PRECIPITATION PREDICTION IN THAILAND BY USING REGRESSION ANALYSIS

YURANAN JAIYEAKYEN

A THEMATIC SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2013

Thematic Paper
entitled

PRECIPITATION PREDICTION IN THAILAND BY USING REGRESSION ANALYSIS

……..
Mr. Yuranan Jaiyeakyen
Candidate

……..……
Lect. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Major advisor

……..……
Asst. Prof. Bunlur Emaruchi
Ph.D. (Environmental Systems Engineering)
Co- advisor

……..……
Prof. Banchong Mahaisavariya,
M.D., Dip. (Thai Board of Orthopedics)
Dean
Faculty of Graduate Studies,
Mahidol University

……..……
Asst. Prof Supaporn Kiattisin,
Ph.D.
(Electrical and Computer Engineering)
Program Director
Master of Science Program in Technology of Information System Management
Faculty of Engineering,
Mahidol University

Thematic Paper
entitled

PRECIPITATION PREDICTION IN THAILAND BY USING REGRESSION ANALYSIS

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Technology of Information System Management)
on
December 26, 2013

……
Mr. Yuranan Jaiyeakyen
Candidate

……
Asst. Prof Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Chair

……
Lect. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Member

……
Asst. Prof. Bunlur Emaruchi,
Ph.D. (Environmental Systems Engineering)
Member

……
Asst. Prof Werapon Chiracharit,
Ph.D. (Electrical and Computer Engineering)
Member

……
Prof. Banchong Mahaisavariya,
M.D., Dip (Thai Board of Orthopedics)
Dean
Faculty of Graduate Studies
Mahidol University

……
Lect. Worawit Israngkul,
M.S. (Technical Management)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

The success of this Thematic Paper is attributed by attentive support from my major advisor, Lect. Waranyu Wongseree, and my co-advisor, Asst. Prof. Lect. Bunlur Emaruchi and Asst. Prof. Supaporn Kiattisin, I would like to express my sincere gratitude for all their invaluable encouragement, guidance, supervision and suggestion throughout this research.

A special thank goes to my co-advisor, Asst. Prof. Supaporn Kiattisin, and Asst. Prof. Bunlur Emaruchi, committee chair for their comments and useful suggestions to this research.

Special acknowledgement goes to all the lecturers and staff of the Technology of Information System Management Program, Faculty of Engineering, Mahidol University for their service and support. I would like to thank all of my friends for their helps and encouragements.

Finally, the most important and indispensable persons are my family members. I am very grateful for their entire support, caring and love throughout my whole life. This has inspired me to fight and beat the obstacles to finish this research as well.

Yuranan Jaiyeakyen

PRECIPITATION PREDICTION IN THAILAND BY USING REGRESSION ANALYSIS

YURANAN JAIYEAKYEN 5437855 EGTI/M

M.Sc. (TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)

THEMATIC PAPER ADVISORY COMMITTEE : WARANYU WONGSEREE, Ph.D., BUNLUR EMARUCHI, Ph.D., SUPAPORN KIATTISIN, Ph.D.

ABSTRACT

A methodology is presented for statistical downscaling the general circulation model (GCM) of GFDL-R30 and HadGEM output to predict precipitation at meteorology stations in Thailand under expected future climatic conditions. The meteorology stations that were sampled for testing for prediction were Chiang Mai, Udon Thani, Bangkok, Ubon Ratchathama, Nakhon Ratchasima and Phuket, provinces from each region in Thailand. The approach involves a combination of data preprocessing, data classification and regression analysis. In the preprocessing step, this research used data with GFDL-R30 which was compared between logarithm transformations, square root transformations and no transformations. The results showed the correlation yield of both algorithms was less than no transformations but they can help for negative numbers in predicted data. In the classification step, it applied to the model for cutoff on days that had less rainfall and heavy rainfall. It was varied from 0 mm to 1.5 mm. The results showed the classification cannot increase strong correlation performance. The last section presents about post-processing comparison between multiple linear regressions and Gaussian process which is a non-linear regression. The result showed Gaussian process generally yielded better results both for correlation and RMSE for every station that was tested. Therefore the Gaussian process is a suitable approach for predicting precipitation in the future without any preprocessing.

KEY WORDS: REGRESSION ANALYSIS / NUMERICAL WEATHER PREDICTION / PRECIPITAION PREDICTION / DATA TRANSFORMATION / CLASSIFICATION / GAUSSIAN PROCESS

39 pages

การทำนายหยาดน้ำฟ้าในประเทศไทยด้วยวิธีการวิเคราะห์สมการถดถอย

PRECIPITATION PREDICTION IN THAILAND BY USING REGRESSION ANALYSIS

ยุรนันท์ ใจเยือกเย็น 5437855 EGTI/M

วท.ม. (เทคโนโลยีการจัดการระบบสารสนเทศ)

คณะกรรมการที่ปรึกษาสารนิพนธ์: วรัญญู วงษ์เสรี, Ph.D., บันลือ เอมะรุจิ, Ph.D., สุภาภรณ์ เกียรติสิน, Ph.D.

บทคัดย่อ

งานวิจัยฉบับนี้ได้นำเสนอกระบวนการลดขนาดทางสถิติของ global climate models (GCMs) ของ GFDL-R30 และ HadGEM ไปสู่สถานีตรวจอากาศของประเทศไทยเพื่อทำการพยากรณ์ปริมาณน้ำฝนภายใต้สภาพอากาศที่คาดการณ์ไม่ได้ในอนาคต สถานีตรวจอากาศที่ถูกนำมาใช้ทดสอบการพยากรณ์คือ เชียงใหม่ อุดรธานี กรุงเทพ อุบลราชธานี นครราชสีมา และภูเก็ต ซึ่งเป็นจังหวัดจากภูมิภาคต่างๆของประเทศไทย วิธีการที่ใช้ในทดสอบรวมไปถึงขั้นตอนการเตรียมข้อมูล การแบ่งกลุ่มข้อมูล และการวิเคราะห์ด้วยสมการถดถอย ในขั้นตอนการเตรียมข้อมูลงานวิจัยได้ใช้ข้อมูล GCMs ของ GFDL-R30 เพื่อเปรียบเทียบระหว่างการใช้ล็อการิทึมและสแควร์รูทเพื่อเปรียบเทียบกับการไม่แปลงข้อมูล ผลการทดลองพบว่าค่าสหสัมพันธ์ของทั้งสองวิธีการน้อยกว่าการไม่ทำการแปลงข้อมูล แต่การแปลงข้อมูลสามารถช่วยให้ผลการทำนายไม่เป็นตัวเลขติดลบได้ ในขั้นตอนการแบ่งแยกข้อมูล เป็นวิธีการที่ถูกนำมาประยุกต์ใช้เพื่อแบ่งแยกข้อมูลระหว่างวันที่ฝนตกน้อยกับวันที่ฝนตกหนัก โดยทำการปรับขั้นตั้งแต่ปริมาตร 0 -1.5 มิลลิเมตร ผลการทดสอบพบว่าการแบ่งแยกข้อมูลไม่สามารถเพิ่มค่าสหสัมพันธ์ได้เช่นกัน ในส่วนสุดท้ายเป็นการนำเสนอในส่วนของการทำประมวลผลข้อมูลเปรียบเทียบกันระหว่างวิเคราะห์การถดถอยเชิงเส้นและกระบวนการแบบเกาส์ซึ่งเป็นการวิเคราะห์การถดถอยไม่เชิงเส้น ผลการทดสอบพบว่ากระบวนการแบบเกาส์ให้ผลการทดสอบที่ดีกว่าทั้งค่าสหสัมพันธ์และรากที่สองของผลรวมของกำลังสองของค่าความคลาดเคลื่อนในทุกๆ สถานีที่ทำการทดสอบ ดังนั้นกระบวนการแบบเกาส์จึงเป็นวิธีการที่เหมาะสมที่สุดที่จะใช้ในการทำนายปริมาณน้ำฝนในอนาคตโดยไม่ต้องมีการทำการแปลงข้อมูลแต่อย่างใด

39 หน้า

CONTENTS

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (cont.)

CHAPTER I
INTRODUCTION

1.1 Background and statement of problems

Climate changes due to the increasing of greenhouse gases in the atmosphere produced by human activities effects to the environment. Recent impacts such as seasons are shifting, more frequent hot days and nights, temperatures are climbing and sea levels are rising. These changes will impact our food supply, water resources, infrastructure, ecosystems, and even our own health.

Much of operational weather and long-range (climate) forecasting has a statistical basis. GCMs and global climate models are widely applied for weather forecasting. However GCMs are usually run at coarse grid resolution (in the order of 300s of kilometers) and as a result they are inherently unable to present local sub-grid scale features and dynamics.

This research explored statistical downscaling the outputs of GFDL-R30 in Thailand area by using multiple linear regression and gaussian process approach validated using cross validation to find correlation between variable from GCMs and observed precipitation data to predict precipitation in target station.

1.2 Objectives of study

1. To compare correlation performance of pre-processing approaches for data transformations method between no transformations, logarithm transformations and square root transformations.

2. To compare correlation performance between models which apply classification and without apply classification.

3. To compare correlation performance of regression analysis between multiple linear regression and gaussian process.

1.3 Scope of work

1. To study regression analysis by using linear regression and gaussian process

2. Examine testing results by using correlation coefficient (R) and root mean square error (RMSE).

3. Study only observation data from meteorology station in Thailand by using GFDL-R30 and HadGEM as the variable and observation data from Meteorological station to prediction.

1.4 Result

1. Data Transformation method can't increase strength of model's correlation the both of Logarithm transformation and Square root transformation. No transformation are best of performance.

2. Classification method can't increase strength of model's correlation as well.

3. Gaussian process, non-linear regression model yielded correlate stronger than multiple linear regressions.

CHAPTER II
LITERATURE REVIEW

2.1 Numerical weather prediction (NWP)

Numerical weather prediction is weather forecasting by using complex computer programs on power of computers depend on current weather conditions. NWP widely applied to climate change in currently. The atmospheric variable for weather prediction is meteorological information such as temperature, pressure, rainfall and wind covering land, ocean areas, and the upper atmosphere. The quality of the analysis depends on mathematical equations/algorithms and observations that used as input to the forecasts. Numerical weather prediction is at the core of the complex set of forecast services, and more accurate forecasts depend on progress in numerical weather prediction more than on anything else [1]. Result of NWP model can generate either short-term weather forecasts or longer-term in the future with grid resolution 15-100km.

2.2 Model Output Statistics (MOS)

Model Output Statistics is derived from weather forecasting model output as the variables statistical approach by using post-processing method. Not only numerical weather prediction, MOS can obtained prior surface weather observations or geoclimatic information as predictors. The predictand for MOS is historical record of weathering observation data from the targets. The population screening regression procedure that widely use is multiple linear regression [2]. The MOS accuracy generally better than pure statistical model or a pure numerical model (NWP).

2.3 Global climate models (GCMs)

A Global climate models (GCMs) is a mathematical model of the general circulation of atmosphere, ocean, cryosphere and land surface. GCMs and global climate models are widely applied for weather forecasting, understanding the climate, and projecting climate change. The typically GCMs grid resolution of between 250 and 600 km. The values of the predicted variables, such as surface pressure, wind, temperature, humidity and rainfall are calculated at each grid point over time, to predict their future values The originally created by Syukuro Manabe and Kirk Bryan at the Geophysical Fluid Dynamics Laboratory in Princeton, New Jersey. The GCMs calculation produces a large amount of truncation error, which grows with time. So there are problems with the case of long-range forecasts and climate simulation [3].

GCMs have model structure both of atmospheric GCMs (AGCMs) and oceanic GCMs (OGCMs). The coupled of AGCM and an OGCM can calling as atmosphere-ocean coupled general circulation model (CGCM or AOGCM). This is kind of GCMs on each application.

- A simple general circulation model (SGCM),
- Atmospheric GCMs (AGCMs)

Model the atmosphere (and typically contain a land-surface model as well) and impose sea surface temperatures (SSTs). A large amount of information including model documentation is available from AMIP.

- Oceanic GCMs (OGCMs)

Model the ocean that may or may not contain a sea ice model. For example, the standard resolution of HadOM3 is 1.25 degrees in latitude and longitude, with 20 vertical levels, leading to approximately 1,500,000 variables.

- Coupled atmosphere–ocean GCMs (AOGCMs)

They thus have the advantage of removing the need to specify fluxes across the interface of the ocean surface. These models are the basis for sophisticated model predictions of future climate, such as are discussed by the IPCC.

Table 2.1: Example of GCMs from each institution.

GCMs	Institution	Country
GFDL	Geophysical Fluid Dynamics Laboratory, National Oceanic Atmospheric Administration (NOAA)	The United States.
GISS	Goddard Institute for Space Studies, National Aeronautics and Space Administration (NASA)	The United States.
HadCM and UKMO	Hadley Center Meteorological Office	United Kingdom
EKHAM	Max Planck Institute	Germany
CCCma and CGCM	Canadian Center for Climate Modeling and Analysis	Canada
CSIRO Mk II	Commonwealth Scientific and Industrial Research Organization	Australia
IPSL	Institute Pierre Simon Laplace	France
MRI	Meteorological Research Institute	Japan

2.4 Climate Downscaling

Downscaling is a methods to obtaining local scale of global climate models (GCMs). In generally, Output from GCMs have a grid resolution around 150-300km. This gird size can't reflect impact of local area so they need downscaling to small grid size around 50 km or less. There're two main forms of downscaling technique exist. One form is dynamical downscaling, where output from the GCM is used to drive a regional, numerical model in higher spatial resolution. And another one is statistical downscaling, where a statistical relationship is established from observations between large scale variables such as surface pressure, wind, temperature, humidity and rainfall.

Figure 2.1: Grid size 300x300km. on Thailand Map

2.4.1 Dynamical downscaling

Dynamical downscaling is the method to fits output from GCMs into regional area or regional climate model (RCM). The typically grid resolution around 10-50km. Propose of dynamical downscaling is reflect how global patterns affect local weather conditions. This method need computer that high performance.

2.4.2 Statistical Downscaling

Statistical Downscaling have same propose as dynamical downscaling. This methods established from observations, like surface pressure, wind, temperature, humidity and rainfall as the variable and using output from GCMs as predictands.

Because the statistical approach requires less computational effort than dynamic downscaling, it offers the opportunity for testing scenarios for many decades or even centuries, rather than the brief “time slices” of the dynamical downscaling approach.

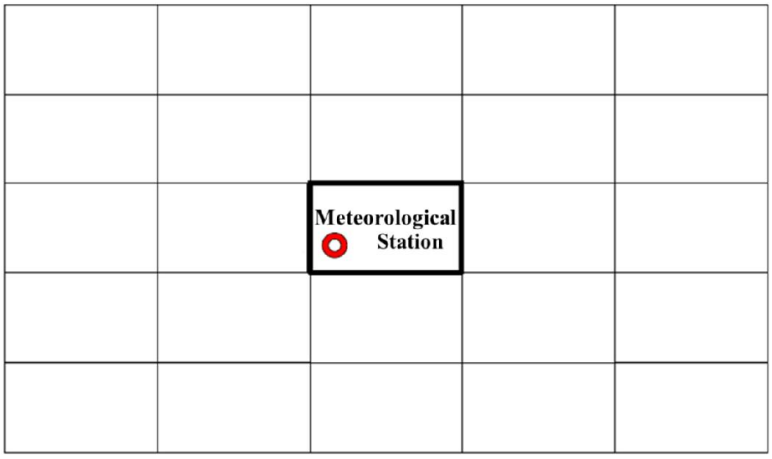

Figure 2.2: Potential grid around target station

2.5 Related research

Sailor et al presented for downscaling NCAR Model output to predict surface wind speeds in the wind power industry at Texas and California by using multi-layer perceptron (MLP) method. This research applied the methodology to three test sites. All sites demonstrated significant differences in downscaled wind speeds relative to the raw output from the GCMs (Yielded the strongest model with a validation data set correlation coefficient of 0.90) [5].

Huth compared 5 statistical downscaling methods are canonical correlation analysis (CCA), singular value decomposition (SVD) and 3 multiple regression models: stepwise screening of the predictor's PCs (stepwise regression), regression of the PCs without screening (full regression) and stepwise screening of the predictor's gridpoint values (pointwise regression) The performance of the methods was evaluated using cross-validation and quantified in terms of root-mean-squared error (RMSE) and found the pointwise regression proved to be the best method of those considered [6].

Dibike and Coulibaly presents an application of temporal neural networks for downscaling NCEP and CGCM1 model output to predict daily precipitation and temperature in the Saguenay watershed in Canada by using multiple linear regression and ANN. This search show time lagged feed-forward network (TLFN) can be an effective method for downscaling daily precipitation and temperature data as compared to the commonly used method (MR) [7].

Tripathi et al compares two methods of ANNs are Support Vector Machine (SVM) and Multilayer Perceptron to predict precipitation on monthly time scale in India with Coupled Global Climate Model (CGCM2). The result show SVMs provide a promising alternative to conventional artificial neural networks for statistical downscaling and are suitable for conducting climate impact studies [8].

E. Eccel prediction for minimum temperature in fruit-growing region of the Italian Alps, based on the output of two different NWPs (ECMWF T511–L60 and LAMI-3) They implemented a multivariate linear regression on the output at the grid points surrounding the target area, and two non-linear models based on machine learning techniques: Neural Networks and Random Forest. The performance of the post-processing models described above has been assessed in three different aspects are accuracy, mean absolute error (MAE) and Correlation. The best performing method is Random Forest [9].

From the related research, the widely post-processing method that was used in weather prediction is linear regression. After A.D.2000, non-linear such as neural network was adopt to this application. The performance of the methods was evaluated using cross-validation.

So this research will apply both of linear regression and non-linear regression. And we will propose another one non-linear regression is Gaussian process.

CHAPTER III
RESEARCH METHODOLOGY

This chapter presented processes for model developing and evaluation. There're major step as follows:

1. Data Selection
2. Data Pre-Processing
3. Development
4. Evaluation

3.1 Data Selection

We use Geophysical Fluid Dynamics Laboratory (GFDL) from The United States to trains our model. Scientists at GFDL develop and use mathematical models and computer simulations to improve our understanding and prediction of the behavior of the atmosphere, the oceans, and climate. GFDL scientists focus on model-building relevant for society, such as hurricane research, prediction, and seasonal forecasting, and understanding global and regional climate change.

Table 3.1: Detail of GFDL-R30 model used to predict climate change.

Scope	Detail
GCMs model	GFDL-R30
Base years	A.D.1965-1990
Forecast years	A.D.2010-2029 and A.D.2040-2059
Size	Latitude 2.20° x Longitude 3.75°
Time resolution	Monthly
Predictor variable	- Temperature - Atmospheric Pressure - Precipitation - Solar Radiation - Evaporation

Another one GCMs model that was used is this research is HadGEM3. HadGEM3 stands for the Hadley Centre Global Environment Model version 3. The HadGEM model results are drawn from the IPCC AR4 20C3M scenario, which incorporates the observed increase in greenhouse gas concentrations.

Table 3.2: Detail of HadGEM model used to predict climate change

Scope	Detail
GCMs model	HadGEM
Base years	Dec 1, 1949 to Nov 30, 2005 (56 years)
Time resolution	Daily
Predictor variable	- Daily-Mean Near-Surface Wind Speed - Sea Level Pressure - Precipitation - Near-Surface Specific Humidity - Near-Surface Air Temperature - Daily Maximum Near-Surface Air Temperature - Daily Minimum Near-Surface Air Temperature

Climate data used in this study measure from 119 Thai meteorological stations covering the area to look at average temperature, max min-temperature, relative humidity, sunshine duration ,atmospheric pressure, wind speed and amount of precipitation.

Table 3.3: Detail of observation data from Metrological Station

Scope	Detail
Data Range	Jan 1, 1951 to Dec 31, 2011 (60 years)
Metrological Station	6 Stations in Thailand
Time resolution	Daily
Observation Data	Precipitation

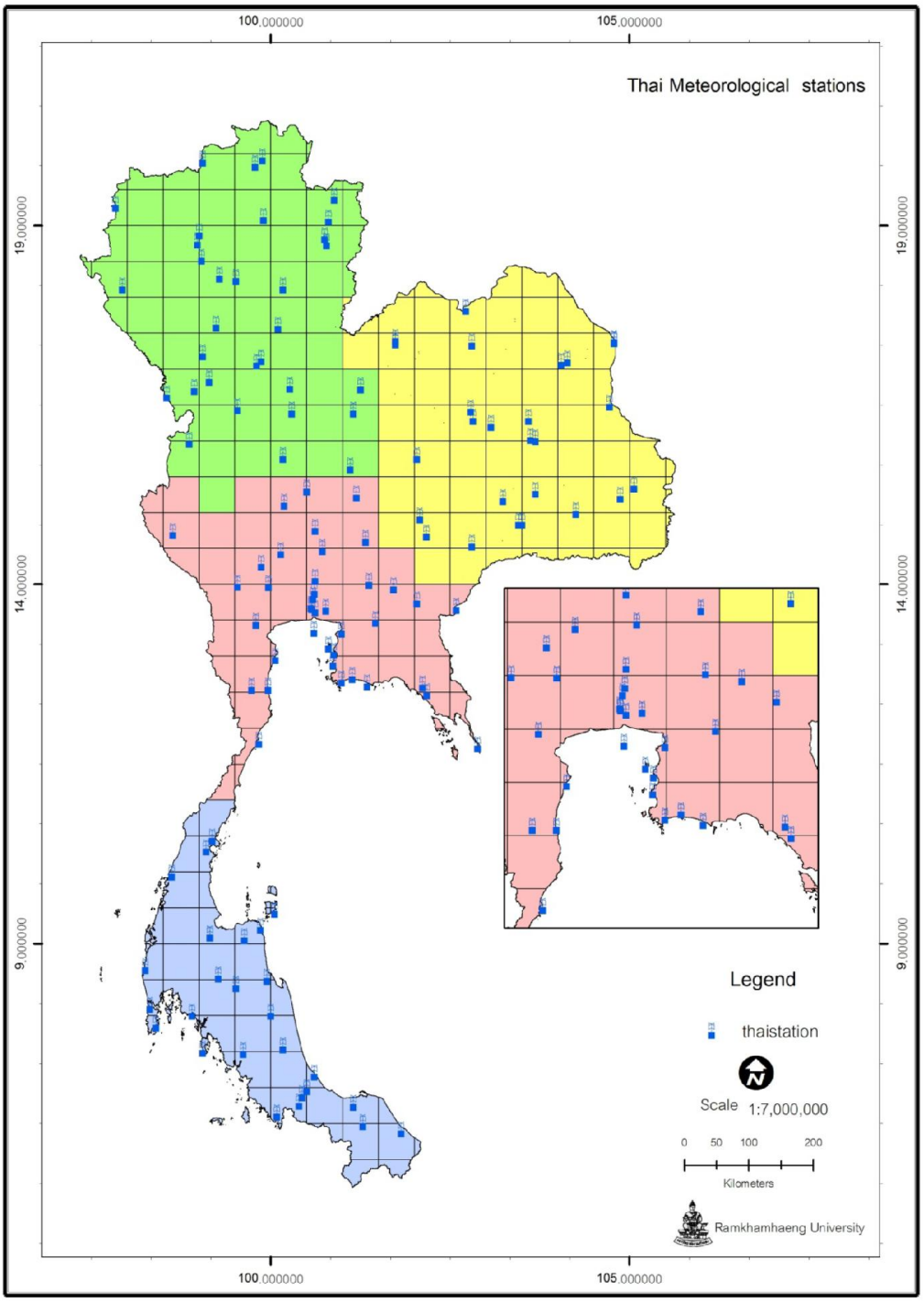

Figure 3.1: [10] All Thai meteorological stations

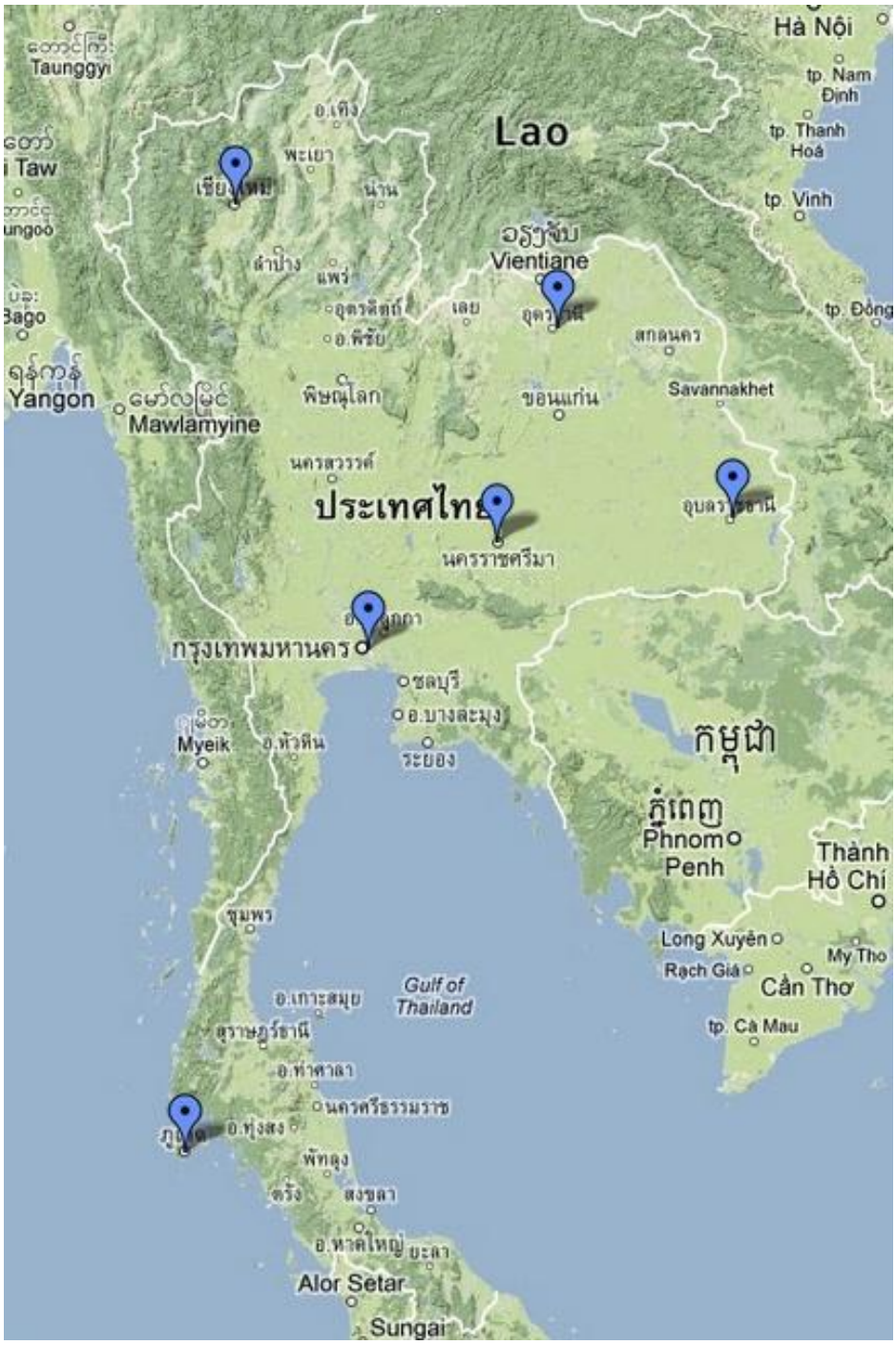

Figure 3.2: Thai meteorological station for training

In this section we imported both of GCMs data and observation data into Microsoft SQL server express 2012 through Microsoft SQL server management studio 2012 and query target data of each station for experimental.

Microsoft SQL Server is a relational database management system from Microsoft. The original SQL Server code was developed by Sybase; in the late 1980s. MS SQL Server working on T-SQL (Transact –SQL) such us store and retrieve data. T-SQL is programming command developed by Sybase. Microsoft add more feature for T-SQL for more ability. MS SQL Server can working on local, intranet and another computer across internet.

Figure 3.3: Microsoft SQL Server express 2012, tooling for data preparation

SQL Server Management Studio (ssms) is a software from Microsoft for configuring, managing, and administering Microsoft SQL Server express 2012. The mainly propose of SQL Server Management Studio is graphical tools as monitoring portal for retrieve data and store new data to database. The freely version of SQL Server Management Studio is “express” version that can download couple with Microsoft SQL server express 2012.

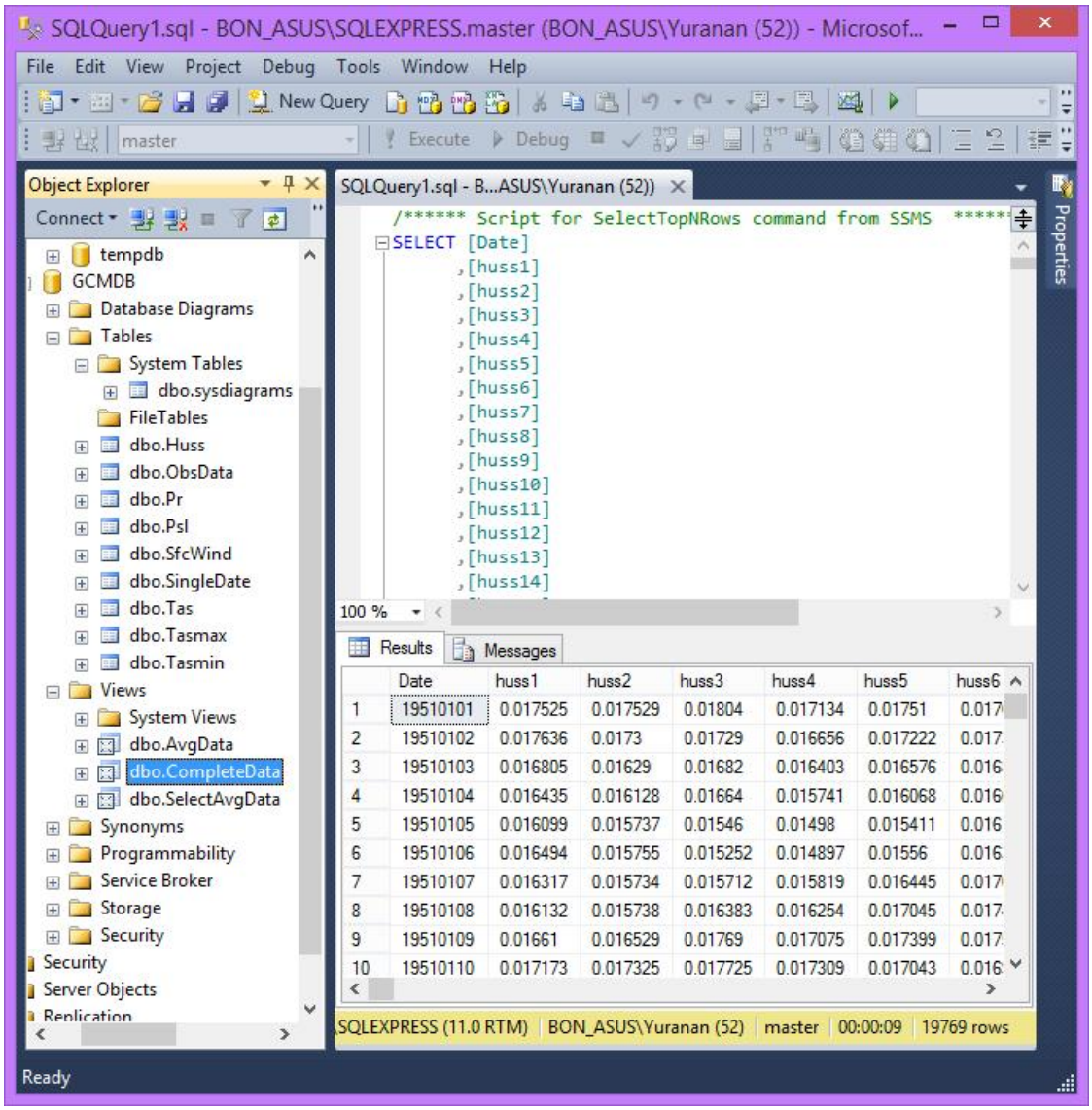

Figure 3.4: SQL Server Management Studio 2012 for data query

3.2 Data Pre-Processing: Data transformation

There're some variables are not normally distributed and therefore do not meet the assumptions of parametric statistical tests. Transforming the data will make it fit the assumptions better. In this research, we use 3 type of data transformation are log transformation and square root transformation and to find optimum model.

Precipitation data have zeros numbers on day that no rainfall. This data can't take the log so we must add a constant to each number to make them positive and non-zero. However we have to subtract the constant from the results also.

Table 3.4: Formula which using for transformation

Transformation technique	**Pre-Formula**	**Post-Formula**
Log Transformation	Y = Log(Y+1)	Y = Exp(Y)-1
Square root Transformation	Y = Square(Y+1)	Y = Power(Y,2)-1

We use Matlab R2012b for data pre-processing. Matlab is a numerical software that developed by MathWorks. Matlab allow user create programming algorithms as the matrix such as manipulation and plotting. So it suitable to convert batch data like this GCMs.

.

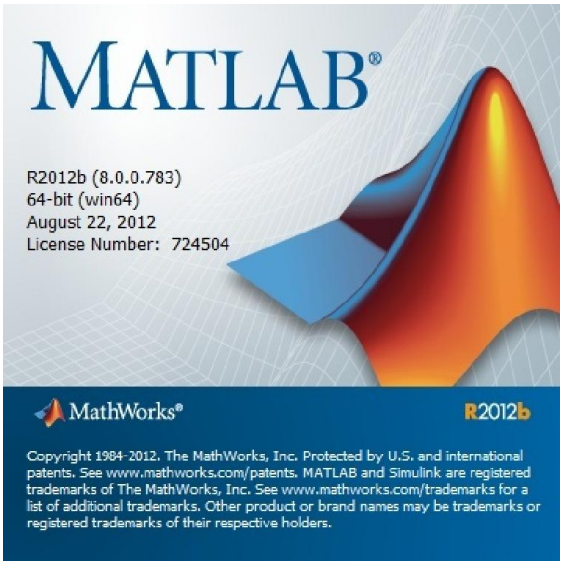

Figure 3.5: Matlab R2012b Software for Pre-Processing

3.3 Data Pre-Processing: Data Classification

Data Classification is a management tool for categorization type of data into their groups by using statistics. Data classification has close type to data clustering. The different point of both methods is data classification using for predictive while data clustering using for descriptive.

The classification step start at classify dataset into groups for category training. Categorical dependent variable that depend on one or more predictor variables into group. Then creating a descriptive model by using data set each group. The outcome of creation is the model used to categorize new items in the created classification system. An algorithm used for classification called the “classifier”.

These are the measures of effectiveness of the classifier as below [11].

- Predictive accuracy: How well does it predict the categories for new observations?

- Speed: What is the computational cost of using the classifier?

- Robustness: How well do the models created perform if data quality is low?

- Scalability: Does the classifier function efficiently with large amounts of data?

- Interpretability: Are the results understandable to users?

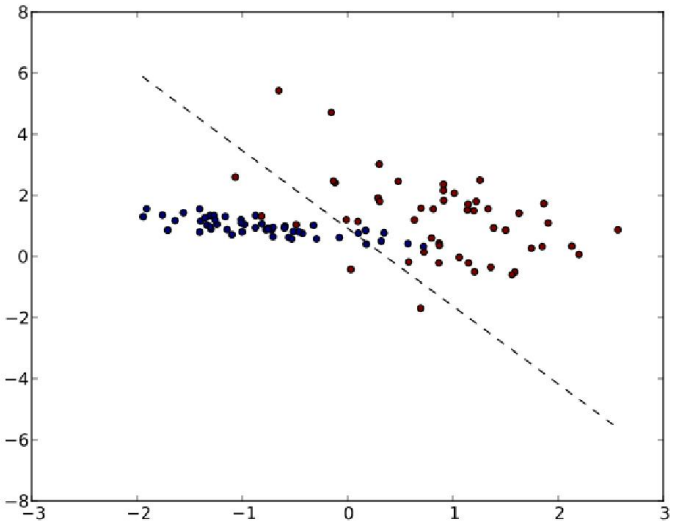

Figure 3.6: Example of data classification technique

In data classification phase, we use Matlab R2012b, numerical computing environment and fourth-generation programming language from MathWorks as well.

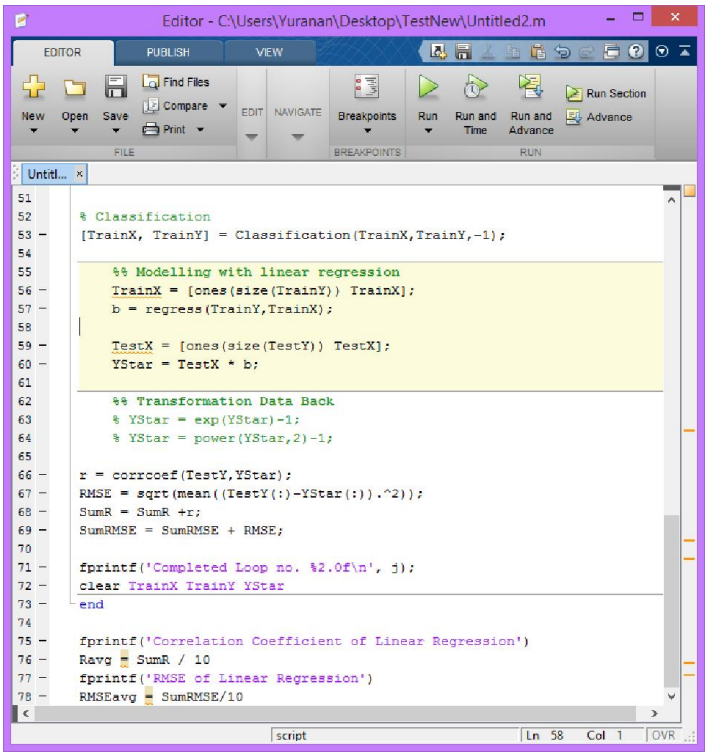

Figure 3.7: Matlab R2012b used for classification data

3.4 Development

Much of statistical weather forecasting is based on the statistical procedure known as linear, least-squares regression. There are three method was using in this research are linear regression and guassian process.

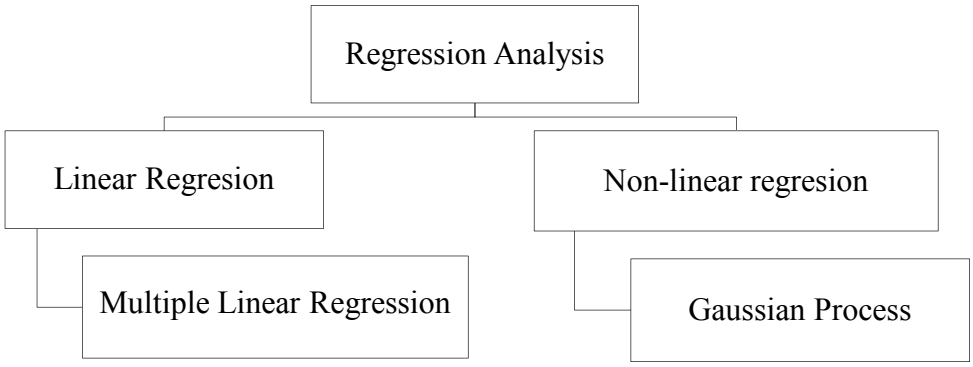

Figure 3.8: Thai meteorological station for training

3.4.1 Simple linear regression

Simple linear regression is a statistical technique to describe the relationship between two variables. The first variable is independent or explanatory or predictor. Symbol x used for represent this variable type. And the second variable is dependent or outcome or predictand by used symbol y. Relationship between two variable can shows in scatter plot and draw straight line to estimate trend of data. The scatter plot allows us to visually inspect the data prior to running a regression analysis. Its mathematical can be presents in equation as below

$$\hat{y} = a + bx$$

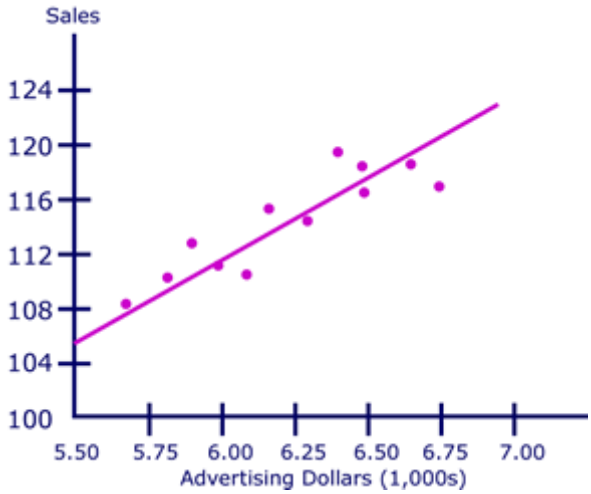

Figure 3.9: Simple linear regression [12]

There are gap between point and straight line, it is error of predictors of the dependent variables. Value on straight line is predictors from model. The circumflex (“hat”) accent signifies that the equation specifies a predicted value of y. And use “e” to indicates error of vertical distances between the data points and the line, also called errors or residuals, are defined as

$$e_i = y_i - \hat{y}\,(X_i)$$

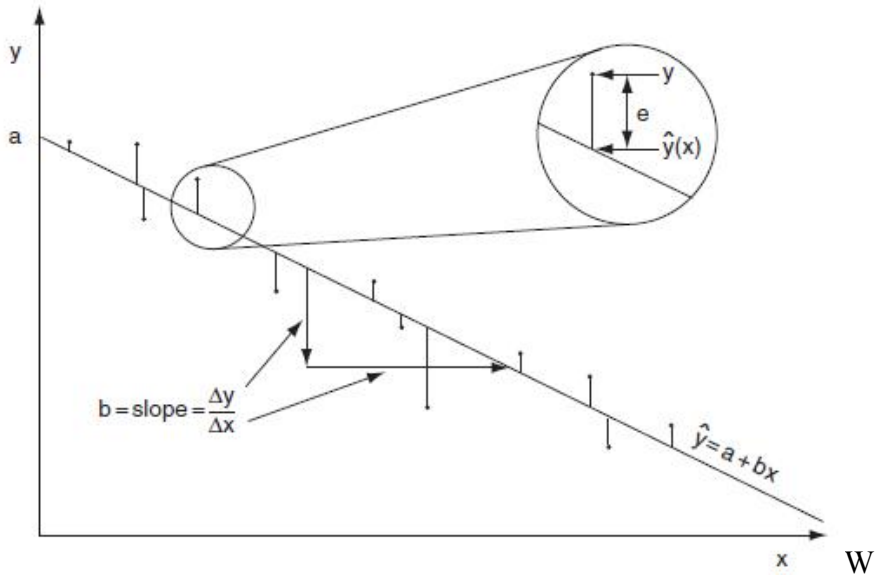

Figure 3.10: The residuals between the points and the line from linear regression [13]

3.4.2 Multiple linear regression

Multiple linear regression (MLR) is a statistical technique to describe the relationship between a dependent variable and independent variables like simple linear regression. But there are independent variables more than only one. X_1, X_2, X_3…..X_k used for represent number of independent variables (simple linear regression then reduces to the special case of K = 1). The prediction equation becomes

$$\hat{y} = b_0 + b_1x_1 + b_1x_2 + \ldots + b_kx_k$$

Each of the K predictor variables has its own coefficient, analogous to the slope, b, in equation. For notational convenience, the intercept (or regression constant) is denoted as b_0 rather than as a, as in equation. These K+1 regression coefficients often are called the regression parameters. In addition, the parameter b_0 is sometimes known as the regression constant.

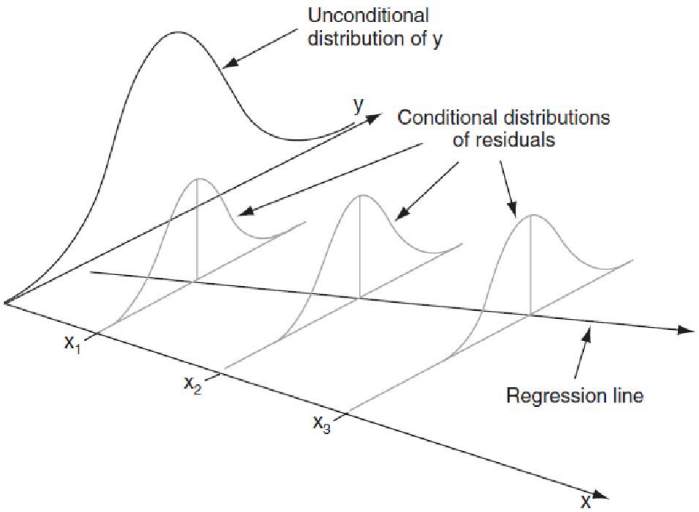

Figure 3.11: Multiple linear regression

3.4.3 Non-Linear Regression

- Gaussian Process

Gaussian process is a statistical modelling which realizations consist of random values associated with every point in a range of times. Moreover, every finite collection of those random variables has a multivariate normal distribution.

A gaussian process is a stochastic process Xt, t ∈ T, for which any finite linear combination of samples has a joint Gaussian distribution. More accurately, any linear functional applied to the sample function Xt will give a normally distributed result. Notation-wise, one can write X ~ GP(m,K), meaning the random function X is distributed as a GP with mean function m and covariance function K. When the input vector t is two- or multi-dimensional a Gaussian process might be also known as a Gaussian random field.

Some authors assume the random variables Xt have mean zero; this greatly simplifies calculations without loss of generality and allows the mean square properties of the process to be entirely determined by the covariance function K.

$$X \sim N(\mu, \sum)$$

There are a number of common covariance functions

- Constant :

$K_c(x, x') = C$

- Linear:

$K_L(x, x') = x^T x'$

- Gaussian Noise:

$K_{GN}(x, x') = \sigma^2 \delta_{x,x'}$

- Squared Exponential:

$K_{SE}(x, x') = \exp(-\frac{|d|^2}{2l^2})$

- Ornstein–Uhlenbeck:

$K_{OU}(x, x') = \exp(-\frac{|d|}{l})$

- Matérn:

$K_{Matern}(x, x') = \frac{2^{1-v}}{\tau(v)}(\frac{\sqrt{2v}|d|}{l})^v K_v(\frac{\sqrt{2v}|d|}{l})$

- Periodic:

$K_P(x, x') = \exp(-\frac{2sin^2(\frac{d}{2})}{l^2})$

- Rational Quadratic:

$K_{RQ}(x, x') = (1 + |d|^2)^{-\alpha}, \alpha \geq 0$

We use Weka (Waikato Environment for Knowledge Analysis), popular suite of machine learning software to support our experiment. Weka written in Java, developed at the University of Waikato, New Zealand

Figure 3.12: Weka Application

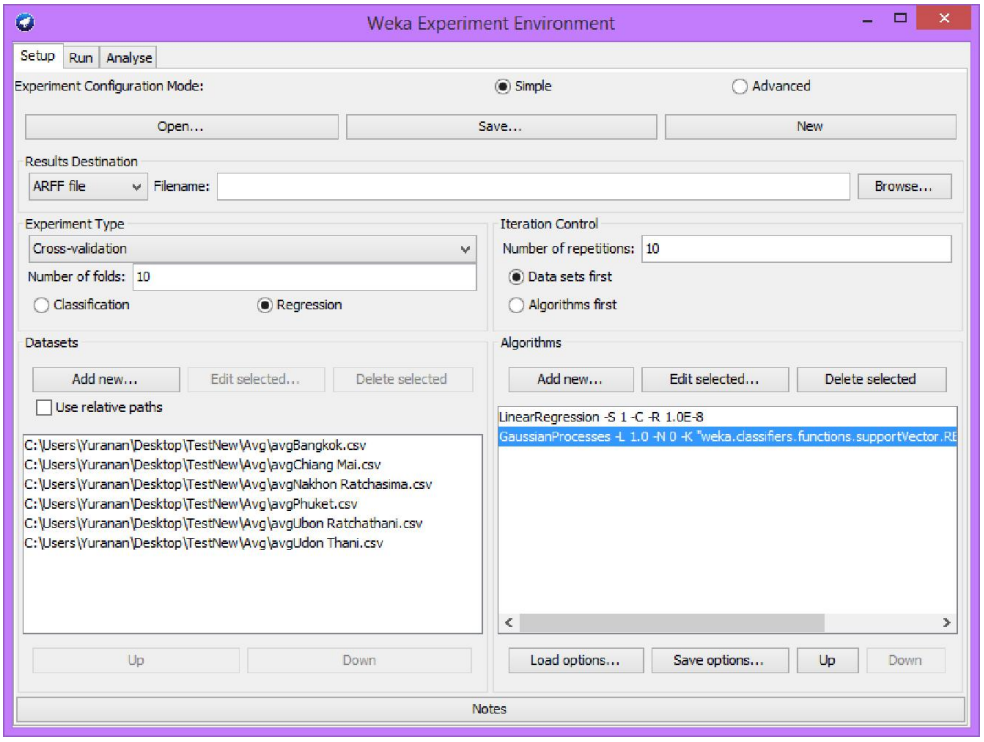

Figure 3.13: Weka Experiment Environment, the application enables the user to create, run, modify,and analyse experiments

3.5 Evaluation

3.5.1 Performance measurement

- Correlation Coefficient

We apply Correlation Coefficient (R) to measure strength and direction of the linear relationship between variables. The value of Pearson's correlation coefficient, unlike all other estimates for continuous dependent variables, is influenced by the distribution of the independent variable in the sample.

$$\text{Correlation coefficient} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

- Root Mean Square Error (RMSE)

Root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample.

$$\text{Root mean-squared error} = \sqrt{\frac{(X_1-Y_1)^2+\ldots+(X_n-Y_n)^2}{n}}$$

3.5.2 Method

- Cross Validation

Cross-validation is a statistical technique for assessing how the results of a statistical analysis will generalize to an independent data set. In cross validation, data will dividing into two segments. The first segments used to learn or train a model and the other used to validate the model. The cross validation form used in this research is k-fold cross-validation and represent k with 10. This is other form of cross valiation.

- K-fold cross-validation
- 2-fold cross-validation
- Repeated random sub-sampling validation
- Leave-one-out cross-validation

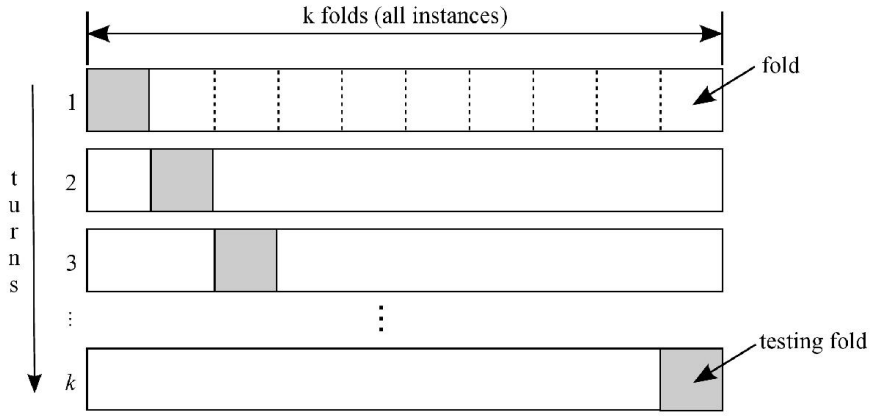

Figure 3.14: Cross validation technique for validate data

In stratified k-fold cross-validation, k is sized of segments or folds. k-1 folds are used for learning and remaining for testing. In data mining and machine learning 10-fold cross-validation (k= 10) is the most common.

CHAPTER IV
RESULT AND DISCUSSION

In order to compare different downscaling approaches, all algorithms each approaches considers all grid points that have the potential to influence the target meteorological station. The performance of each approached has been assessed in two different aspects as below.

1. Correlation between predicted and observed precipitation (R).
2. Root mean square error (RMSE) that measure of the difference between predicted values by a model and the actually values observed from the Real physical sites.

4.1 Data transformation comparison

There are three pre-processing algorithms as applied to target GFDL-R30 with six observation data from real physical sites. The goal of this preliminary analysis is identify the data transformation that yielded the best performance of MLR.

Table 4.1: Data information that used for compare pre-processing methods

Scope	Detail
GCMs information (Predictor)	• GCMs: GFDL-R30 • Base years: A.D.1965-1990 • Time resolution : Monthly • Size: Latitude 2.20° x Longitude 3.75° • Predictor variable : - Temperature, Atmospheric Pressure, - Precipitation, Solar Radiation, Evaporation
Observation data from Meteorological station. (Dependent)	Chiang Mai, Udon Thani, Bangkok, Ubon Ratchathani, Nakhon Ratchasima, Phuket

The process flow is shows in Fig. 4.1 and the results are summarized in Fig. 4.2 and 4.3.

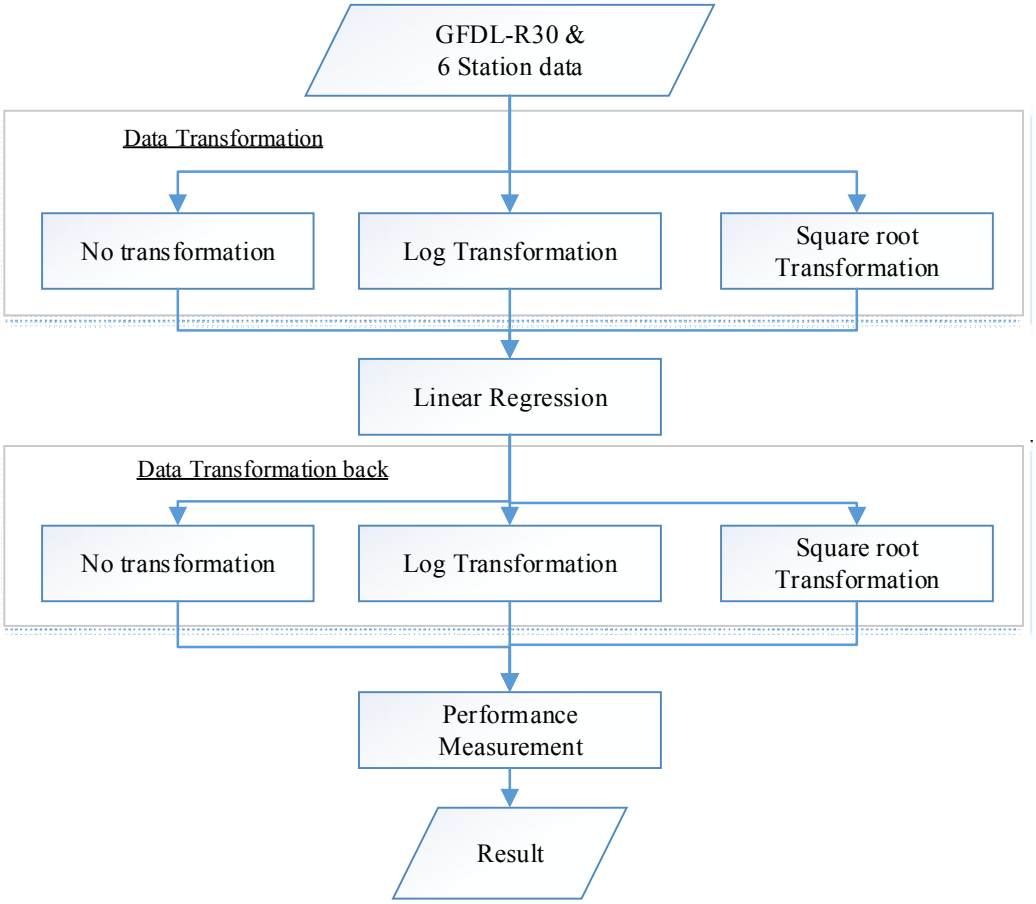

Figure 4.1 Process flow for pre-processing methods comparison

The process stating with data preparation from SQL Server 2012 Express. Then we classify observation data to three group are no transformation, logarithm transformation and square root transformation. Each group contains with real meteorological data from Chiang Mai, Udon Thani, Bangkok, Ubon Ratchathani, Nakhon Ratchasima and Phuket. All of groups was preprocess by multiple linear regression and transform data back to normal. After that we compare predicted data and original for correlation coefficient and RMSE. The result was validate by 10-Fold cross validation again.

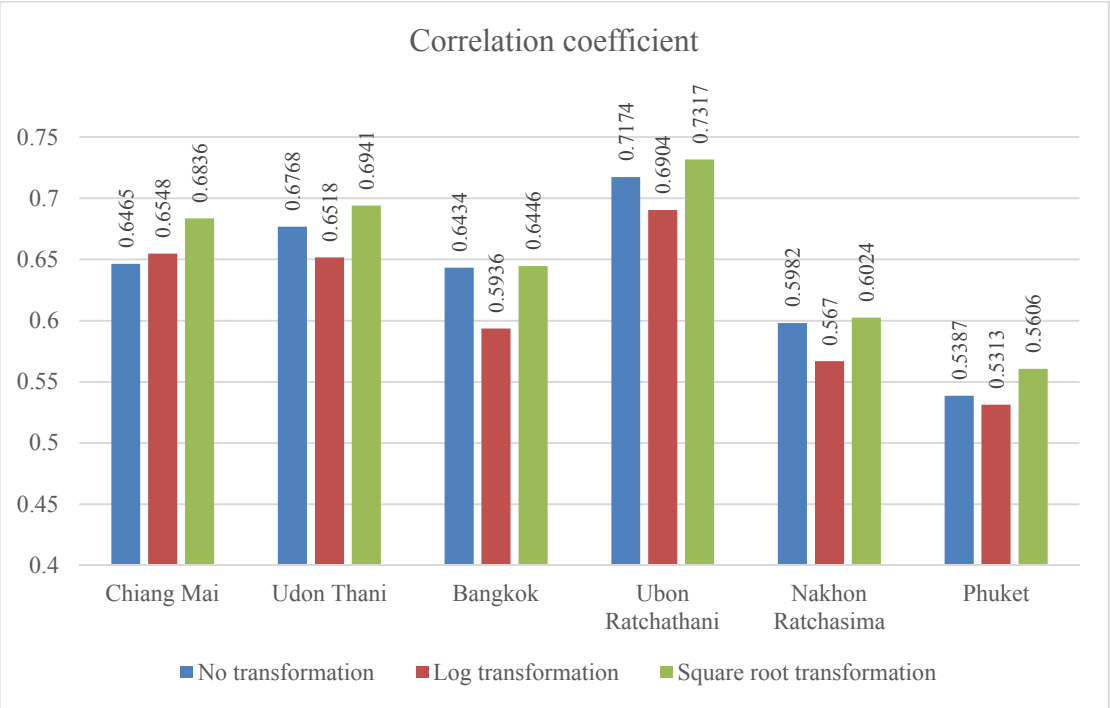

Figure 4.2 Correlation coefficient comparison between each data transformation algorithms

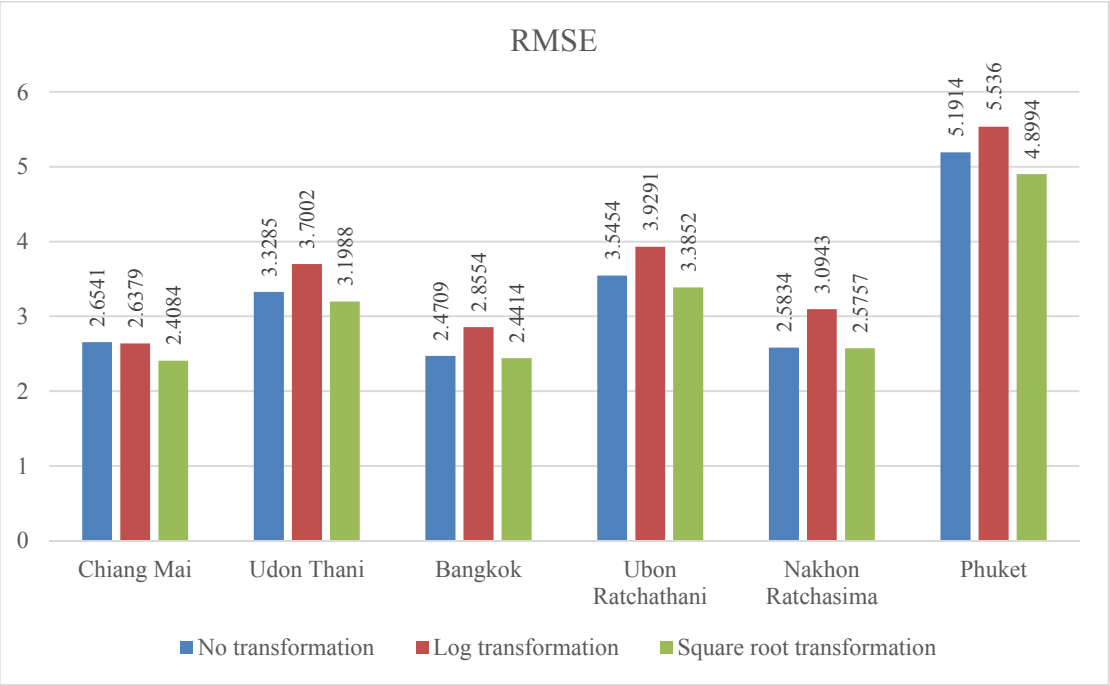

Figure 4.3 Root Mean Square Error comparison between each data transformation algorithms

From the result, we found log transformation give a correlation worse than no transformation but square root transformation give the slightly better correlation than no transformation data. However we can conclude that all algorithms of data transformation don't significant to increase strength and relationship for observation data and GCMs. So we will select no transformation to apply with our model.

4.2 Data classification

In this section we changed GCMs model from GFDL-R30 (Monthly data) to HadGEM (Daily) to increase data frequency for modeling. However HadGEM contains grid too much for modeling. So we have to capture only grid that cover Thailand area from Lat 5.625 x Long 96.5625 to Lat 20.625 x Long 105.9375. Total 78 Grids. This is detail of HadGEM GCMs that used to this experiment.

Table 4.2 Data information that used for classification experimental

Scope	Detail
GCMs information (Predictor)	• GCMs: HadGEM • Data Range: Dec , 1949 to Nov , 2005 • Time resolution : Monthly • Size: Latitude Lat 5.625 x Long 96.5625 to Lat 20.625 x Long 105.9375 (78 Grids) • Predictor variable : Daily-Mean Near-Surface Wind Speed, Sea Level Pressure, Precipitation, Near-Surface Specific Humidity, Near-Surface Air Temperature, Daily Maximum Near-Surface Air Temperature, Daily Minimum Near-Surface Air Temperature
Observation data from Meteorological station. (Dependent)	• Chiang Mai • Bangkok • Phuket

After capture grid to cover Thailand area, we applied classification technique to divided data to the day that no rainfall and the day that rainfall. In Thailand, Meteorological Department classified the intensity rule of rainfall divided

into four classes (unit: mm): light rain (0.1-10), moderate rain (10.1 – 35), heavy rain (35.1 – 90), very heavy rain (> 90.1).

Figure 4.4 Rain gauge used to gather and measure the amount of liquid precipitation over a set period of time

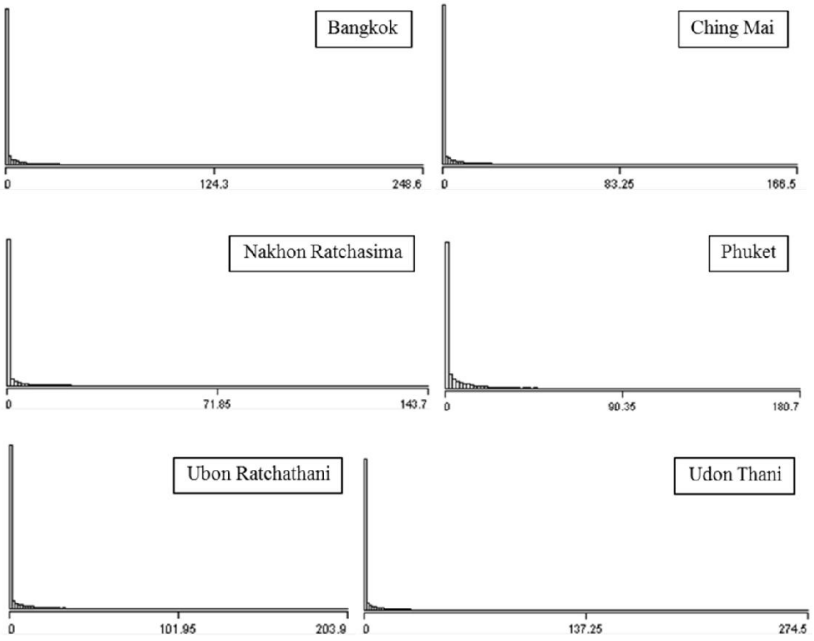

Figure 4.5 Histogram compare observation data from each station

Classification is the technique that used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. We apply classification technique cut off precipitation with level 0, 0.3, 0.6, 0.9, 1.2 and 1.5mm then remodeling multiple linear regression. This is process flow for apply classification to model.

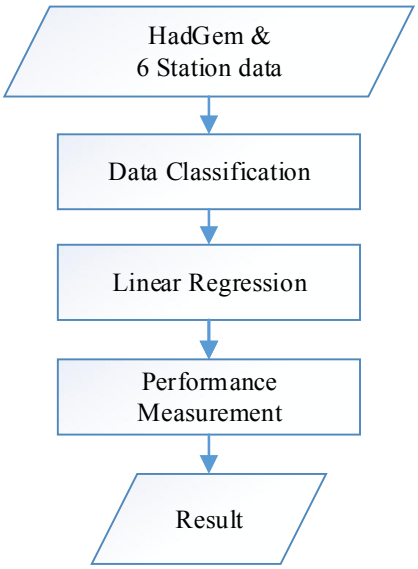

Figure 4.6 Process flow after apply classification to model

The process nearly same with previously section. It started on query HadGEM GCMs and observation from MS SQL Server 2012. After that we vary cutoff level for classify data to light rain and heavy rain. All of groups was preprocess by multiple linear regression and transform data back to normal. After that we compare predicted data and original for correlation coefficient and RMSE. The result was validate by 10-Fold cross validation again.

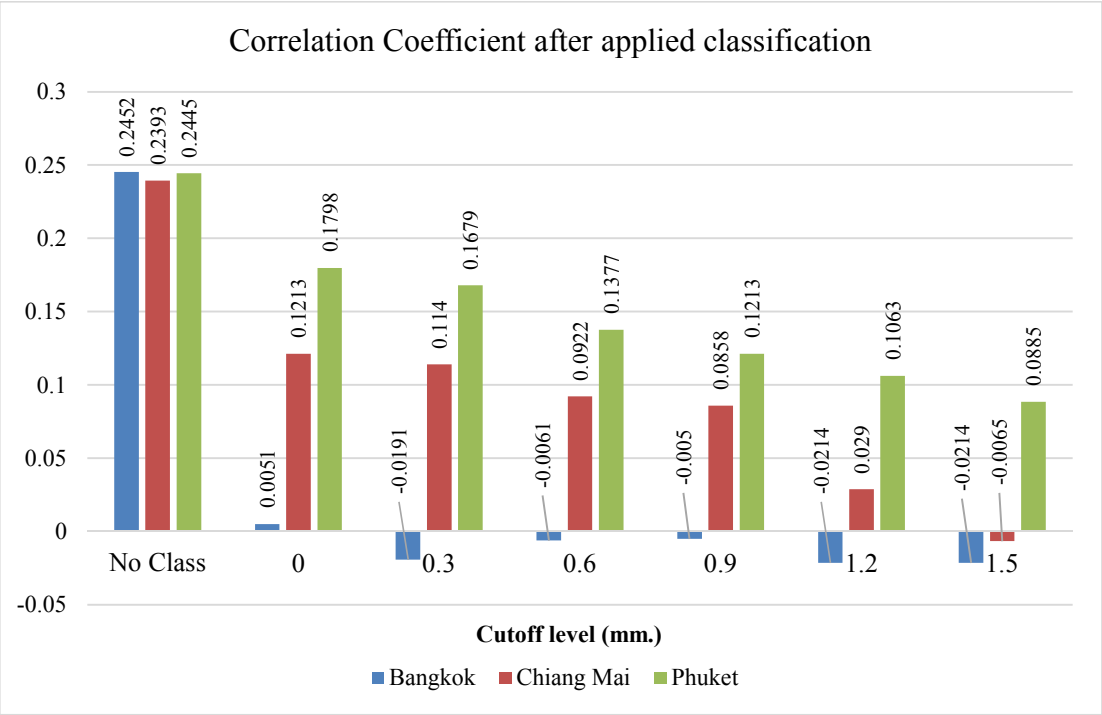

Figure 4.7 Correlation coefficient after applied classification with each cutoff level

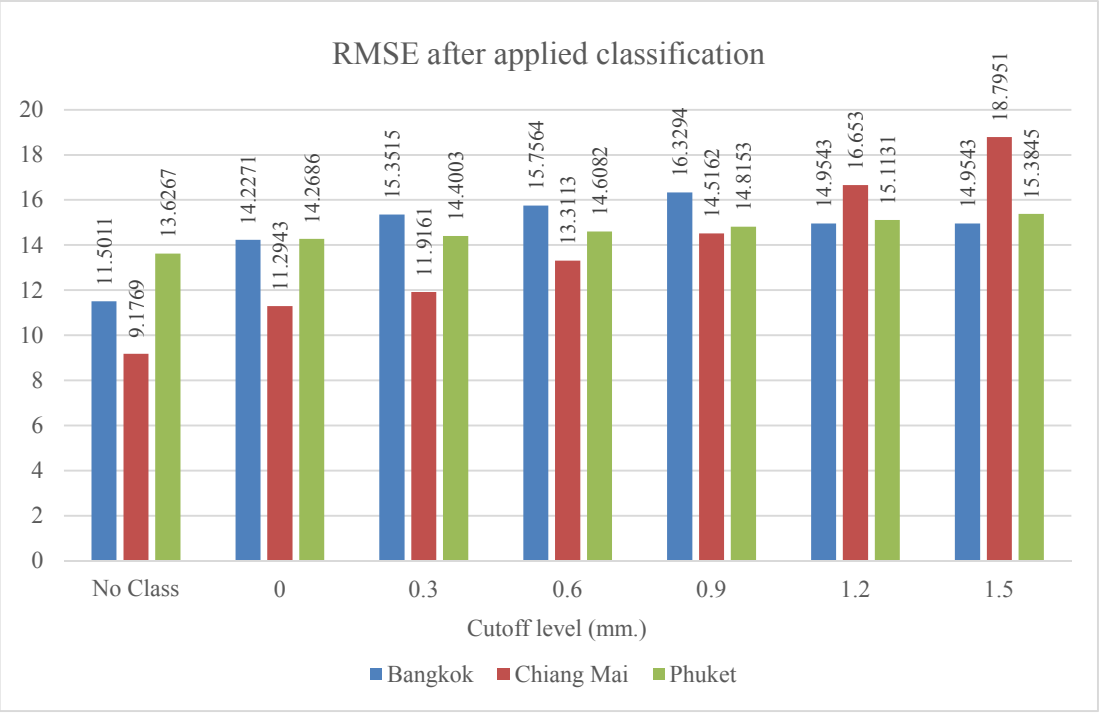

Figure 4.8 RMSE after applied classification with each cutoff level

4.3 Development result

The previous section assessed the performance of models when applied pre-processing and classification can't increase correlation strong between observation and predicted data. So in this section we will select post-processing for comparison are multiple linear regression and gaussian process.

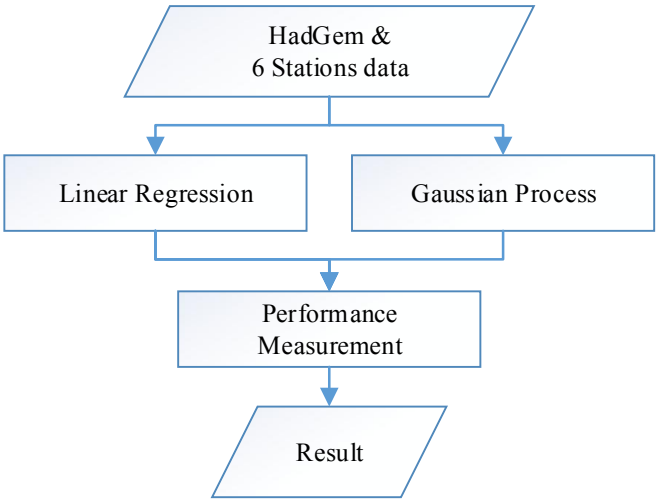

Figure 4.9: Process flow for post-processing comparison

The process starts at data preparation from MS SQL server 2012. After that we pass the data to linear regression and gaussian process for modeling. The predicted data was compared performance by correlation coefficient and RMSE.

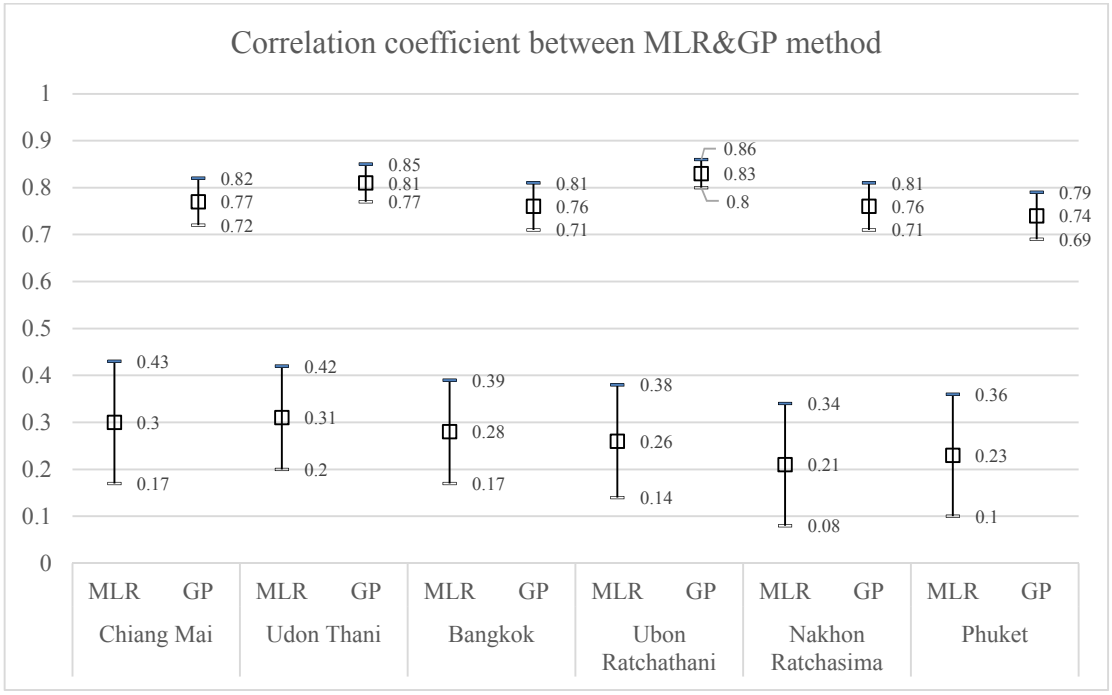

Figure 4.10 Correlation coefficient after apply cross validation compare between each method

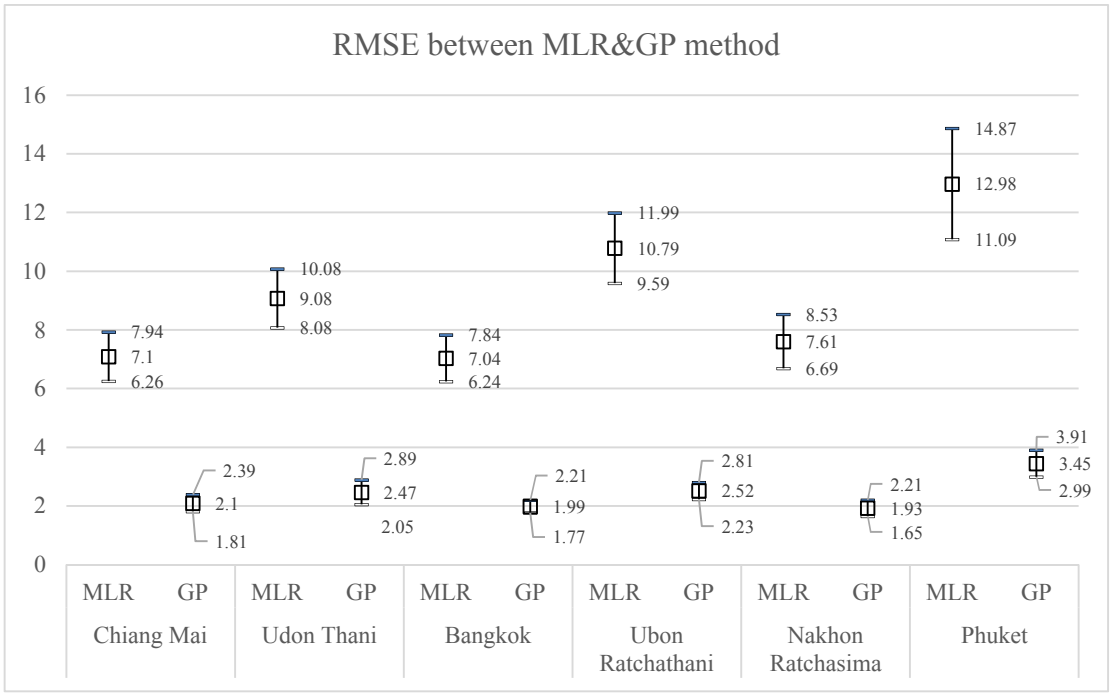

Figure 4.11 RMSE after apply cross validation compare between each method

From the experimental data, we found correlation from GP higher than 0.5 on every meteorological station that represent high correlation while MLR less than 0.4 mm. Root Mean Square Error is same direction with correlation. RMSE from GP method still lower than MLR

4.4 Discussion

4.4.1 Selecting pre-processing algorithm

In this research, we select data transformation for testing GFDL-R30 GCMs from Geophysical Fluid Dynamics Laboratory, The United States. Because of precipitation data is not normally distributed and therefore do not meet the assumptions of parametric statistical tests. We expect to transforming the data will make it fit the assumptions better.

From the experimental result, data transformation both of logarithm transformation and square root transformation can't increase relation between observation data and prediction data. Because of performance measurement both of correlation co and RMSE give the result worse than no transformation data.

Although no data transformation give high correlation between actual precipitation and prediction but the result of prediction contain with negative number too much (precipitation data can't be negative number in the real world). While logarithm transformation give prediction data in negative lower and nearly zero in square root transformation.

No transformation		
Actual Precip	**Prediction**	**Error**
0	-1.029	-1.029
0	-0.152	-0.152
0	-0.241	-0.241
0	-0.104	-0.104
0	-0.854	-0.854
0	-0.903	-0.903
0	-0.086	-0.086

square root transformation		
Actual Precip	**Prediction**	**Error**
0	0.591	0.591
0	0.305	0.305
0	0.477	0.477
0	0.254	0.254
0	0.45	0.45
0	0.283	0.283
0	0.057	0.057

Figure: 4.12 comparison prediction data between no transformations and square root transformation

4.4.2 Classification Data

In this research, classification was applied to daily observation data for cutoff data from the day that on rainfall. Propose of classification is to increase strength on development phase. Although we cutoff light rain data from all data and prediction only heavy rain. However result from the experiment show strength of observation data and prediction data lower than without classification.

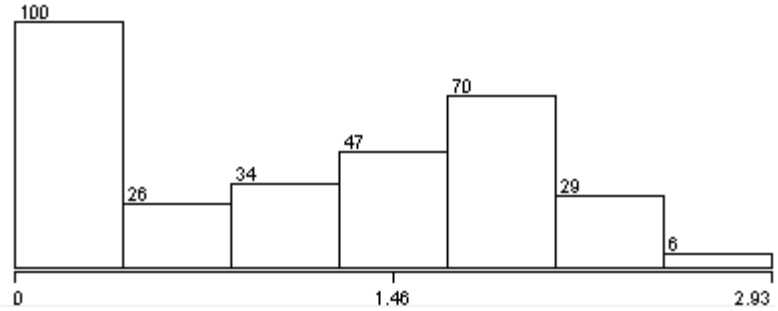

Figure 4.13: Histogram of average monthly precipitation.

We can conclude this result of this experiment to two topic as below.

1. Natural of precipitation data: light rain can help to increase yield for correlation.
2. Size of data after classified less than before classify so it effect to our modeling.

CHAPTER V
CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This thematic has presented a methodology for downscaling precipitation predictions comparison of linear and non-linear MOS methods, aim at the prediction of precipitation on target meteorology station in Thailand, based on the output of two different GCMs (GFDL-R30 from Geophysical Fluid Dynamics Laboratory and HadGEM from Hadley Center Metrological Office). The both of global circulation models captured precipitation factor are temperature, pressure, humidity and wind speed as the grid point around study station.

In order to compare different downscaling approaches, the performance of each approached has been assessed in two different aspects are correlation coefficient (R) and root mean square error (RMSE) between predicted and observed precipitation.

The first step is preliminary study on preprocessing data. The technique was used for preprocessing data in this study is data transformation. Data transformation will make precipitation data that not normally distributed to meet the assumptions of parametric statistical tests. The three data transformation algorithms was applied to target GFDL-R30 with six observation data from real physical sites. The result showed this approach can't increase correlation between observation and predicted data.

Next step on this thematic is classification data. Propose of classification is divided light rain and heavy rain from all data. The result still showed this technique can't help to improve performance of model.

This last step is comparison of data post-processing approaches. The gussian process, non-linear regression was compared with linear regression

5.2 Recommendation

- **Feature Selection on linear regression**

On linear regression, we don't select any feature selection method to apply on the model. The initial result that was tested found feature selection can help to improve correlation result of linear regression.

So we recommend feature selection for selection attribute or grid that has potential effect to precipitation on target station.

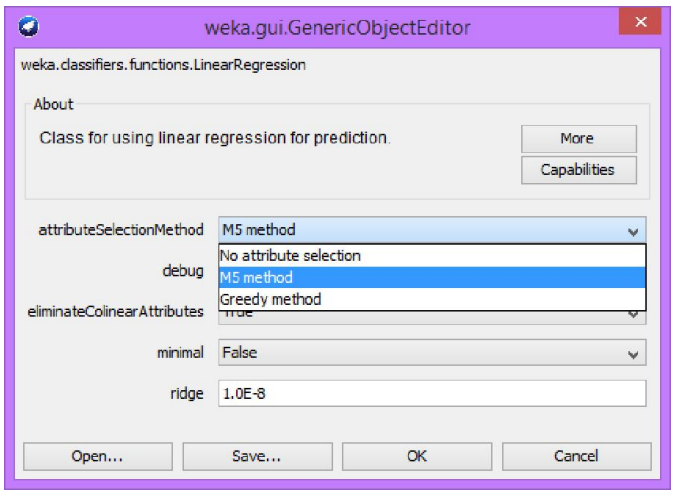

Figure 5.1: Attribute Selection Method of Linear Regression in Weka

REFERENCE

1. Sailor et al, T. Hu, X. Li, J.N. Rosen. (2000). A neural network approach to local downscaling of GCM output for assessing wind power implications of climate change.
2. Radan Huth. (1999). Statistical downscaling in central Europe: evaluation of methods and potential predictors
3. Dibike and Coulibaly, Paulin Coulibaly. (2006). Temporal neural networks for downscaling climate variability and extremes.
4. Tripathi et al. (2006). Downscaling of precipitation for climate change scenarios A support vector machine approach.
5. E. Eccel. (2007).Prediction of minimum temperatures in an alpine region by linear and non-linear post-processing of meteorological models
6. กัณฑรีย์ บุญประกอบ. (2010). การสร้างภาพจำลองของการเปลี่ยนแปลงภูมิอากาศในประเทศไทยโดยการย่อส่วนแบบจำลองภูมิอากาศโลก
7. Golfarelli, M. & Rizzi, S. (2009). Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill Osburn. ISBN 0-07-161039-1
8. "Statistical Sampling and Regression: Simple Linear Regression". (2013). [Online]. Available:http://ci.columbia.edu/ci/premba_test/c0331/s7/s7_6.html.
9. D.S. Wilks. (2006). Statistical methods in the atmospheric sciences. p.181

BIOGRAPHY

NAME	Mr. Yuranan Jaiyeakyen
DATE OF BIRTH	30 July 1984
PLACE OF BIRTH	Nakhonpathom, Thailand
INSTITUTIONS ATTENDED	Mahidol University, 2004-2007 Bachelor Degree of Engineering (Industrial Engineering) Mahidol University, 2010-2013 Master of Science (Technology of Information System Management)
HOME ADDRESS	42/4 Moo 3 Malaiman Road, Tambon Tap Luang, Amphoe Mueang Nakhon Pathom 73000 Tel. 080-910-6525 E-mail : yuranan@gmail.com