from their storage. However, it is not easy to develop rainfall forecasting model with high performance because climate change cause abnormal rainfall pattern. To develop the performance of forecasting model, the understanding of rainfall variability caused by large-scale atmospheric variables (LAV) is essential. From previous study, LAV (i.e. surface air temperature, sea level pressure, surface zonal and meridian winds) were identified as independent predictors for rainfall forecasting models in Chi River Basin [3]. Therefore, this study aimed to use identified LAV [3] to develop rainfall forecasting models in the Northeast of Thailand (Chi River Basin).

1.1 Study basin

The Chi River Basin (Fig. 1) is located in the northeast of Thailand between 15°30' -17°30'N latitude and 101°30'-104°30' longitude. It covers an area of 49,476 km^2. The Chi River originates in Chaiyaphoom Province and flows through 14 provinces to merge with the Mun River in Ubolrachathani Province. The total length of the Chi River is 830 m.

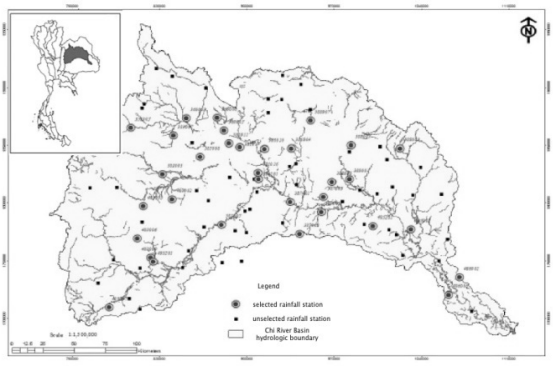

Figure 1.Chi River Basin

1.2 Climatology in the study basins

From Figure 2 the averaged monthly rainfall from 1975 - 2011 in the Chi River Basin ranged from 3.9 - 251.1 mm. The annual rainfall was estimated of 1,231 mm. The average temperature is highest in May and lowest in December. The wet season in the basin covers from May-October with two peaks in May and August or September. The wet season can be divided into two periods: the pre-monsoon season that lasts from May-July and the monsoon season that lasts from August- October. The pre-monsoon season is influenced by the southwest monsoon originating from the Indian Ocean in May. Subsequently, the ITZC causes the wind moving the north and covering Thailand and central China in mid-June or early-July. The ITZC moves back to Thailand in late- July, which causes a peak in August or September.

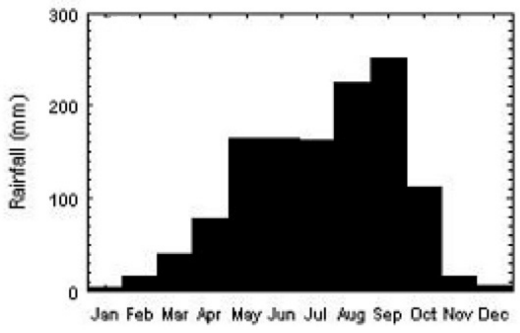

Figure.2 Monthly rainfall variation in Chi River Basin

2. Research methodology

2.1 Data collection

Monthly rainfall data in the Chi River Basin were obtained from the Thailand Meteorological Department (TMD). The rainfall stations are located in the northeast of Thailand (Fig 1), which are in and around the study basins. There are 34 selected stations in the Chi River Basin. The station selection was done based on the incomplete data less than 10% of rainfall data and the consistency of duration in all selected stations. The monthly rainfall from 1975-2011 were used in this study.

The monthly data of large-scale atmospheric variables (LAV) from 1948 to the present at the surface level and higher levels are provided by the National Centers for Environmental Prediction (NCEP). The observed data of LAV are presented in grid cells with a size of 2.5° latitude x 2.5° longitude [4]. The LAV that were used in this study included the surface air temperature (SAT), sea level pressure (SLP), zonal (u) and meridional (v) winds at the surface level. The monthly data from 1975 - 2011 of four variables were applied in the statistical analysis.

2.2 Identification predictor for rainfall forecasting model

To analyze the statistical relationships between rainfall in the study basins and the large-scale atmospheric variables (LAV) such as surface air temperature (SAT), sea level pressure (SLP), zonal wind speed (U) and meridional wind speed (v), the correlation maps were applied. The correlation map based on Pearson's r [5] is an interactive plot and analysis (example showing in Figure 3. It is provided by the Earth System Research Laboratory (ESLR) of the National Oceanic and Atmospheric Administration (NOAA) [6]. LAV were selected by criteria based on the significant relationships at 95% confidence level, the significance of a correlation coefficient is tested using Fisher's Transformation [5]. Location and group of identified predictors for 3-month rainfall forecasting in Chi River Basin were shown in Table 1 and Table 2 respectively. For Dec-Feb rainfall forecasting model, relationship between rainfall and LAV was not found but historical rainfall data was used as predictor for model

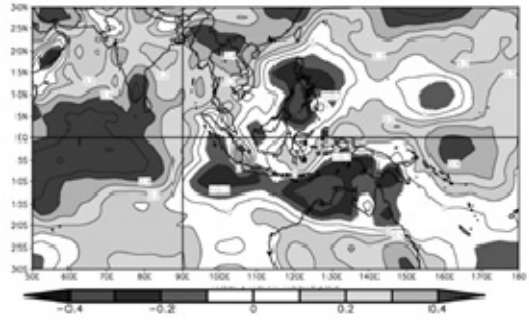

Figure 3. Correlation maps between June-August rainfall and August-October SAT at 10 months lag time

Table 1 Location of identified predictors

Variable	Location	
	Latitude	Longtitude
SAT1	12.5-20°S	60-70°E
SAT2	0-10°S	70-80°E
SAT3	0-10°S	160-170°E
SAT4	0-10°S	80-90°E
SAT5	17.5-22.5°N	170-180°E
SLP1	0-10°N	90-100°E
SLP2	10-20°N	140-150°E
SLP3	5-15°N	110-120°E
SLP4	0-10°N	55-65°E
SLP5	0-10°S	160-175°E
SLP6	0-10°N	90-100°E
SLP7	0-10°N	55-65°E
U1	25-35°N	140-150°E
U2	0-10°N	105-115°E
U3	2.5°S-7.5°N	160-170°E
U4	12.5-20°S	90-100°E
U5	12.5-20°S	95-105°E
U6	0-10°S	170-180°E
U7	12.5-20°S	70-80°E
U8	15-25°S	95-105°E
V1	0-7.5°N	120-130°E
V2	5-10°S	60-70°E
V3	20-25°N	125-135°E
V4	20-25°N	130-140°E
V5	0-5°S	60-70°E
V6	12.5-20°N	65-72.5°E

Table 2 The Identified predictors for 3-month rainfall for

Periods	Potential Predictors
Jan-Mar	U1
Feb-Apr	SAT1, SLP1, SLP2, U2, U3, V1, V2
Mar-May	SAT1, SLP2, SLP3, U2, U3, V2, V3
Apr-Jun	V4
May-Jul	U4
Jun-Aug	SAT2, SLP4, SLP5, U5, V5
Jul-Sep	SAT2, SAT3, SLP5, SLP6, U5, U6, V6
Aug-Oct	SAT3, SAT4, SLP4, SLP5, U6, U7, V6
Sep-Nov	SAT5, SLP6, U6, V6
Oct-Dec	SAT5, SLP7
Nov-Jan	U8
Dec-Feb	-

2.3 Combination case of predictors

For k multiple independent variables, there are 2k – 1 possible combination cases. To ensure that there is no redundancy, an optimal subset of variables is selected by a selection method. The Generalized Cross Validation (GCV) with the leave-one-out technique was applied in this study to select an optimal subset of predictors. The GCV estimates the error from a developed regression following the Eq. (1). A combination case of predictors is selected among all possible cases based on the minimum GCV.

$$GCV = \frac{\sum_{i=1}^{n} (y_i - y'_i)^2}{(1 - m/n)^2} \quad (1)$$

where y_i-y'_i is the error from the developed regression using each combination case, n is the number of data points, and m is the number of parameters.

2.4 The modified k-nearest neighbor (k-nn) model

The k-nn model is the non-parametric approach. it has been developed to improve the performance of the parametric regression, which is a function to fit the relationship between dependent (y) and independent (x) variables. Nonparametric regression does not require a prior assumption of relationship between two data sets. The fitting function (f) can locally capture the relationship using a small set of neighbours (k) at a given point (xi). So, the function is flexible and able to describe the relationship better than parametric regression. The k-nn model method can be divided into two process: fitting regression and simulation.

1) The steps of fitting regression of the modified k-nn model are described as follows: fitting process, the size of the neighbors (k) and the order of the polynomial (p), which is normally 1 or 2, are

selected and associated with the combination of k and p so as to obtain minimum GCV. The GCV is estimated by Eq. (1) (where y_i-y'_i is the error from the developed regression using k and p). After the fitting process, the dependent variables (y) according to the developed fitting regression are then estimated and called mean estimations ($\bar{y}_1$, $\bar{y}_2$, $\bar{y}_3$,..., $\bar{y}_n$). Then, the residuals (e_1,e_2,e_3,...,e_n) are computed.

2) The step of simulation of the modified k-nn model: The forecast of a dependent variable is required the new independent variable (x_{new}). The mean estimation ($\bar{y}_{new}$) is calculated from the developed regression. Then the residual (ei) was randomly selected the by using a weight function, presented in Eq. (2) and the a residual was added to $\bar{y}_{new}$. The distance between x_{new} and all the x_i can be calculated by Eq. (3). Finally, the residual process has been repeated as many times as required to achieve a number of simulation (N (N=300 is used in this study))

$$W(j)=\frac{1/j}{\sum_{i=1}^{k}(1/i)} \quad (2)$$

where W(*j*) is a weight of a neighbour of x_{new} and its distance from x_{new} falls in the j^{th}-rank, and k is the size of neighbour which can be different from k in the fitting process.

$$d_i=\sqrt{\sum_{j=1}^{m}(x_{new,j}-x_{i,j})^2} \quad (3)$$

where *i*=1, 2, 3, ..., *n*, and *m* is the number of independent variables.

2.5 Evaluation of model performance

The modified k-nn model was developed and evaluated from 1977 – 2011 (35 years) using identified predictors as independent variables. The leave-one-out cross validation was applied at all points of observation observed data. Note that there were 300 simulations of each year obtained from the modified k-nn model. The criteria used to evaluate the performance of the modified k-nn model are (a) the annual variability of 3-month rainfall; and (b) the likelihood skill score (LLH).

a) The annual variability of 3-month rainfall were presented in a box plot of 300 simulations of each year. The solid line presented the observed rainfall data.

b) The likelihood skill score (LLH) was used to evaluate a statistical stochastic model in terms of capturing the PDF of climatology. LLH varies from 0.0 to +3.0 in this study. The score of +1.0 indicates similarity between the model performance and reference simulated climatology. A score of less than +1.0 indicates the weaker performance of the model compared to climatology. Otherwise, the score of higher than +1.0 indicates a better performance of model. Steps of calculation are as follows:

(1) Divide the observed rainfall into 3 equal sets based on the specific thresholds, which in this case are the 33rd and 67th percentiles. Rainfall below 33rd percentile is defined as below-normal rainfall while rainfall above 67th percentile is defined as above-normal rainfall. Otherwise, it is defined as normal rainfall.

(2) Calculate the categorical probabilities of climatology, which are the proportion of rainfall in each category. Since historical data is divided into 3 equal sets, the categorical probability of all three categories is 1/3.

(3)Then, in a given year, the N simulated ensembles are also divided into three categories using the same thresholds i.e. the 33^{rd} and 67^{th} percentile. The categorical probabilities of ensembles in a given year, which are the proportion of rainfall ensembles in each category, are computed. Subsequently, LLH is estimated using Eq. (4)

$$\mathrm{LLH} = \frac{\prod_{t=1}^{n} \hat{P}_{j,t}}{\prod_{t=1}^{n} P_{cj,t}} \quad (4)$$

where n is the number of years, j is the category of the observed value in the year t, $\hat{P}_{j,t}$ with $\hat{P}_{j,t} = (\hat{P}_{1,t}, \hat{P}_{2,t}, \hat{P}_{3,t}, ..., \hat{P}_{k,t})$ is the probability of rainfall ensembles for category j in the year t, where k is the number of categories, and $P_{cj,t}$ is the categorical probability of climatology for category j in the year t, which in this case is the same value for all three categories.

3. Research results and discussion

The modified k-nn model was developed and evaluated from 1977 - 2011 using identified predictors as independent variables. By the leave-one-out cross validation, the modified k-nn model was evaluated separately for simulations of 3-month rainfall (i.e. Jan-Mar, Feb-Apr, Mar-May,..., Dec-Feb). The variability of 3-month rainfall and the anomalous events in the twelve periods can be found from the historical observation. The rainfall ensemble from 1977 to 2011 quite well capture the annual variability of observation Figure 4. Out of the 35 validating year, the modified k-nn model can capture historical observation in Chi River Basins by 19-25 years (around 62%). Although amount of rainfall observation are high variation, the modified k-nn model correctly shown the trend of rainfall.

Figure 5. shows the likelihood skill score of rainfall ensembles during all periods of year from 1977-2011 in the Chi River Basin. The darker shading represents a better performance of the modified k-nn model. Modified k-nn model has better performance for Rainy season which are on Jun-Aug and Jul-Sep rainfall in the Chi River Basin. Moreover, LLH of rainfall ensembles are quite high (dark shading) post 2000. It is important to make a note that variation of large-scale atmospheric significantly impact to rainfall in the Chi river basin.

4. Conclusion

This study developed the statistical models for forecast rainfall in Chi River Basin by apply large-scale atmospheric variables (LAV) as independent variable to Modified K-Nearest Neighbor Model. In addition, those LAV are in both Indian and Pacific Oceans. The model performance are evaluated by using box plot of annual variability of 3-month rainfall and likelihood skill score (LLH). From model evaluation, approximately 62% of historical rainfall data was captured forecasting model. LLH of rainfall ensembles in the Chi River Basin are quite well and better LLH can be found post 2000, especially in Rainy season which are June-August and July-September rainfall. For further study, rainfall isohyets should be applied to separate study area because almost of all previous researchs selected scope of study area as watershed area.

5. Acknowledgements

The authors are thankful to the Hydro and Agro Informatics Institute (HAII), Thailand for the financial support of the research.

6. References

[1] Salee W. Prediction of drought category using loglinear models based on SPI in the Northeast of Thailand [MSc thesis]. Nakorn Prathom: Silpakorn University; 2009. (In Thai).

[2] Weesakul U, Singhratta N, Luangdilok N. Identification of large-scale atmospheric predictors for rainfall forecasting of Chi River Basin (Thailand). In: 5th National Convention on Water Resources Engineering, 5 - 6 September 2013, Le Meridien Chiang Rai Resort, Chiang Rai, Thailand; 2013. (In Thai).

[3] Weesakul U, Singhratta N, Luangdilok N. Identification of large-scale atmospheric predictors for rainfall forecasting of Chi River Basin (Thailand). In: 5th National Convention on Water Resources Engineering, 5 - 6 September 2013, Le Meridien Chiang Rai Resort, Chiang Rai, Thailand; 2013. (In Thai).

[4] Kalnay E, Kanamitsu M.et al. The NCEP/NCAR reanalysis 40-year project. Bull. Am. Meteorol. Soc. 77 (1996) 437–471.

[5] Haan C T. Statistical Method in Hydrology (2nd ed.), Iowa State Press, U.S.A., 2002.

[6] (Earth System Research Laboratory ESRL). May 2011 at http://www.esrl.noaa.gov/psd/data/correlation.

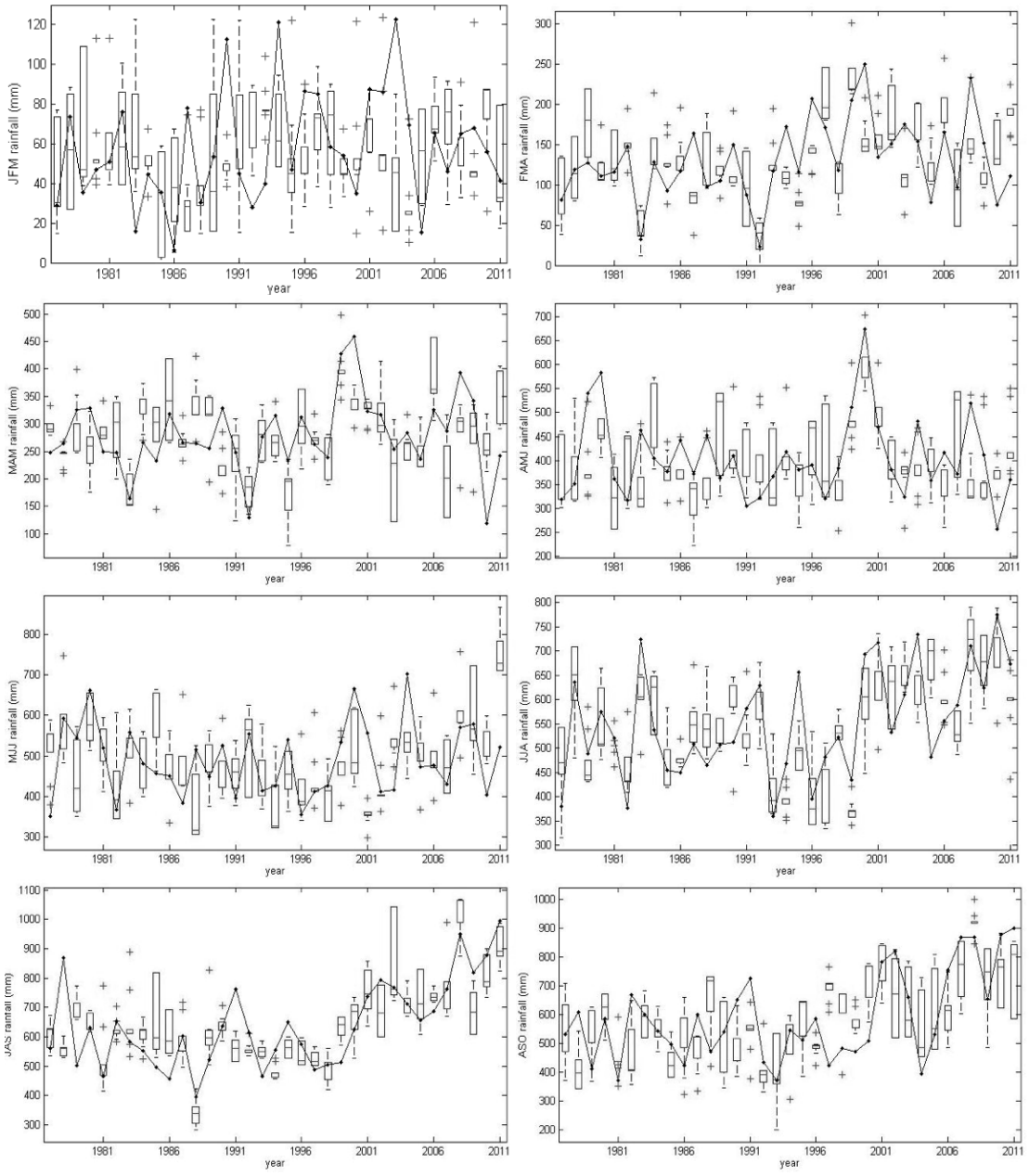

Figure 4 Box plot of 3-month rainfall from 1977 - 2011, estimated from 300 simulations of the modified k-nn model. The solid lines with marks represent the annual observed rainfall

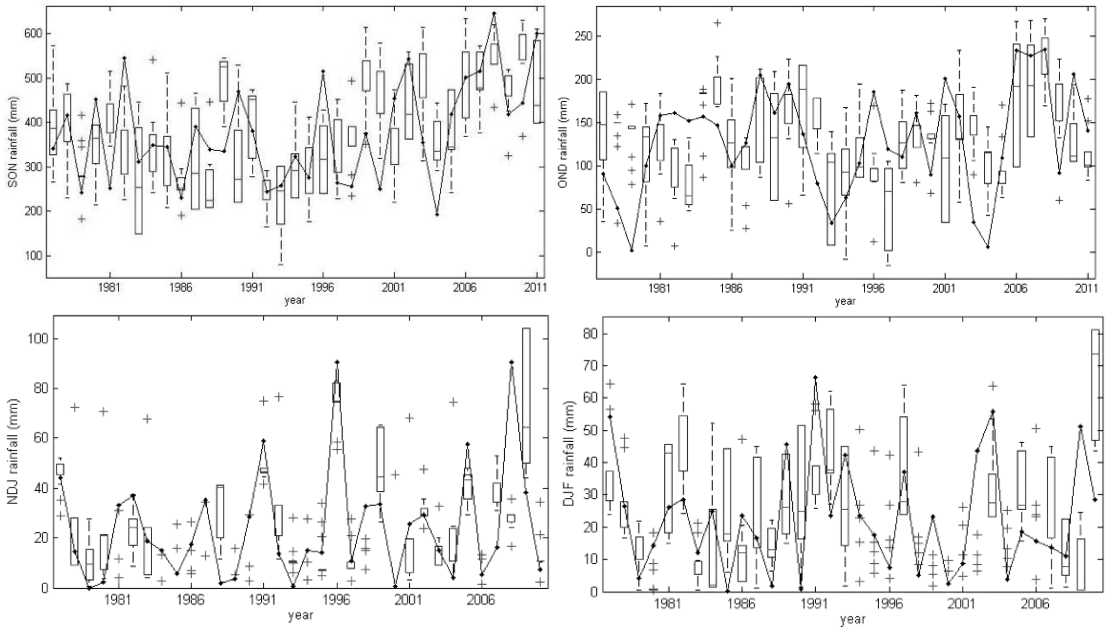

Figure 4(con.): Box plot of 3-month rainfall from 1977 - 2011, estimated from 300 simulations of the modified k-nn model. The solid lines with marks represent the annual observed rainfall

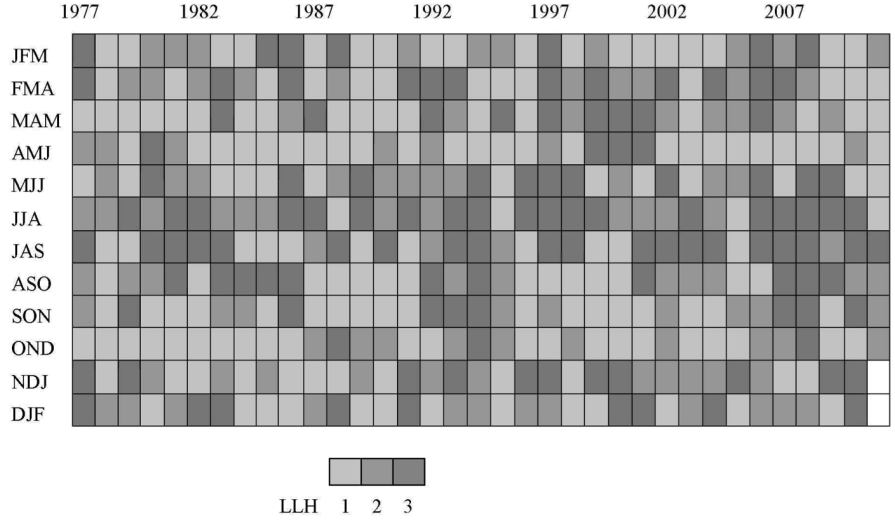

Figure 5 Likelihood skill score of rainfall ensembles in the Chi River Basin from 1977-2011