Rainfall Prediction in the Northeast Region of Thailand using Modular Fuzzy Inference System

Jesada Kajornrit[1], Kok Wai Wong[2], Chun Che Fung[3]
School of Information Technology, Murdoch University
South Street, Murdoch, Western Australia, 6150
Email: j_kajornrit@hotmail.com[1], k.wong@murdoch.edu.au[2], l.fung@murdoch.edu.au[3]

***Abstract*—In water management systems, accurate rainfall forecasting is indispensable for operation and management of reservoir, and flooding prevention because it can provide an extension of lead-time of the flow forecasting. In general, time series prediction has been widely applied to predict rainfall data. The conventional time series prediction models or artificial neural networks can be used to perform this task. However, such models are difficult to interpret by human analyst. From a hydrologist's point of view, the accuracy of the prediction and understanding the prediction model are equally important. This study proposes the use of a Modular Fuzzy Inference System (*Mod* FIS) to predict monthly rainfall data in the northeast region of Thailand. The experimental results show that the proposed model can be a good alternative method to provide both accurate results and human-understandable prediction mechanism.**

Keyword-Rainfall Prediction; Seasonal Time Series; Fuzzy Inference System; Northeast Region of Thailand

I. INTRODUCTION

In water management systems, accurate rainfall forecasting is very important because it can provide an extension of lead-time of the flow forecasting used in reservoir operation and flooding prevention. Many time series prediction models have been developed to perform this task such as the Box-Jenkins (BJ) and Artificial Neural Networks (ANN) [1]. However, such models are difficult to be interpreted by human analyst since the prediction mechanism requires comparatively complex mathematical modeling. From a hydrologist's point of view, the accuracy of prediction and understanding of the prediction model are equally important. Therefore, it is the aim of this study to develop an alternative method to achieve the said objectives.

A Fuzzy Inference System (FIS) uses the process of mapping from a given set of input variables to an output based on a set of human understandable fuzzy rules [2]. FIS has been successfully applied in various applications, such as pattern recognition, data analysis and system control [3], [4]. An advantage of the FIS is that the mechanism of the FIS model could be interpretable by human. As fuzzy rules are closer to human reasoning, the analyst could understand how the model performs prediction. If necessary, the analyst could also make use of his/her knowledge to modify the prediction model [5].

In hydrological time series prediction, FIS is not as popular as the ANN approach because FIS lacks of learning ability. However, taking the advantages of FIS into account, it is worth investigating the use of FIS to the time series prediction problem. Thus, the main objective of this study is to investigate the use of FIS in an effective way for the rainfall time series prediction. This study also proposed the *Modular* FIS (*Mod* FIS) to the time series prediction problem. Such model is easy to interpret by human analysts and it provides the prediction performance as good as BJ or ANN models [7].

This paper is organized as follows; Section 2 discusses related works. The concept of FIS is briefly introduced in Section 3. Case study area and the proposed *Mod* FIS are described in Section 4 and 5 respectively. Section 6 shows the experimental results. Finally, Section 7 provides the conclusion of this paper.

II. RELATED WORKS

Rainfall prediction is relatively more difficult when compares to other climate variables such as temperature. This is because of the highly stochastic nature in rainfall estimation, which consists of complex spatial and temporal features. Coulibaly and Evora [6] compared six different ANNs to predict daily missing rainfall data. Among the different types of ANN, Multilayer Perceptron, Time-lagged Feedforward Network, and Counter-propagation Fuzzy-neural Network provided higher accuracy than the Generalized Radial Basis Function Network, Recurrent Neural Network and Time Delay Recurrent Neural Network. Wu et al. [7] proposed the use of data-driven models with data preprocessing techniques to predict precipitation data in daily and monthly scales. They proposed three preprocessing techniques, namely, Moving Average, Principle Component Analysis (PCA) and Singular Spectrum Analysis to smoothen time series data. Somvanshi et al. [8] compared ANN and Auto -Regressive Integrated Moving Average (ARIMA) for rainfall prediction. He concluded that ANN provided better accuracy than the ARIMA model.

Time series prediction is not only used for precipitation data but also other hydrological data such as streamflow. Wang et al. [9] compared several artificial intelligence models, namely, Auto-Regressive Moving Average (ARMA), ANN, Adaptive Neural-Fuzzy Inference System (ANFIS), Genetic Programming (GP) and Support Vector Machine (SVM) to predict monthly discharge time series. The results indicated that the best performance could be obtained by ANFIS, GP and SVM. Wu et al. [10] compared performance of data-driven model to forecast monthly streamflow. The results showed that ARMA and K-Nearest-Neighbors (KNN) per-

formed prediction better than ANN and its variants when the correlation between input and output was low. Lohani [11] compared ANN, FIS and linear transfer model for daily rainfall-runoff model under different input domains. The results showed that FIS outperformed linear model and ANN. Nayak et al. [12] and Kermani et al. [13] introduced ANFIS model to river flow time series. Jain and Kumar [14] applied conventional preprocessing approaches (de-trended and de-seasonalized) to ANN for streamflow time series data. Overall, FIS itself is not as popular as the ANN or BJ models for time series prediction. Especially, for rainfall time series prediction, applications of FIS are limited. Thus, the primary aim of this study is to investigate an appropriate way to use FIS for rainfall time series prediction problem.

III. Fuzzy Inference System

FIS is a process of mapping given inputs to outputs by using the fuzzy set theory [15]. FIS is an appropriate technique to be applied to the hydrology problem since FIS allows variables "partial true" and/or "partial false", which reflect the uncertainty nature in physical processes. FIS consists of five components: (i) A rule base involves IF-THEN rules mapping the relations between inputs and outputs. (ii) A database that collects the membership functions (MFs) of each input and output variables. (iii) A fuzzification process that fuzzifies crisp inputs into fuzzy set inputs. (iv) A defuzzification process that defuzzifies fuzzy set outputs into crisp outputs. (v) An inference engine that is the logic decision system using IF-THEN the rules from rule base module and membership functions from the database module.

Basically, there are two typical methods to defuzzify fuzzy sets outputs and they are Mamdani [16] and Sugeno [17] approaches. The Mamdani approach defuzzifies output fuzzy sets by finding the centroid of a two-dimensional shape by integrating across a continuously variation function. In the Sugeno approach, output fuzzy sets are in the form of singleton, a fuzzy set with unity membership grade at a singleton point and zero everywhere else on the universe of discourse. The output centroid is calculated by the weighted average method. In this study, the Mandami approach is used because it is intuitive and well suited for human understanding [18].

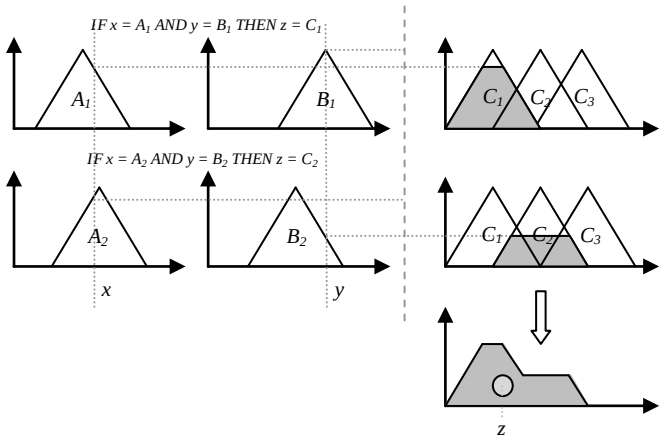


Figure 1. Fuzzy inference system used in this study

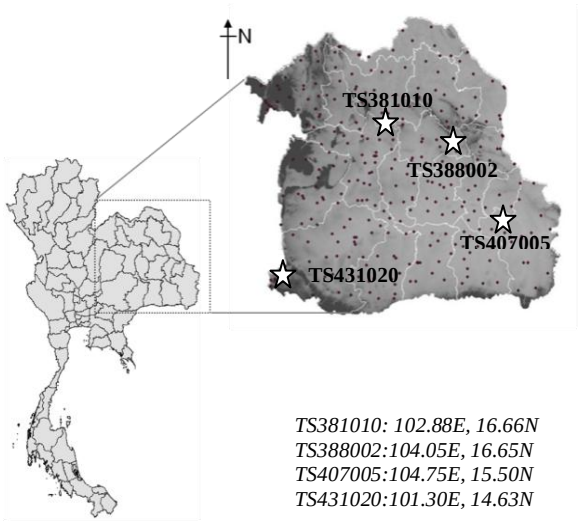


Figure 2. The case study area sites in the northeast region of Thailand

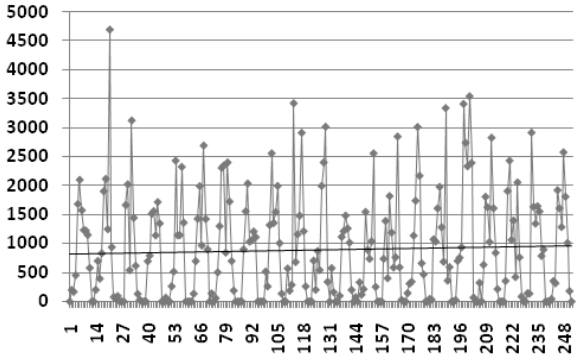
The fuzzy inference process used in this study includes four steps (see Figure. 1): First step, the crisp inputs are fuzzified into fuzzy set inputs based on membership functions, for example, x is fuzzified by A_1 and A_2 membership functions and y is fuzzified by B_1 and B_2. Second step, those fuzzy sets are inferred by IF-THEN rules in the knowledge base and provide the fuzzy set outputs. In this step the *AND* operation in rule base are replaced with *MIN* operator. In Figure 1, for example, C_1 results from rule *"IF $x = A_1$ AND $y = B_1$ then $z = C_1$"* in a certain membership grade. Next step, fuzzy set outputs from each rule are aggregated. In this step, the *Max* operator is adopted. Once all the fuzzy outputs from every fuzzy rule are aggregated, the fuzzy output will be defuzzified into a single crisp output.

IV. Case Study Area and Data

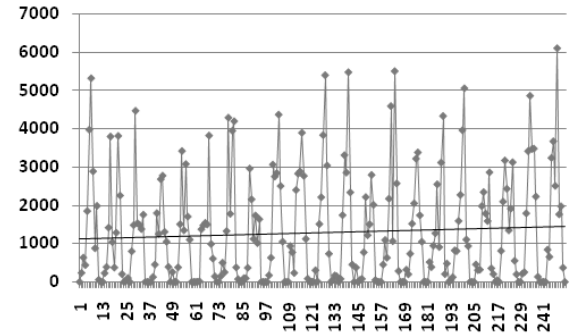
The case study selected sites in the northeast region of Thailand (Figure 2). Four rainfall time series selected in this study are depicted in Figure 3. Table 1 shows the statistics of the datasets. The linear fit (linear line) is used to verify the consistency of time series. It is evident that linear lines are not parallel to horizontal axis, especially in TS388002 and TS431020. Therefore, selecting the time period to create the model must be handled with care. The datasets range from years 1981 to 2001. Since the linear trend appears strongly in earliest period, only data from year 1989 to 1998 are used to create models and data from 1999 to 2001 are used to validate the models.

This study will predict 1 step ahead, that is, 1 month. To validate the models, the Mean Absolute Error (MAE) is used. The mathematic formula of MAE is shown in (1), In addition, the Coefficient of Fit R is also used to assess the results. The performance of the proposed model is compared with conventional BJ models (Autoregressive, AR and Seasonal Autoregressive Integrated Moving Average, also known as SARIMA) and ANN [1], [9], [10], [12], [13], [14].

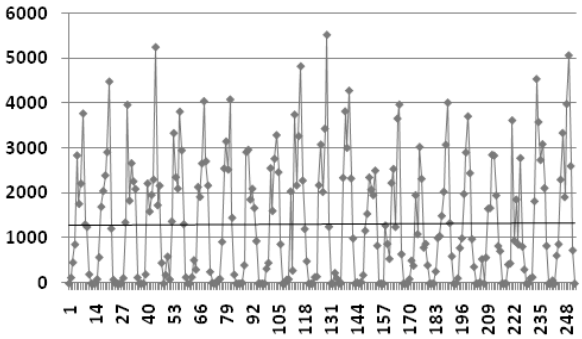
$$MAE = \sum_{i=1}^{m} |Oi - Pi| / m \quad (1)$$



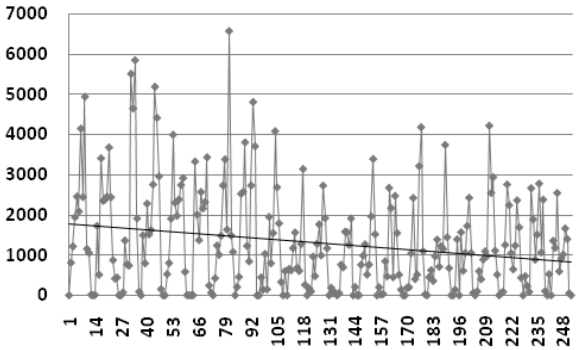
(TS381010)



(TS388002)



(TS407005)



(TS431020)

Figure 3. Four monthly rainfall time series in this study

TABLE I. DATASET' S STATISTIC

Statistics	TS381010	TS388002	TS407005	TS431020
Mean	889.04	1286.28	1319.70	1296.35
SD	922.99	1425.88	1346.80	1289.01
Kurtosis	0.808	0.532	-0.224	1.590
Skewness	1.080	1.131	0.825	1.276
Range	4704	6117	5519	6558
Minimum	0	0	0	0
Maximum	4704	6117	5519	6558

V. MODULAR FUZZY INFERENCE SYSTEM

Based on the modular concept used in [7] and [19], the proposed model consists of twelve monthly FIS sub-models. These monthly sub-models are used to predict the rainfall associated to the months in a year. An advantage of using sub-models is that the historical data will be examined in both seasonal and non-seasonal level at the same time. The architecture of the proposed model is depicted in Figure 4.

In Figure 4, the parameters (MFs) of FIS sub-models are derived from the time series data at the seasonal level, whereas the inputs of FIS sub-models come from the time series data in non-seasonal level. For example, for an two-input and one-output FIS sub-model, $FIS_{(July)}$, this sub-model uses rainfall patterns of *May-June-July* period from every year to create the model at seasonal level. On the other hand, when testing the model, data from May and June in present year are used as the inputs at non-seasonal level and the July data is the output. In the prediction process, inputs will be fed to an associated sub-model and derives the output from the fuzzy inference engine. The methodology used to create FIS models consists of three steps, namely, defining universe of discourse, defining membership functions and constructing fuzzy rules.

The first step is to define the universe of discourse. Twelve universes of discourse are defined according to the month. Ten-year time series data (training data) are overlaid in order to observe the rainfall distribution in every month of a year. Figure 5 illustrates an example of the overlaid ten-year data (TS381010). It can be seen from Figure 5 that in each month of a year, the distribution of rainfall varies within a certain range. Therefore, it is not necessary to define membership function outside the range of rainfall in that month. This could effectively reduce computational time and the number of fuzzy rules.

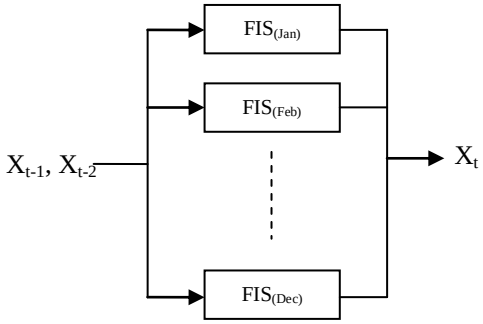


Figure 4. An architectural overview of the proposed model

For each month, the rainfall data are clustered in order to create the MFs. At this point, the K-Mean clustering technique is used. It can be seen from Figure 5 that the distribution of rainfall is small in January and it gradually increases up to August. After August, the rainfall distribution continues to decrease until the end of the year. When considering the distribution of the rainfall data along the year from every dataset, the appropriate K should be: K = 2 for January and December, K = 3 for February and November and K = 5 for the rest of a year. Infrequently, however, K = 6 is used if the rainfall distribution is higher than usual.

The second step is to create the MFs. The triangle membership function is adopted in this study. The advantages of this MF over other MF (for example, Gaussian MF) are that it consumes less computational resources due to its simple form. The equation is given as follows.

$$\begin{aligned} \mu_x(a,b,c) &= 0; && x < a \\ &= (x-a)/(b-a); && a \le x \le b \\ &= (c-x)/(c-b); && b \le x \le c \\ &= 0; && c < x \end{aligned} \qquad (2)$$

Each cluster is mapped to a set of membership function. For example, when the data are clustered into 5 membership functions, those memberships are represented as very low (VL), low (L), medium (M), high (H) and very high (VH). The centroid of each cluster is the peak of the triangular MF, which has full membership function value as 1. The intersected area between consecutive MFs is set to 50 percents.

The last step is to construct the fuzzy rules. The number of inputs has an effect on the overall system. By considering the appropriate number of inputs intuitively coupled with the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of data, the results indicates that using data from first lag and second lag should be the optimal selection. To create fuzzy rules, the input-output pairs of training data are mapped directly to the clusters.

In the process of mapping data to MFs, infrequently, it is possible that some conflicted rules appear. These conflict rules must be resolved or removed. When conflicted rules appear, the rule that occurs more frequently is selected. This criterion can correspond to normal rainfall event. In rare cases, if the numbers of conflicted rules are equal, the rule that occurs latest is selected. This assumption is based on the hypothesis that the latest rainfall event will probably occur again.

VI. Result and Analysis

To evaluate the prediction accuracy of the reported model, the time series data between 1999 and 2001 were used for validation. This period was not included in the model calibration. The MAE and R measures of validation period are shown in Table II and Table III respectively. Figure 6 illustrates both validation measures. According to Figure 6, the order of prediction accuracy of those models are *Mod* FIS > SARIMA > AR > ANN in general, with *Mod* FIS as the highest accuracy. As observed from the results of MAE and R values, these experimental results could be considered consistent. Figure 8 shows

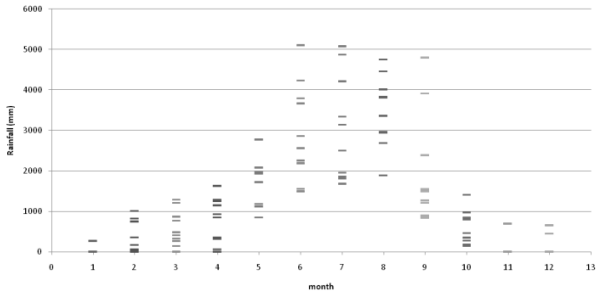


Figure 5. An example of monthly rainfall distribution of overlaid ten-year training data (TS381010)

the plots between the observation values and the predicted values of *Mod* FIS.

AR model is one of the Box-Jenkins' models for time series prediction and it has been commonly used in hydrology studies [14]. In this study, the AR model uses degree 2 because it uses the same input as *Mod* FIS (two previous non-seasonal lags). From Figure 6, the *Mod* FIS show better prediction accuracy than AR model in all the datasets. As mentioned before that rainfall data is highly stochastic in nature, so it is difficult for the linear model such as AR to capture non-linearity in data. In contrast, the *Mod* FIS can capture non-linearity in data because such model derives the prediction results based on fuzzy membership functions and fuzzy rules.

SARIMA model is an improvement over the AR model because it considers the data in both seasonal and non-seasonal levels when the model was created. In the SARIMA model, the prediction is derived from a linear equation and the input data come from seasonal and/or non-seasonal levels. Since SARIMA model use input data from both seasonal and non-seasonal lag, one can assume that the SARIMA model should provide more accurate prediction than the *Mod* FIS that uses data only from non-seasonal level. However, in this experiment, the *Mod* FIS prediction was better than the SARIMA model in 3 out of the 4 datasets. This shows that the proposed method is quite versatile.

TABLE II. MAE Measures

DATA	TS381010	TS388002	TS407005	TS431020
AR	534	923	890	741
SARIMA	503	716	621	524
Mod FIS	454	550	576	563
ANN	612	769	973	819

TABLE III. R Measures

DATA	TS381010	TS388002	TS407005	TS431020
AR	0.463	0.594	0.603	0.281
SARIMA	0.577	0.762	0.782	0.640
Mod FIS	0.649	0.895	0.819	0.578
ANN	0.389	0.769	0.546	0.157

From these experimental results, it can be inferred that seasonal rainfall time series data are not smooth in seasonal level. Even though the differencing method is applied to time series to satisfy the stationary condition, it is still difficult for the SARIMA model to provide accurate predication results. In TS431020, rainfall data is rather highly irregular in non-seasonal level because this station is located in mountainous area. This causes some problems for the proposed *Mod* FIS model to perform effectively because the FIS uses data at non-seasonal level as input.

The most commonly used ANN model in hydrology is the three layer back-propagation neural networks and has been adopted as a comparison technique in this study. The number of hidden nodes is three. In Figure 6, the ANN models showed the lowest prediction accuracy. More experiments were performed to investigate the cause of low performance of the ANN model in this study. One possible reason is that the number of training data is relatively small. To verify this hypothesis, training data is increased to 18 years (1981-1989), the performance of ANN is improved as shown in Figure 7. However, the *Mod* FIS and SARIMA methods are still slightly better when compared to the ANN models. Another possible reason is that the time series used in this study is periodic. Wu et al. [7] has also provided similar observation in their study, in which ANN performed well on daily rainfall data but not in the monthly rainfall data. In the monthly rainfall data, there are only twelve data points in one cycle. Furthermore, the data are not as smooth as those in the seasonal level data.

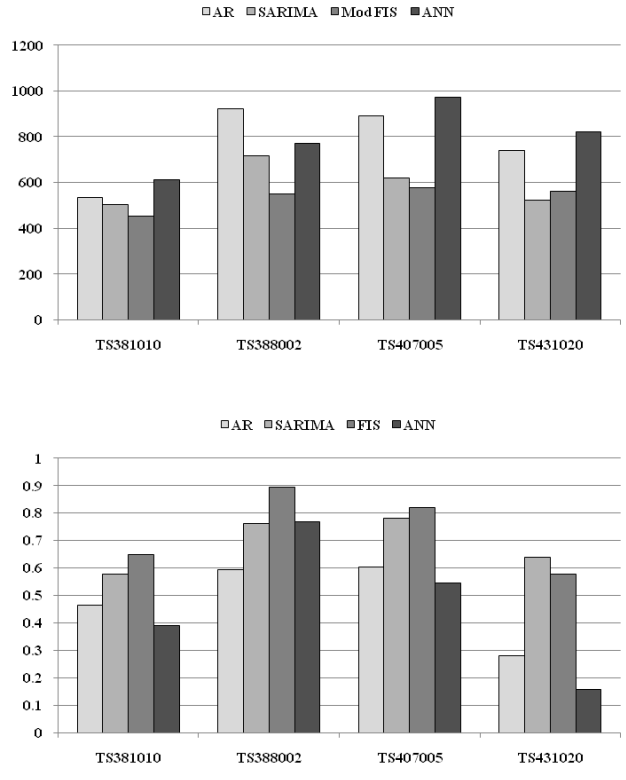


Figure 6. (Top) MAE measure and (Bottom) R measure of validation period

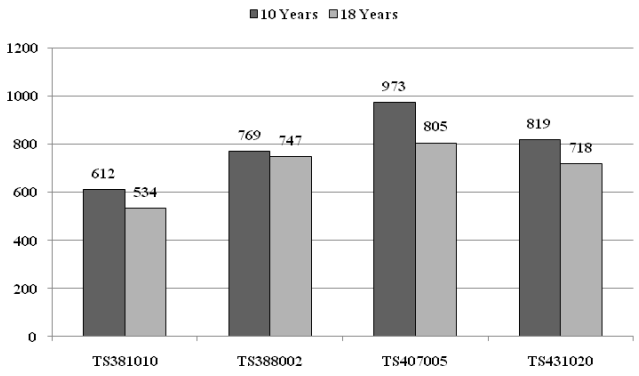
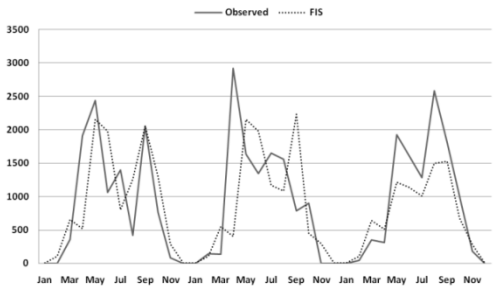


Figure 7. MAE of ANN after increasing training data to 18 years.

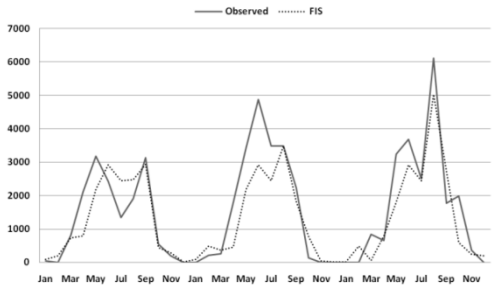
At this point, the performance of the proposed model has been evaluated. Another important feature of the proposed model is that the prediction mechanism is easily interpretable by human analyst through fuzzy rules. However, solely based on this point may not be strong as a significant advantage of the proposed model. Another feature of the proposed model is that it decomposed one model into monthly sub-models to reduce the number of fuzzy rules in the system. Even though the fuzzy rules are close to human reasoning and interpretable, a large number of fuzzy rules may not be appropriate to human analyst. According to "*the curse of dimensionality*", the number of fuzzy rules could increase exponentially when the number of input increases. Supposed a FIS model with five MFs for each input dimensions, the number of complete fuzzy rules will be 5 x 5 x 5 x 5 = 625 rules for 4 inputs. This number of rules may not be easy to be handled in practical. Since the proposed model decomposes one large model into twelve sub-models and use only 2 inputs, the number of fuzzy rules is not more than 25 rules for each sub-model, which is more practical for human analyst.

VII. Conclusion

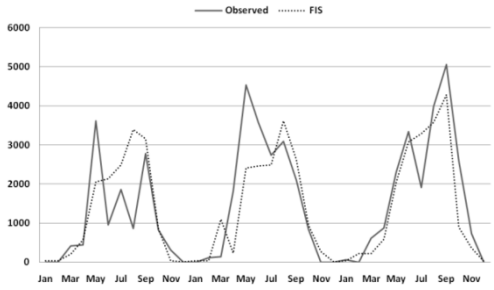
An accurate rainfall forecasting is crucial for reservoir operation and flooding prevention because it can provide an extension of lead-time of the flow forecasting. Many time series prediction models have been applied to give accurate results. However, the prediction mechanism of those models may be difficult to be interpreted by human analyst. This study investigated the use of modular fuzzy inference system to predict monthly rainfall time series in the northeast region of Thailand. The prediction performance of the proposed model was compared to the conventional Box-Jenkins and artificial neural networks models. The experimental results showed that the proposed model can be a good alternative method to provide accurate prediction. Furthermore, the prediction mechanism can be interpreted through fuzzy rules. The following are some directions for future works in order to enhance the proposed model. First, since the proposed FIS model lacks of self-learning ability, it may cause a problem when the dataset are large. In this case, the proposed FIS needs a supplementary procedure to create the model. Second, the number of MFs is still defined intuitively by the expert's experience; is it possible to define MFs automatically and appropriately from the characteristics of time series data?



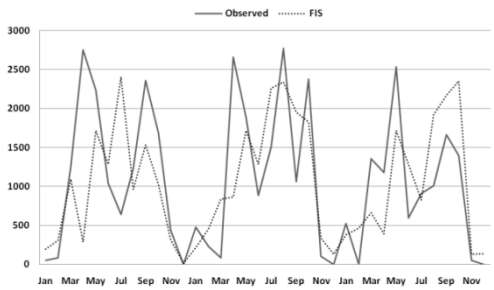
(TS381010)



(TS388002)



(TS407005)



(TS431020)

Figure 8. Monthly rainfall between observation and predicted value of *Mod* FIS models

REFERENCES

[1] H. Raman and N. Sunilkumar, "Multivariate modeling of water resources time series using artificial neural network*," Hydrological Sciences –Journal- des Sciences Hydroligiques*, vol. 40, pp.145-163, 1995.

[2] Z. F. Toprak, et al., "Modeling monthly mean flow in a poorly gauged basin by fuzzy logic," *Clean*, vol. 37, no. 7, pp. 555-567, 2009.

[3] S. Kato and K. W. Wong, "Intelligent Automated Guided Vehicle with Reverse Strategy: A Comparison Study," in Mario Köppen, Nikola K. Kasabov, George G. Coghill (Eds.) Advances in Neuro-Information Processing, Lecture Notes in Computer Science, Springer-Verlag, Berlin Heidelberg, 2009, pp. 638-646. 2009"

[4] K. W. Wong, and T. D. Gedeon, "Petrophysical Properties Prediction Using Self-generating Fuzzy Rules Inference System with Modified Alpha-cut Based Fuzzy Interpolation", Proceedings of The Seventh International Conference of Neural Information Processing ICONIP, pp. 1088-1092, November 2000, Korea.

[5] K. W. Wong, P. M. Wong, T. D. Gedeon, C. C. Fung, "Rainfall Prediction Model Using Soft Computing Technique," Soft Computing, vol 7, issue 6, pp. 434-438, 2003

[6] P. Coulibaly and N. D. Evora, "Comparison of neural network methods for infilling missing daily weather records." Journal of Hydrology, vol. 341 pp. 27-41, 2007.

[7] C. L. Wu, K. W. Chau, and C. Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques." Journal of Hydrology, vol. 389, pp.146-167, 2010.

[8] V. K. Somvanshi, et al., "Modeling and prediction of rainfall using artificial neural network and ARIMA techniques." J. Ind. Geophys. Union, vol. 10, no. 2, pp. 141-151, 2006.

[9] W. Wang, K. Chau, C. Cheng and L. Qiu, "A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series." Journal of Hydrology, vol. 374, pp. 294-306, 2009.

[10] C. L. Wu and K. W. Chau, "Data-driven models for monthly streamflow time series prediction." Engineering Applications of Artificial Intelligence, vol. 23, pp. 1350-1367, 2010.

[11] A. K. Lohani, N. K. Goel and K. K. S. Bhatia, "Comparative study of neural network, fuzzy logic and linear transfer function techniques in daily rainfall-runoff modeling under different input domains." Hydrological Process, vol. 25, pp. 175-193, 2011.

[12] P. C. Nayak, et al., A neuro-fuzzy computing technique for modeling hydrological time series, Journal of Hydrology, vol. 291. pp. 52-66, 2004.

[13] M. Z. Kermani and M. Teshnehlab, "Using adaptive neuro-fuzzy inference system for hydrological time series prediction." Applied Soft Computing, vol. 8, pp. 928-936, 2008.

[14] A. Jain and A. M. Kumar, "Hybrid neural network models for hydrologic time series forecasting." Applied Soft Computing, vol. 7, pp. 585-592, 2007.

[15] L. A. Zadeh, "Fuzzy Sets," Inform and Control, vol. 8, pp. 338 – 353. 1965.

[16] E. H. Mamdani, and S. Assilian, "An experiment in linguistic synthesis with fuzzy logic controller," International journal of man-machine studies, vol. 7 no. 1, pp.1-13, 1975.

[17] M. Sugeno, "Industrial application of fuzzy control," North-Holland, Amsterdam. 1985.

[18] Fuzzy logic toolbox™ 2.0 – User's guide R2011b, Matlab

[19] C. C. Fung, K. W. Wong, H. Eren, R. Charlebois, and H. Crocker, "Modular Artificial Neural Network for Prediction of Petrophysical Properties from Well Log Data," in IEEE Transactions on Instrumentation & Measurement, 46(6), December, pp.1259-1263, 1997.