

Using multiple linear regression model to estimate thunderstorm activity

W Suparta^{1,*} and W S Putro¹

¹ Space Science Centre (ANGKASA), Institute of Climate Change, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia.

Email: wayan@ukm.edu.my

Abstract. This paper is aimed to develop a numerical model with the use of a nonlinear model to estimate the thunderstorm activity. Meteorological data such as Pressure (P), Temperature (T), Relative Humidity (H), cloud (C), Precipitable Water Vapor (PWV), and precipitation on a daily basis were used in the proposed method. The model was constructed with six configurations of input and one target output. The output tested in this work is the thunderstorm event when one-year data is used. Results showed that the model works well in estimating thunderstorm activities with the maximum epoch reaching 1000 iterations and the percent error was found below 50%. The model also found that the thunderstorm activities in May and October are detected higher than the other months due to the inter-monsoon season.

Keywords: Multiple linear regression model, Meteorology data, Thunderstorm estimation

1. Introduction

Mesoscale convective systems (MCS) are large precipitation proportion in the convective system. MCS is an ensemble of thunderstorms with high precipitation index of 100 km or more in one direction [1]. The thunderstorm phenomena are categorized as the most deadly natural hazard in the world. Thunderstorm phenomena are electrical storms where cumulonimbus cloud produced lightning during heavy rain and thunder during precipitation. The estimation of thunderstorm frequency over MCS area is important to commercial space vehicle launch operation and power utilities. Suparta et al. [2] reported that the MCS activity in Tawau has high precipitation level that increased during the winter season (December, January, and February) and decreased during the summer season in June, July, and August. This showed that the MCS contributed to thunderstorm level in the Tawau area and hazardous for space activities.

Based on the impact of thunderstorm activity over Tawau area, this study aimed to estimate the thunderstorm frequency. Previous studies have shown that the simulation of thunderstorm activity is unable to evaluate the impact of a fault during an individual storm event [3, 4]. Based on these previous studies, it is suggested that Jacobi algorithm is applied in the multiple linear regression (MLR) to estimate thunderstorm activities. Six combination parameter inputs from meteorological parameter were constructed to develop a linear equation. From the meteorology data provided at each point of the Earth, the model can be easily implemented to other locations.



2. Methodology

2.1. Data and Location

Tawau is located in Sabah at East Malaysia (see Figure 1). The climate condition in this area is relatively hot and wet with average shade temperature about 26°C to 29°C at noon and falling to around 23°C at night. Looking the relationship between meteorology parameter and thunderstorm occurrence, the parameter inputs such as Pressure (P), Temperature (T), Relative Humidity (H), Cloud (C), Precipitable Water Vapor from radiosonde (RSPWV), and precipitation (Pr) were used to evaluate the variation of climate characteristic for thunderstorm activity. Data was taken over one year (1 January ~ 31 December 2012) for each parameter. The observation data were taken from Climate & Hydrological Section – Malaysian Meteorological Department (MetMalaysia) at TWU station. Moreover, RSPWV data and precipitation data were taken from the University of Wyoming website (<http://weather.uwyo.edu/upperair/sounding.html>) and the NASA website (<http://gdata1.sci.gsfc.nasa.gov>), respectively.

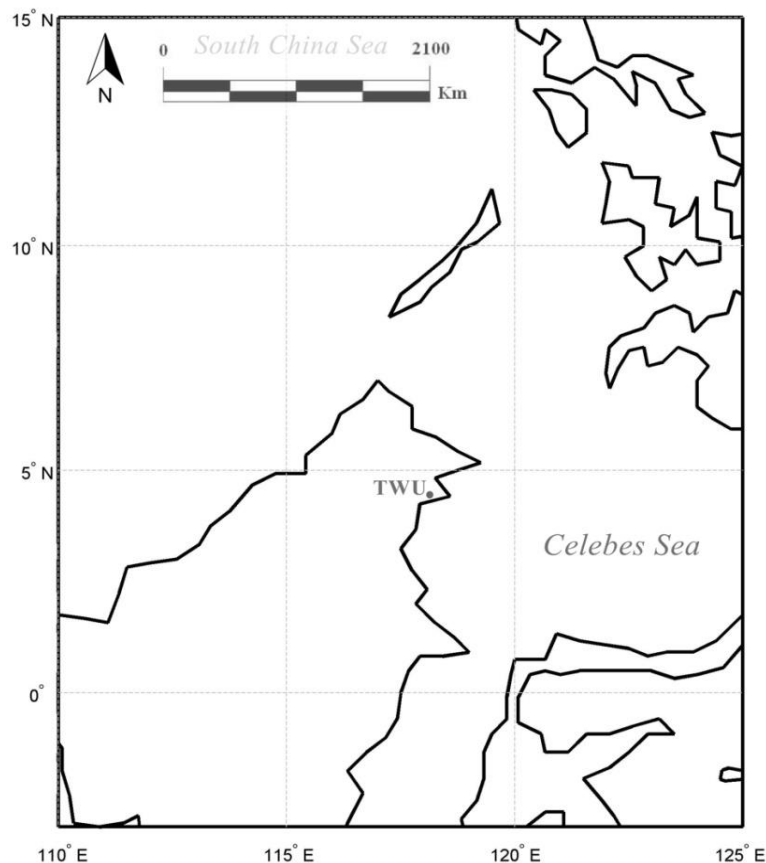


Figure 1. The location of TWU station for meteorological data observation

2.2. Data Processing

In this study, daily data were processed using MATLAB program to filter and enhance the quality of raw data. Figure 2 showed the flowchart of data processing for thunderstorm identification using meteorology parameters. The combinations of parameters were proposed to create a linear equation. The training data were processed to obtain estimation model with 1,000 iterations using the Jacobi algorithm. Five equation models were suggested to estimate the thunderstorm frequency based on the three seasons in 2012 which were summer, winter, and transition season (inter-monsoon). In this work, to design configuration input and output observation data using MLR method, the correlation

and regression analysis to obtain R-square value between input (meteorological data, e.g. P , T , H , C , RSPWV, and Pr) and output (thunderstorm data) is performed.

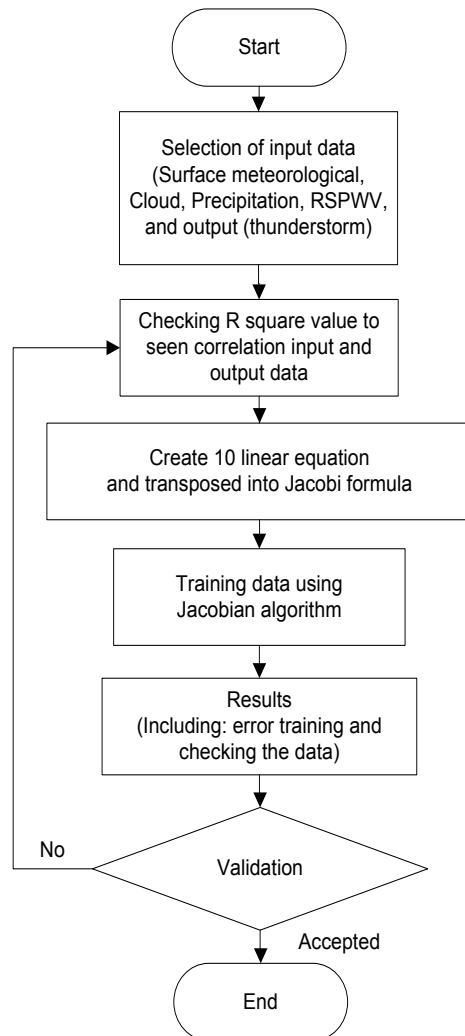


Figure 2. Flowchart of data processing and identification

3. Results and Discussion

3.1. Meteorology Parameter

Figure 3 showed the variation of surface meteorology (P , T , and H), Cloud, RSPWV, and Pr from 1 January 2012 to 31 December 2012 as inputs the model. The variation of the first three of the parameter (P , T , and H) is observed in April, June and September reached the pressure to above 1010 mbar, June and August reached the temperature to above 28°C, January, February, and December reached the relative humidity to above 85%. Furthermore, the last three of parameter showed an extreme value of cloud density, RSPWV, and precipitation in June and July reached to above 7 oktas, February and December reached to above 80 mm, and February, May, and October reached to above 55 mm/day, respectively. On the other hand, winter I and winter II from Figure 3 is estimated from the month of January – February and December, respectively. The summer is defined from June to August and the other months are intermonsoon. The minimum precipitation and high temperature indicated that the clear day was obtained in the middle of July. In addition, the rainy day during the winter season occurred in January and February 2012 and December.

2.3 Multiple Linear Regressions

The multiple linear regressions (MLR) equation can be described as an *algebraic equation* in which each term either a constant or the product of a constant and a single or more variable (linear equation). The multiple linear regressions are useful to characterize estimation model for any phenomena based on mathematics methods. In general, the MLR is written as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i \quad (1)$$

where the least-squares estimates b_0 , b_1 , ... b_p are usually computed by statistical software, and the e_i are the residuals.

2.4 Jacobi Algorithm

The Jacobi method or Jacobi iterative method is an algorithm for determining the solutions of a diagonally dominant system of linear equations [5]. Each diagonal element in this method is solved accurately, and an approximate value is plugged in until the iteration process achieving convergent. The Jacobi method was used to find the best iteration by using a linear equation over six combination parameters. The Jacobi algorithm is expressed in Equation 2.

$$B = a_{x_{11}} + a_{x_{12}} + a_{x_{13}} + \dots + a_{x_n} \quad (2)$$

where a_{xn} and B are not equal to zero for predictor and target parameters, respectively.

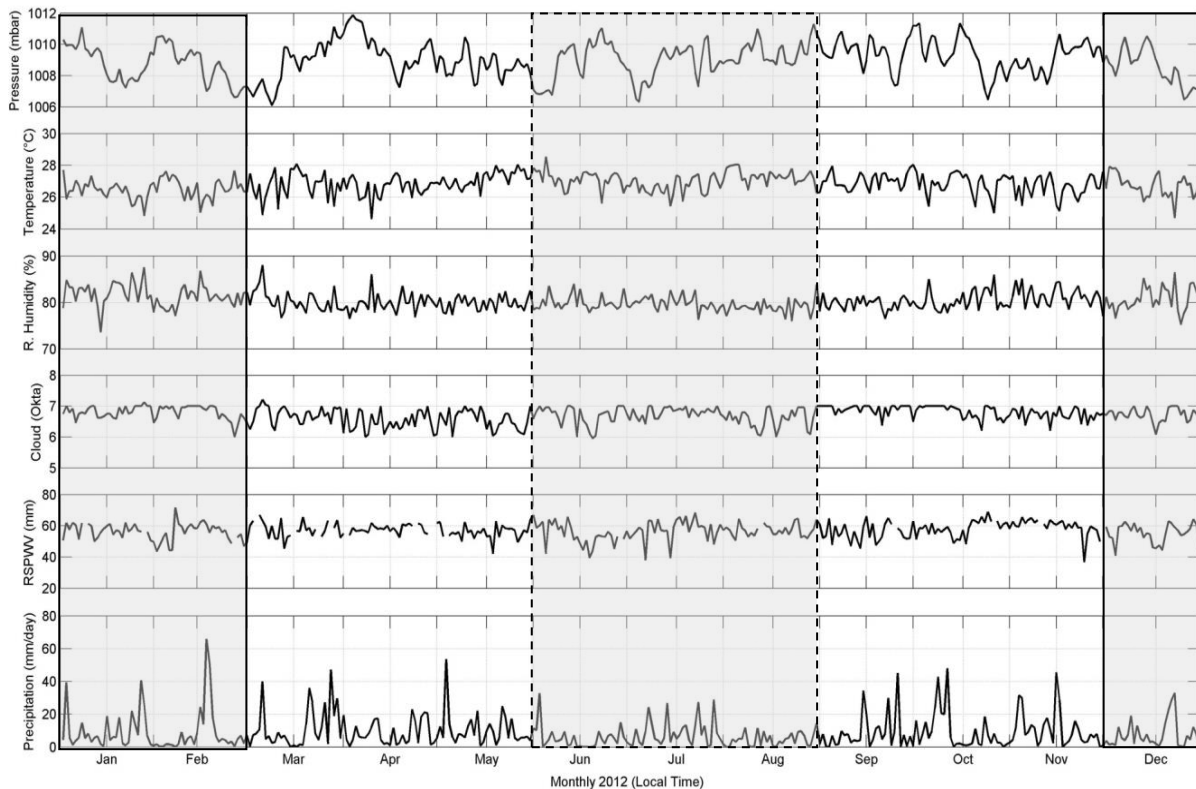


Figure 3. The variation of meteorology parameter over Tawau area. The shaded area shows the winter I, summer, and winter II, respectively.

3.2. Thunderstorm Estimation

A standard error regression (S) and R-squared (R-sq) values are examined to show the best relationship between input and output developed for estimation. Results show that the configuration input with relative humidity and RSPWV obtained a good result as compared to another configuration with S and R-sq are obtained below 50% and 6%, respectively. After the best configuration input was obtained, the input with H and RSPWV are proposed to develop a MLR equation by using the observation data in 2012. The selection configuration inputs were processed to find the maximum iteration. Table 1 shows the MLR equation for three seasons in 2012 that was processed using Jacobi method. The maximum iteration for $x_1 \sim x_2$ is presented in two - three steps with two configurations for each season. The Python software was used to obtain the estimation model. In this study, 1000 iteration (max) was used to find the optimum input model for estimation process.

Table 1. The estimation of H and RSPWV using Jacobi algorithm for the season in 2012

Season	The Jacobi Equation	Solution
Winter I (JF) month	$3.48 = 0.0462 x_1$ $3.33 = 0.0514 x_1 - 0.0104 x_2$	$x_1 = 75.324$ $x_2 = 52.08$
Summer (JJA) month	$-2.64 = -0.0294 x_1$ $-2.58 = -0.0294 x_1 + 0.00092 x_2$	$x_1 = 89.795$ $x_2 = 61.19$
Winter II (Dec) month	$3.49 = 0.0470 x_1$ $3.90 = 0.0465 x_1 + 0.0080 x_2$	$x_1 = 74.255$ $x_2 = 55.89$
Transition I (MAM) month	$1.60 = 0.0248 x_1$ $1.61 = 0.0264 x_1 - 0.00217 x_2$	$x_1 = 64.516$ $x_2 = 42.96$
Transition II (SON) month	$2.12 = 0.0325 x_1$ $2.06 = 0.0371 x_1 - 0.00736 x_2$	$x_1 = 65.230$ $x_2 = 48.92$

Note that x_1 and x_2 stand for H and RSPWV, respectively

The estimation of parameter H and RSPWV from two configuration equations for each season was obtained using Jacobi algorithm. However the negative result in x_2 (RSPWV) indicated that the value is out of range. Furthermore, the estimation result using MLR equation is presented in Figure 4. As can be seen on the figure, five equations were used in the Jacobi methods which were winter I, transition I, summer, transition II, and winter II. Based on the investigation, the minimum and maximum RSPWV and precipitation are 38 ~ 82 mm and 0 ~ 76 mm/day, respectively. The estimation result of thunderstorm frequency for each season was obtained by using rounding number technique. For example, if the output reached < 0.5 would mean that there is no event (value = 0) while output at > 0.5 indicate that an event may occurred (value = 1).

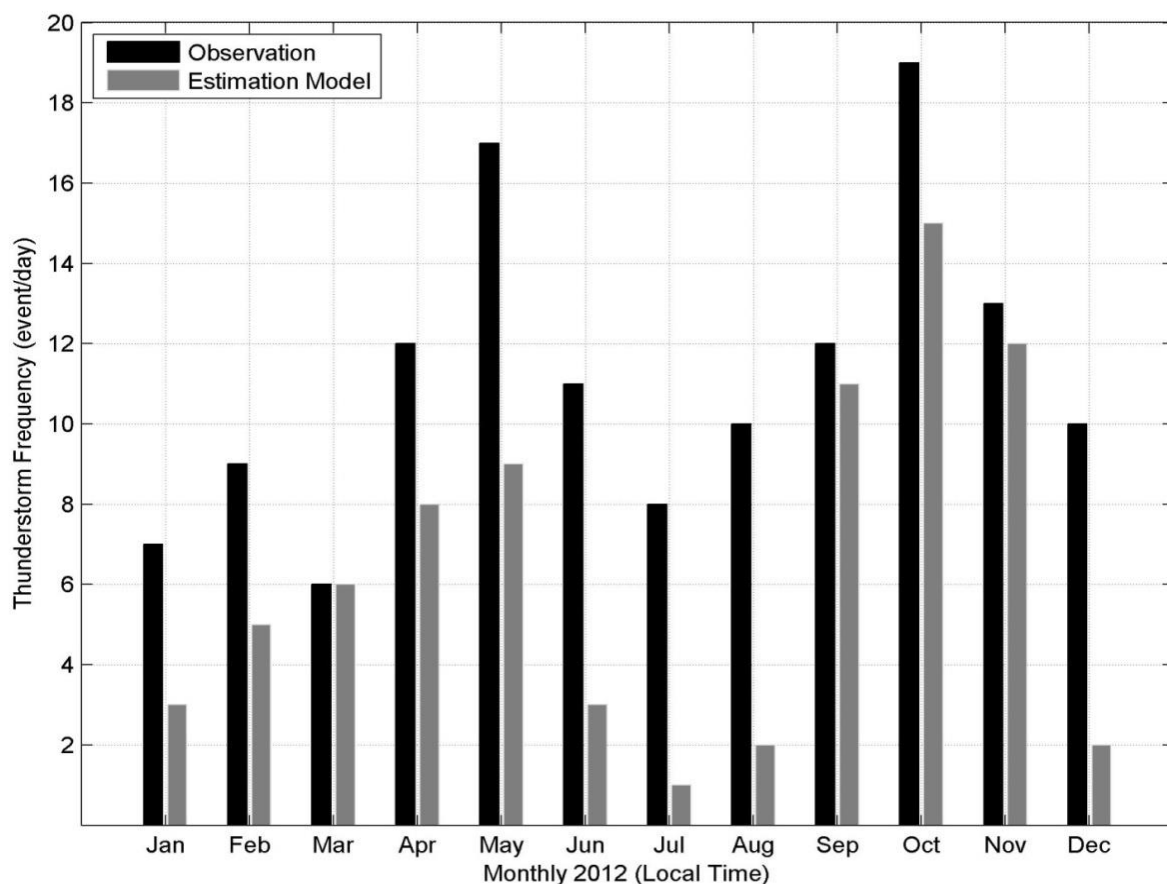


Figure 4. The estimation model of thunderstorm activity over TWU station

In order to evaluate the performance of the MLR model in estimating thunderstorm activity, the MLR model was compared with *Dvorak* technique and Adaptive Neural Fuzzy Inference System (ANFIS) with *Fuzzy Clustering Method* (FCM). Figure 5 shows the comparison between the three models. The model used 1,000 epoch and five layers (fixed layer ANFIS) and found with maximum error $< 1\%$. From the figure, it can be seen that the three models comparison provides a strong relationship (> 0.75 at the 99% confidence level), where ANFIS FCM reached the highest correlation followed by the *Dvorak* technique and the Numerical model. In addition, all three methods reached minimum error (see Table 2) as indicated by root mean square error (RMSE), mean absolute error (MAE), and percent error (PE). ANFIS FCM provides good results followed by the *Dvorak* technique and the MLR model. From the comparison analysis, the PE for MLR is worse than the other models which possibly due MLR cannot fully capture the linearization for 1,000 iterations.

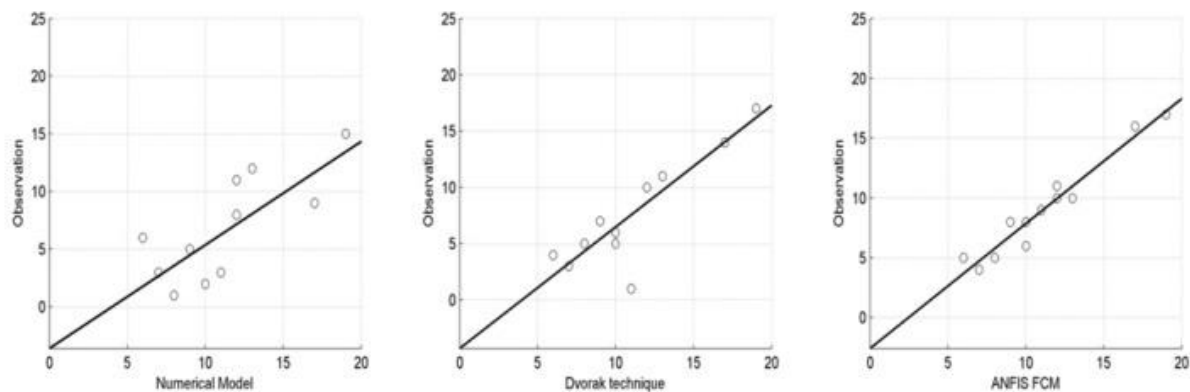


Figure 5. Comparison of observation data between (a) Numerical Modeling, (b) *Dvorak* technique, and (c) ANFIS FCM model

Table 2. The comparison analysis of thunderstorm activity in the TWU station in 2012

Method	Correlation (R^2) value	RMSE	MAE (%)	PE (%)
MLR Model	0.8911	4.7697	4.4167	39.7288
<i>Dvorak</i> technique	0.9352	3.8944	3.6667	36.4322
ANFIS FCM	0.9812	1.8930	1.7500	17.2097

4. Conclusion and Future Work

A MLR model using Jacobi method was successfully carried out to estimate thunderstorm activity over Tawau Area for the year of 2012. The model was developed based on the two configuration inputs (H and RSPWV) and one output of thunderstorm. The thunderstorm activity in Tawau was found highest in the inter-monsoon season in March ~ May (winter to summer) and September ~ November (summer to winter), respectively. More than 12 events per month were obtained in this season indicating that the region is very convective. The comparison between MLR model, *Dvorak* technique, and ANFIS FCM found that the MLR model have percent error almost 40%, which is not suitable for estimation of complex thunderstorm. Finally, the ANFIS model is more advantage in constructing the estimation of thunderstorm activity, and therefore, it is suggested as a prediction model in the next studies.

Acknowledgements

This study is funded by the Ministry of Science, Technology and Innovation (MOSTI) under Science Fund 01-01-02-SF1100 grant and UKM Publications Rewards: GP-K017155. We would like to thank the Malaysian Meteorological Department (MetMalaysia) for providing the meteorological data, NASA website for providing the precipitation data, and the Department of Atmospheric Science, University of Wyoming for providing the RSPWV data.

References

- [1] Houze R A Jr 2004 Mesoscale Convective Systems *Reviews of Geophysics* **42** 8755-1209
- [2] Suparta W, Putro W S, Singh M S J, Asillam M F 2015 Characterization of GPS and Meteorological Parameters for Mesoscale Convective Systems Model over Tawau, Malaysia *Advanced Science Letters* **21**(2) 203-206
- [3] Albar A, AL-Khalaf A K and Abdel-Basset H 2015 Radar Rainfall Estimation of a Severe Thunderstorm over Jeddah *Atmospheric and Climate Sciences* **5** 302-316
- [4] Balijepalli N, Venkata S, Richter Jr C, Christie R and Longo V 2005 Distribution system reliability assessment due to lightning storms *IEEE Trans. Power Delivery* **20** 2153-2159
- [5] Yang X I A and Mittal R, 2014 Acceleration of the Jacobi iterative method by factors exceeding 100 using scheduled relaxation *Journal of Computational Physics* **274** 695-708