KKU Engineering Journal

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Seasonal rainfall forecast for cropping pattern planning using a modified k-nearest neighbor model

Uruya Weesakul*, Nkrintra Singhratta and Phailin Yodpongpiput

Faculty of Engineering, Thammasat University, Pathum Thani 12120, Thailand

Received June 2016
Accepted July 2016

Abstract

Drought phenomena have recently occurred in Thailand causing severe damage to the agricultural sector, especially in the Central, Northern and Northeastern regions. A reliable seasonal rainfall forecast model is needed to provide useful information for effective crop planning and water resource management. The study aims to develop a seasonal rainfall forecast model using a stochastic model k-nearest neighbor technique for an upcoming year. The Mun River basin, located in the Northeastern region, was selected as a case study. Monthly rainfall data from 152 stations, distributed throughout the river basin, were collected over a period of 37 years from 1975 to 2011. Analysis of correlation between large scale atmosphere variables (LAV) around the study basin and seasonal rainfall over the river basin revealed that the surface air temperature (SAT), sea level pressure (SLP), surface zonal wind (U) and surface meridional wind (V) over the China Sea and Pacific Ocean influence seasonal rainfall over the basin. These four LAV variables were used as predictors in a modified k-nearest neighbor model to forecast seasonal rainfall. The likelihood skill score (LLH) was adopted as a technique to evaluate model performance. A test of model performance, using seasonal rainfall for a period of 37 years (1975 to 2011), revealed that the model is able to predict seasonal rainfall with a reliability of around 60%, providing sufficient information for appropriate crop pattern planning in the area.

Keywords: K-nearest neighbor, Large-scale atmospheric variables

1. Introduction

Thailand experiences both flood and drought disasters more frequently during the past two decades. It is always claimed that climate change is main cause of rainfall variability in Thailand which consequently resulting in more and more severe flood and drought damages. Generally Thai economy depends on the spatial and temporal distribution of rainfall [1]. Around 50% of labor force is in agricultural sector. Agricultural productivity in the country relies directly on monsoon rainfall bringing water to support crop growth. Rain is the most important factor for rice production in this south-east Asian region. The position of Thailand, locating in tropical latitudes, subjects to high inter-annual and seasonal variability in rainfall [2]. Change in spatial and temporal distribution pattern of rainfall can cause moisture stress to variety of crops leading to disastrous effects on the people and economy.

There are a number of studies focusing on the temporal and spatial characteristics of rainfall in different parts of the world. There are also a variety of studies focusing on impact of climate change on rainfall variability, shift in rainfall pattern and trends of rainfall in different parts of the world.

These studies reveal that rainfall characteristics highly depend on the climate regions, seasons, selected rainfall parameters, and that degree of impact of climate change varies from region to region. For example, spatial and temporal analysis of annual rainfall variability in Turkey was studied by Turks [3]. Spatial analysis of precipitation trends in the region of Valencia (East Spain) was conducted by De Luis et al [4] spatial distribution of seasonal rainfall trends in a western Mediterranean area was also analyzed by Gonzalez-Hildago JC, et al in 2001 [5]. Spatial distribution of seasonal rainfall trends in Sicily (1921-2000), was studied by Cannarozzo M, et al in 2006 [6]. In 2008, Zhang Q, et al analyzed spatial and temporal variability of extreme precipitation during 1960-2005 in the Yangtze River basin and also evaluated possible association with large-scale circulation [7]. In 2009, Zhang Q, et al continued their study on observed changes of drought/wetness episodes in the Pearl River basin, China, using the standardized precipitation index and aridity index [8]. Spatial and temporal variability of precipitation during 1957-2005 over China was also studied by Zhang Q, et al in 2009 [9]. In the same year, Zhang Q et al, continued their study on trends and abrupt changes of precipitation maxima in the Pearl River basin, China [10]. Zhang Q et al recently analyzed precipitation extremes in a karst region: a case study in the Guizhou Province, southwest China [11], in 2010. In the Hawaiian Islands, changes in precipitation extremes wave studied by Chu P-S, et al in 2010 [12].

*Corresponding author. Tel.: +6626 133 1202
Email address: wuruya@engr.tu.ac.th

doi: 10.14456/kkuenj.2016.23

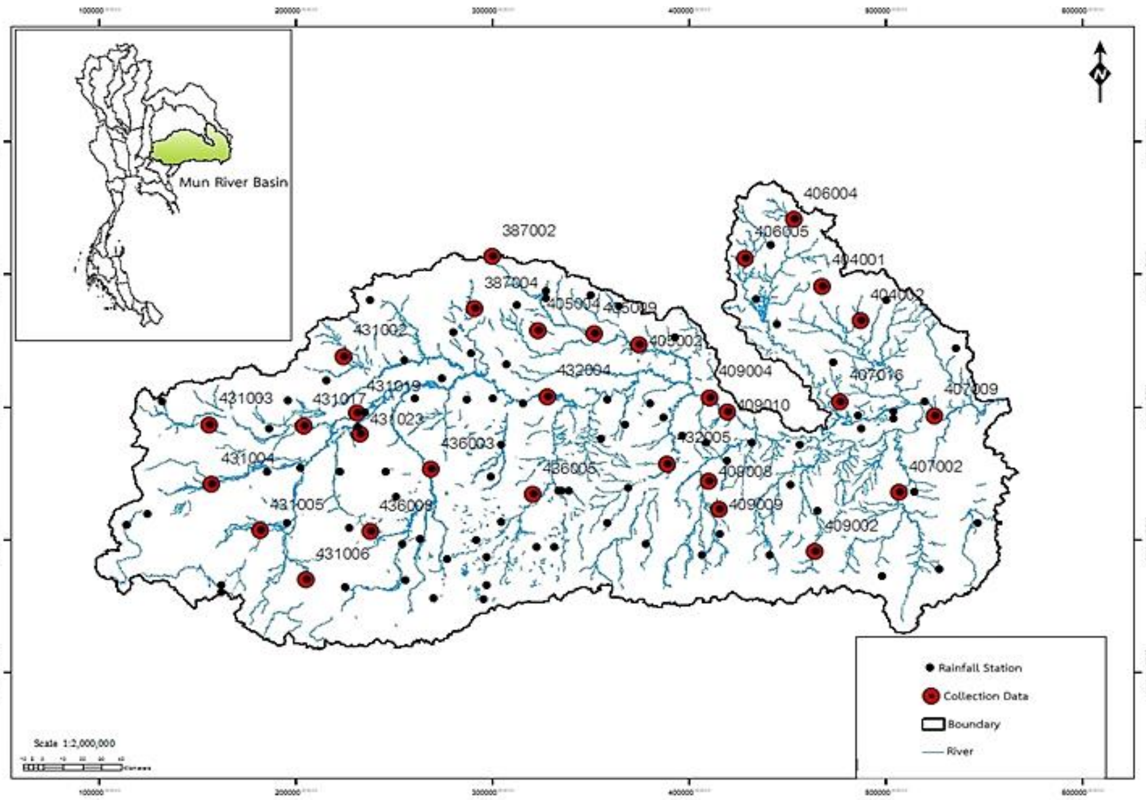

Figure 1 Distribution of Rainfall Station over Mun river basin

It is noted from the studies mentioned above that is not any individual method being capable to reveal different statistical properties of rainfall variability, each method has its own strength and weakness, and different methods should be conducted in analysis to component each other.

Besides, statistical approached methodology mentioned above, General Circulation Models (GCMs) have been generally used as a tool for evaluation of the effect of climate change on rainfall process at a local site. These GCMs have been recognized to be able to represent reasonably well the main features of the global distribution of basin climate parameters for regional scale. Linkage of the greenhouse effect and climate change was described by Mitchell JFB [13]. United Nations Regional Environment Programmer [UNEP], also explained how global warming affect the world [14]. Causes and effects of global warming were described and forecasted for the future by Maslin Mih 2007 [15]. National Intelligence Council [NIC] also predicted the impact of climate change to 2030 [16]. NASA also mentioned about global warming situation as “This ups and downs of global warming” [17]. Chen TC and Yoon JH (2000) investigated inter-annual variation of summer monsoon rainfall in Indochina and also explained its possible mechanism [18]. In 2005, Singhrattna N et al analyzed inter-annual and inter-decadal variability of summer monsoon season for rainfall in Thailand [19]. Later in 2011, Singharatta N. et al. determined the effects of climate change on summer monsoon season rainfall in the upper Chao Phraya River Basin in Thailand [20]. The study used stochastic statistical model to simulate rainfall for a long term planning in water resource and an adaptation strategy to deal with future climate change. It has been found from study that due to the atmospheric oceanic circulation, the pre-monsoon and monsoon season rainfall in the upper Chao Phraya River Basin was influenced by climate change not only in the variability of rainfall but also in the frequency of extreme event. If has been further illustrated in such study that under a condition of doubling CO2 concentration by 2100. The effect of climate change on rainfall variability result in decreasing trends of May-June-July rainfall and August-September-October rainfall. The decreasing trends of summer monsoon rainfall will certainly affect water resource planning and water-related activities in the study area. However due to the large spatial and temporal variation of monsoon rains, different trends of rainfall may occur in different parts of Thailand. Another study on “rainfall Forecast in Northeast of Thailand Using Modified K-Nearest Neighbor” was also conducted in 2014 [21], model evaluation revealed that it can be used to forecast 3 month rainfall over Chi river basin with approximately 62% accuracy.

Due to high spatial variation of rainfall in this region, it is therefore interesting to apply, this Modified K-Nearest Neighbor model to forecast seasonal rainfall in the Mun river basin, located in the lower part of northeast region of Thailand. This is to investigate and verified whether the model can be used over the entire region or not. This study aims to specifically develop rainfall forecast model using modified k-nn technique to predict seasonal rainfall over the Mun river basin.

2. Study area and data collection

In order to investigate the consistency of modified k-nn model in prediction of seasonal rainfall, the Mun river basin, located adjacent to the Chi river basin which was already tested [21], was selected as study area. The Mun river basin has a total drainage area of about 71,000 km3, with 152 rainfall stations distributed over the river basin, as show in Figure 1.

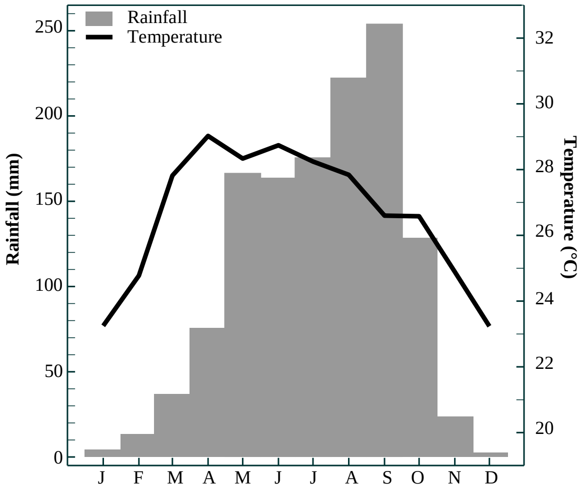

Figure 2 Average Monthly Rainfall and Monthly Temperature in Mun river basin during 1975-2011

Rainfall data quality and availability were checked before analysis. It has been found that only 30 stations, having continuous rainfall data with record length greater than 30 years whose missing data is less than 5 %. These 30 stations, as shown in Figure 1, are used as representatives rainfall stations over the river basin.

Preliminary analysis of time distribution of monthly rainfall over one year was conducted using these 30 rainfall stations for period of 37 years from 1975-2011, as shown in Figure 2. It is obviously presented in Figure 2 that there are two main seasons in this region, one is a wet season beginning from April to October, this wet season is influenced by wet South Asian monsoon, the other one is dry season beginning from November and ending in February. This dry season is influenced by dry air stream or dry northeast monsoon.

For large-Scale Atmospheric Variables (LAV) Data, observed monthly data is available at website of the National Centers for Environmental Prediction (NCEP). The observed data of LAV is presented in grid cells with a size of 2.5° latitude x 2.5° longitude or approximately 227 x 227 km^2 [22]. The monthly observed LAV data is available from 1948 to present, these data from 1975-2011 is analyzed to identified the most influence variables which have impact on the amount of seasonal rainfall over the Mun river basin [23]. It has been found that amongst variety of LAV, Surface Air Temperature (SAT), Sea Level Pressure (SLP), Zonal (U) and Meridional winds (V) at surface level have significant relationships with amount of 3-month rainfall over the Mun river basin [23]. In this study, such 4 LAV are used as predictors in our season rainfall forecast model.

3. Rainfall forecast model

A variety of Rainfall Forecast models are available for researchers, this study adopt a stochastic statistical model to be developed for seasonal rainfall forecast. The k-nn model, as a non-parametric approach model, is selected to as a tool due to its convenient for use under available of hydrologic and meteorological data concerned over the study area.

The k-nn model has been developed to improve the performance of the parametric regression, which is a function to fit the relationship between dependent (y; or seasonal rainfall in our case) and independent (x; or LAV in our case) variables. Non parametric 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 neighbors (k) at a given point (Xi).

Therefore, the function is able to describe the relationship better and more flexible than parametric regression. There are two main steps in the calculation procedure for k-nn model development, which are firstly fitting regression process and the simulation process.

For the fitting regression process, the size of the neighbors (k) and the order of the polynomial (p), which is generally 1 or 2, are selected and associated with the combination of "k" and "p" in order to obtain minimum Generalized Cross Validation (GCV) with the leave-one-out technique, which will allow to select an optimal subset of predictors of the model.

For k multiple independent variables, there are "2k-1" possible combination cases. In order to ensure that there is no redundancy, an optimal subset of variables is selected by GCV with leave-one-out technique. The GCV estimates the error from a developed regression, which can be expressed as:

$$GCV = \frac{\sum_{i=1}^{n} \frac{\left(y_i - y_i^{'}\right)^2}{n}}{\left[1 - m/n\right]^2} \tag{1}$$

Where $(y_i - y'_i)$ is the error from the developed regression using each combination case, n is the number of data, m is the number of parameters.

Following calculation of GCV in eq. (1), a combination case of predictors can be selected amongst all possible cases based on the minimum GCV.

After the GCV is estimated by eq. (1), (where $y_i - y'_i$ is the error from the developed regression using k and p), the dependent variables (y) according to the developed fitting regression are then estimated and called "Mean Estimations" $(\bar{y}_1, \bar{y}_2, \bar{y}_3, \dots, \bar{y}_n)$. Then The "Residuals" ($e_1$, e_2, e_3, …, e_n) are computed.

The second step of calculation procedure is "Simulation". For the simulation process of the modified k-nn model, the forecast of a dependent variable, ($\bar{y}_{new}$, in our case is seasonal rainfall over the Mun river basin), is required the new independent variable (X_{new}, in our case is SAT, SLP, U and V). The mean estimation ($\bar{y}_{new}$) is calculated from the developed regression.

Then randomly the "Residual" (e_i) is selected by using a weight function [W(j)], which can be presented as :

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

Where W(j) is a weight of a neighbor of X_{new} and its distance from X_{new} fall in the "jth" rank, and k is the size of neighbor which can be different from k in the fitting process.

The $\bar{y}_{new}$ is than added with the "RESIDUAL" (e_i). The distance between X_{new} and all the X_L can be adjusted by using the following equation:

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

Where i=1,2,3,…n and n is the number of independent variables.

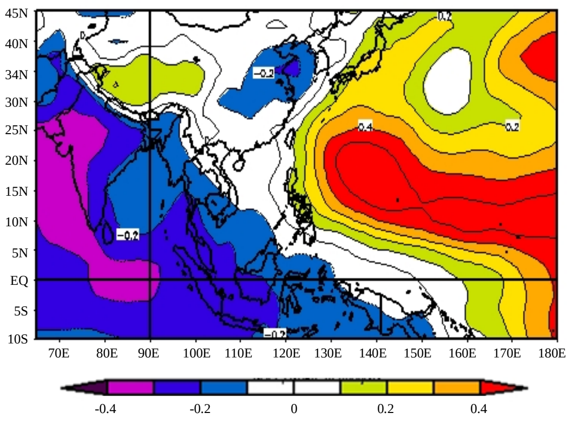

Figure 3 Correlation maps between March-May rainfall and August-October SLP at 7 months lag time

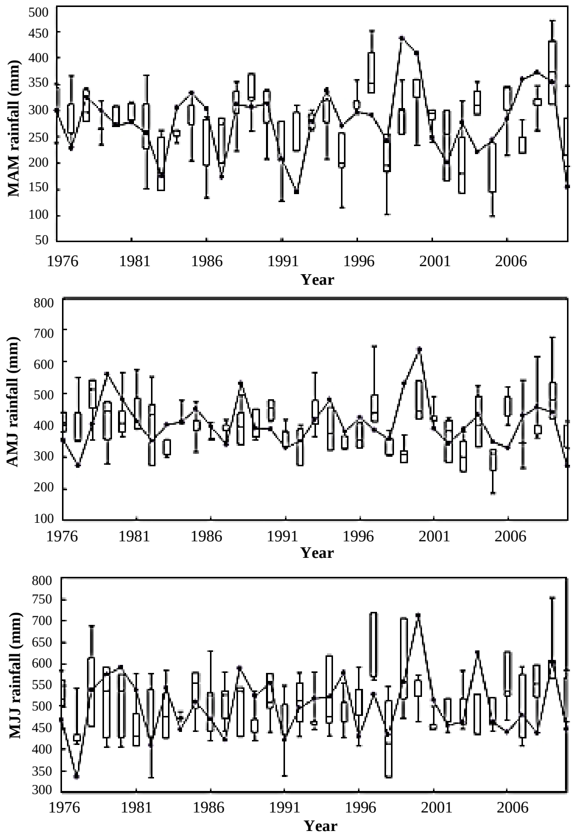

Figure 4 Box plot of forecasting of seasonal rainfall (MAM, AMJ, MJJ) over Mun river basin from 1975-2011

The procedure of calculation of RESIDUAL (e_i) , X_{new}, $\bar{y}_{new}$, $W_{(j)}$ and d_i can be repeated "N" time. In this study the number the number of simulation "N" is equal to 300. Finally the predicted seasonal rainfall is calculated "N" times, therefore the modified k-nn model is able to predict seasonal rainfall ahead, associated with probability of occurrence.

4. Forecasting of seasonal rainfall over the Mun river basin

Generally the preprocess before development of rainfall forecast model is to identity of most influence variables related to the variation of seasonal rainfall over the study river basin. Weesakul U, et al [23] investigate statistical relationships between seasonal rainfall over the Mun river basin and the Large-Scale Atmospheric variables (LAV), using correlation map base on Pearson's *r* [24]. An interactive plot and analysis of correlation between 3-month rainfall over Mun river basin and LAV around Thailand was conducted using monthly rainfall and LAV data from 1975-2011. This process of calculation was manipulated via website of Earth System Research Laboratory (ESRL) of the National Oceanic and Atmospheric Administration (NOAA) [25]. An example of results of analysis is shown in Figure 3, presenting correlation coefficient between 3 month rainfall (March-May) over Mun river basin and August-October SLP at 7 months lag time [23]. It has been found from such study [23] that, the most influence predictors in rainfall forecast model should be SAT, SLP, U and V as shown in Table 1 and Table 2. These predictors are used in rainfall forecast model, to forecast seasonal rainfall with the leading time from one month to one year.

An example of Seasonal Rainfall Forecast over the Mun river basin is shown in Figure 4 presenting trend of seasonal rainfall during 1975-2011. Box plot of forecast seasonal rainfall can be manipulated with the leading time from one month to one year. The results of this prediction show that the model is capable to provide a trend of next year rainfall whether it is a wet year, normal year or dry year. Such information will be useful guideline for appropriate cropping pattern planning leading to effect water resources management and reducing damage of crop.

Table 1 Location of identified predictors

Variable	Location	
	Latitude	Longtitude
SAT1	5°N-15°N	85°E-95°E
SAT2	15°N-25°N	120°E-130°E
SAT3	20°N-30°N	130°E-140°E
SAT4	5°S-0°E	115°E-125°E
SAT5	5°N-15°N	115°E-125°E
SAT6	25°N-30°N	100°E-110°E
SLP1	10°N-20°N	140°E-150°E
SLP2	40°N-45°N	110°E-120°E
SLP3	5°S-5°N	110°E-130°E
SLP4	35°N-40°N	70°E-80°E
u1	25°N-30°N	125°E-130°E
u2	5°N-15°N	130°E-140°E
u3	10°S-5°S	70°E-80°E
u4	30°N-40°N	100°E-120°E
u5	25°N-35°N	105°E-115°E
v1	15°N-25°N	135°E-145°E
v2	15°N-25°N	125°E-130°E
v3	5°S-5°N	70°E-80°E
v4	0°-5°N	120°E-130°E
v5	30°N-35°N	115°E-125°E

Table 2 The identified predictors for 3-month rainfall forecast

Periods	Potential Predictors
Jan-Mar	SAT1, U1, V1
Feb-Apr	SAT2, SLP1, U2, V2
Mar-May	SAT2, SLP1, U2, V2
Apr-Jun	SAT3, SLP1, U2, V2
May-Jul	SAT3, SLP1, V2
Jun-Aug	SAT4, SLP3, U3, V3
Jul-Sep	SAT5, SLP1, U4, V4
Aug-Oct	SAT6, SLP3, U5
Sep-Nov	SLP3
Oct-Dec	-
Nov-Jan	-
Dec-Feb	SLP4, U5, V5

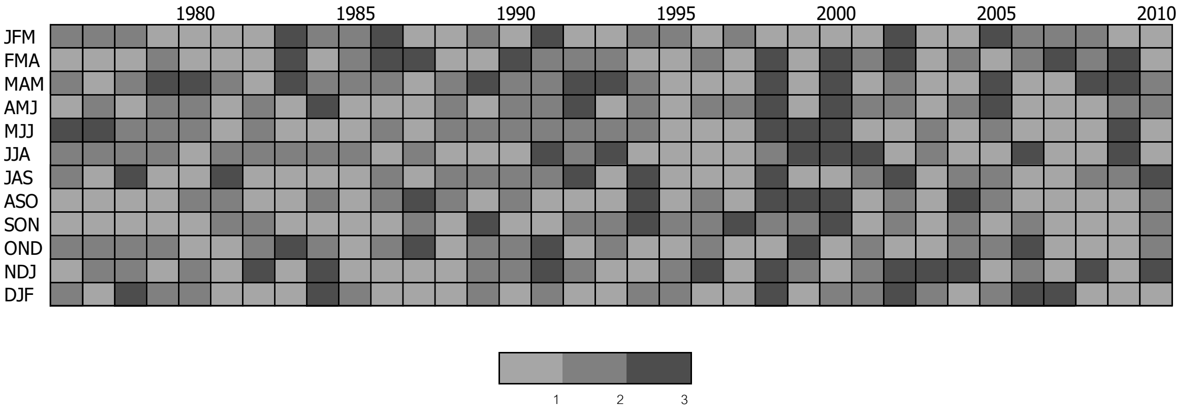

Figure 5 Evaluation of performance of model in forecasting of 3 months rainfall over Mun river basin

5. Evaluation of model performance

In this study, the likelihood skill score (LLH) is selected as a technique to evaluate the stochastic statistical k-nn model. The likelihood skill score (LLH) presents how will the estimated value capture the Probability Density Function (PDF) of the climatology. In this study, the values of LLH vary from 0.0 to +3.0. The score of +1.0 indicates similarity between the model performance and reference simulated climatology. A score of less than +1.0 indicate the weaker performance of the model compare to climatology. Otherwise the score of higher than +1.0 indicates a better performance of model.

In this study, the observed rainfall is divided into 3 equal sets based on the defined 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.

In order to calculate value of LLH, the categorical probabilities of climatology has to be calculated. The categorical probabilities of climatology are the proportion of rainfall in each category. Since in this study, historical data is divided into 3 equal sets, the categorical probability of all three categories is 1/3.

For a given year, the N simulated ensembles are also divided into 3 categories using the same thresholds i.e. the 33th and 67th percentile. The categorical probabilities of ensembles in a given year, which are the proportion of rainfall ensembles in each category, are computed. Subsequently, LLH can be calculated using the following equation:

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

Where n is the number of years,

J is the category of the observed rainfall in year

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

In this study, LLH values were computed for all of forecasted seasonal rainfall during the period of 1975 to 2011. Figure 5 presents the result of model performance evaluation in forecasting 3 months rainfall. If reveals that the model can be used to present seasonal rainfall with the reliability of around 60%.

6. Conclusions

An attempt of forecasting of seasonal rainfall over the Mun river basin, the poorest region of Thailand, was done in study. The model is aimed to be able to forecast seasonal rainfall with leading time of year ahead. The stochastic statistic k-nn model was adopted in study using LAV as predictors in the model. Amongst a variety of LAV, SAT, SLP, U and V over China Sea and Pacific Ocean are main predictors in the model due to their high correlations with monthly rainfall over the Mun river basin. The evaluation of model performance, using LLH technique, was conducted. It reveals that the model is able to forecast seasonal rainfall over the Mun river basin with the leading time from one month to one year with reliability of around 60% that rainfall ensemble well with the observed ones. It is therefore a useful tool to provide preliminary information for appropriate cropping pattern planning, which will consequently lead to effective water resources planning and management in the drought area. This would reduce crop damage due to drought and moisture stress. However more improvement in selection and modification of predictors as well as prediction techniques should be further developed in order to obtain more reliable forecast.

7. Acknowledgements

The authors would like to thank the Hydro and Agro Informatics Institute (HAII), Thailand for research fund and valuable suggestion for the model development.

8. References

[1] Paxson CH. Using weather variability to estimate the response of saving to transitory income in Thailand. Am Econ Rev. 1992;82(1):15-32.

[2] Boonchabun K, Tych W, Chappell NA, Lorsirirat K, Pa-Obsaeng S. Statistical modeling of rainfall and river flow in Thailand. J Geol Soc India. 2004;64: 503-16.

[3] Turks M. Spatial and temporal analysis of annual rainfall variations in Turkey. Int J Clim. 1996;6: 1057-76.

[4] De Luís M, Raventós J, González-Hidalgo JC, Sánchez JR, Cortina J. Spatial analysis of precipitation trends in

the region of Valencia (East Spain). Int J Climatol. 2000;20:1451-69.

[5] Gonzalez-Hildago JC, De Luís M, Raventos S, Sanchez JR. Spatial distribution of seasonal rainfall trends in a western Mediterranean area. Int J Clim Int J Climatol. 2001;21:843-60.

[6] Cannarozzo M, Noto LV, Viola F. Spatial distribution of rainfall trends in Sicily (1921-2000). Phys Chem Earth. 2006;31:1201-11.

[7] Zhang Q, Xu C-Y, Chen YD. Spatial and temporal variability of extreme precipitation during 1960-2005 in the Yangtze River basin and possible association with large-scale circulation. J Hydrol. 2008;353: 215-27.

[8] Zhang Q, Xu C-Y, Zhang ZX. Observed changes of drought/wetness episodes in the Pearl River Basin, China, using the standardized precipitation index and aridity index. Theor Appl Climatol. 2009;98:89-99.

[9] Zhang Q, Xu C-Y, Zhang Z, Chen YD, Liu C-L. Spatial and temporal variability of precipitation during 1951–2005 over China. Theor App Climatol. 2009;95:53-68.

[10] Zhang Q, Xu C-Y, Becker S, Zhang ZX, Chen YD, Coulibaly M. Trends and abrupt changes of precipitation maxima in the Pearl River Basin, China. Atmosph Sc Lett. 2009;10:132-44.

[11] Zhang Q, Xu C-Y, Zhang ZX, Chen X, Han Z. Precipitation extremes in a karst region: a case study in the Guizhou Province, Southwest China. Theor App Climatol. 2010;101:53-65.

[12] Chu P-S, Chen YR, Schroeder TA. Changes in precipitation extremes in the Hawaiian Islands in a warming climate. J Climat. 2010;23:481-90.

[13] Mitchell JFB. The ‘greenhouse’ effect and climate change. Rev Geophys. 1989;27:115-39.

[14] UNEP (United Nations Environment Programme). How will global warming affect my world?. Geneva: UNEP; 2003.

[15] Maslin M. Global warming: causes, effects and the future. Minneapolis, MN: MBI Publishing Company LLC; 2007.

[16] NIC (National Intelligence Council). Southeast Asia and pacific islands: the impact of climate change to 2030. Spec Rep NIC2009-006D. Washington, DC: NIC; 2009.

[17] NASA. The ups and downs of global warming [Internet]. 2010. Available from: www.nasa.gov /topics/earth/features/ ups Downs Global Warming. html

[18] Chen TC, Yoon JH. Inter-annual variation in Indochina summer monsoon rainfall: Possible mechanism. J Clim. 2000;13:1979-86.

[19] Singhrattna N, Rajagopalan B, Kumar KK, Clark M. Inter-annual and inter-decadal variability of Thailand summer monsoon season. J Clim. 2005;18:1697-708.

[20] Singheattna N, Mukand SB. Changes in summer monsoon rainfall in the upper Chao Phraya River Basin, Thailand. Clim Res. 2011;49:155-68.

[21] Weesakul U, Singhratta N, Luangdilok N. Rainfall forecast in northeast of Thailand using modified K-nearest neighbor. KKU Eng J. 2014;41(1):1-10.

[22] Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, et al. The NCEP/NCAR reanalysis 40-year project. Bull Am Meteorol Soc. 1996; 77(3):437-71.

[23] Weesakul U, Singhratta N, Yodpongpiput P. Statistical relationships between large-scale atmospheric variables and rainfall in Mun River Basin (Thailand). 5th National Convention on Water Resources Engineering; 2013 Sep 4-7; Chiang Rai, Thailand. Thailand: Engineering Institute of Thailand; 2013.

[24] Haan CT. Statistical Method in Hydrology. 2nd ed. USA: Iowa State Press; 2002.

[25] ESRL (Earth System Research Laboratory) [Internet]. 2011. Available from: http://www.esrl.noaa.gov /psd/data/correlation.