CHAPTER 2

Methodology

This chapter presents graphical and statistical methods. These methods contain the time series plot, auto-correlation function plot, bubble chart, linear regression analysis and correlation, time series analysis, factor analysis and multivariate linear regression analysis. For graphical and statistical analyses, they were carried out using R program (R Development Core Team, 2009).

2.1 Data source and data management

There are two different data sources for these studies, one for climate change and the other one for solar radiation absorption analysis.

Temperatures

Data used in the climate change study were obtained from the Climate Research Unit (CRU), United Kingdom (CRU, 2009), (http://www.cru.uea.ac.uk/) and described in detail by Brohan *et al.* (2006). CRU provides monthly temperature anomalies for 5^o by 5^o latitude-longitude grid-boxes on the earth's surface, based on data collected from weather stations, ships, and more recently satellites. These data were obtained as excel format which were modified and entered into computer text files suitable for analysis. The anomaly data is raw monthly temperature subtracted by monthly average temperature from 1961 to 1990. Thus the selected data for this study was converted to raw temperature by adding back the monthly average temperature from 1961-1990 in each grid-box to temperature anomaly data. There are more than 200 missing data in this area thus four 5^o by 5^ogrid-boxes were combine to be 10^o by 10^o grid-boxes. This South-East Asia data incorporates 40 regions of 10^o by 10^o grid-boxes which were

designed like ice-blocks in an igloo. This type of grid-box can show correlation among 6 different sides of adjoining grid-boxes. These areas are located in latitude 25^{o} S to 25^{o} N and longitude 75^{o} E to 160^{o} E, and compose of all or part of 11 Southeast Asian countries including Northern Australia, Western India, Bangladesh, Nepal, Bhutan, Southern China, the Indian Ocean and the west of the Pacific Ocean, as shown in Figure 2.1.

The data is the monthly temperature averages for 10^{o} by 10^{o} latitude-longitude grid-boxes on the Earth's surface in Southeast Asia over a 100 year period (1200 months) during 1909 to 2008. Temperatures were studied for three overlapping 36 year periods; the first period: 1909 - 1944, the second period: 1941 - 1976, and the third period: 1973 – 2008.

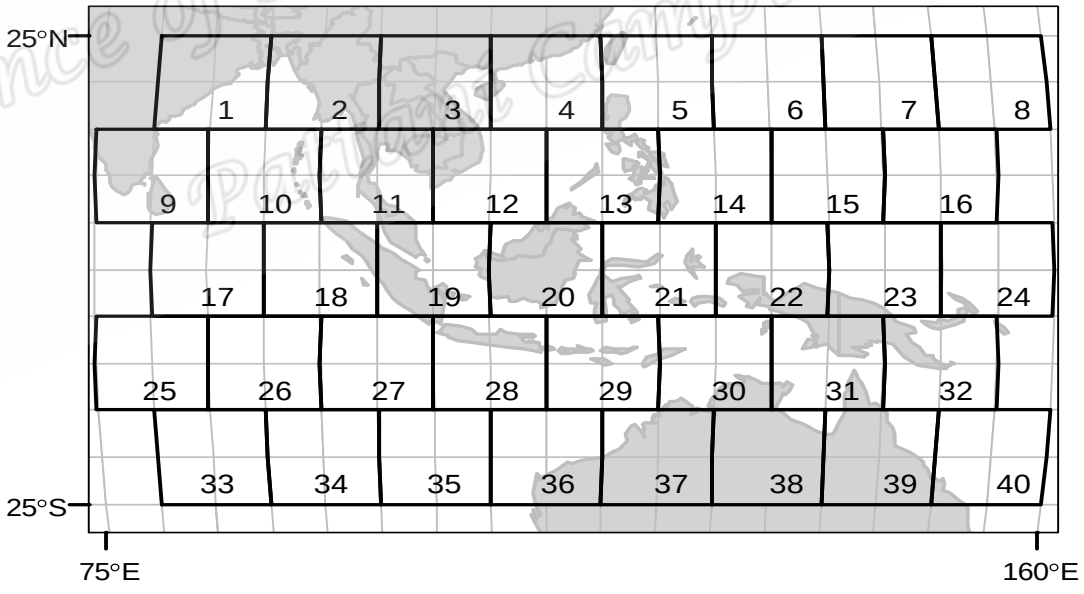

Figure 2.1 The study area for climate change

Solar radiation energy

Data used in the solar radiation energy study were obtained from the Australian Bureau of Meteorology (http://www.bom.gov.au/climate/data/). This department provides daily and monthly climate statistics, weather observations, and solar exposure data files for 144 stations in Australia over the years of 1990 to 2012. Stations were selected for our study, as shown in Figure 2.2. Daily solar data observed (in mega joules/square metre) are provided at the station. We omitted one observation in each leap year, i.e. the observation on February 29th, to maintain the same amount of observations for each year. For our study, there are 365 observations in each year and therefore there are 8,395 observations for each station over 23 years.

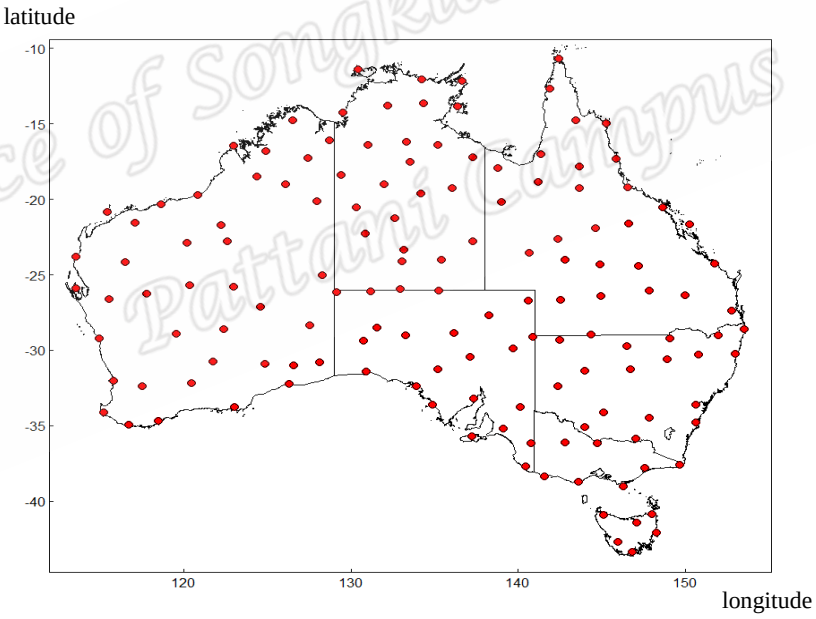

Figure 2.2 The study area for solar radiation energy

2.2 Variables

For climate change study, dependent variables are average monthly temperatures in each grid-box and independent variables are times (months). For solar radiation

energy study, dependent variables are averaged daily solar observed values at the station and independent variables are times (days).

2.3 Study diagrams

As this study had the two study diagrams for temperature change and solar radiation absorption, we have two study diagrams as follows:

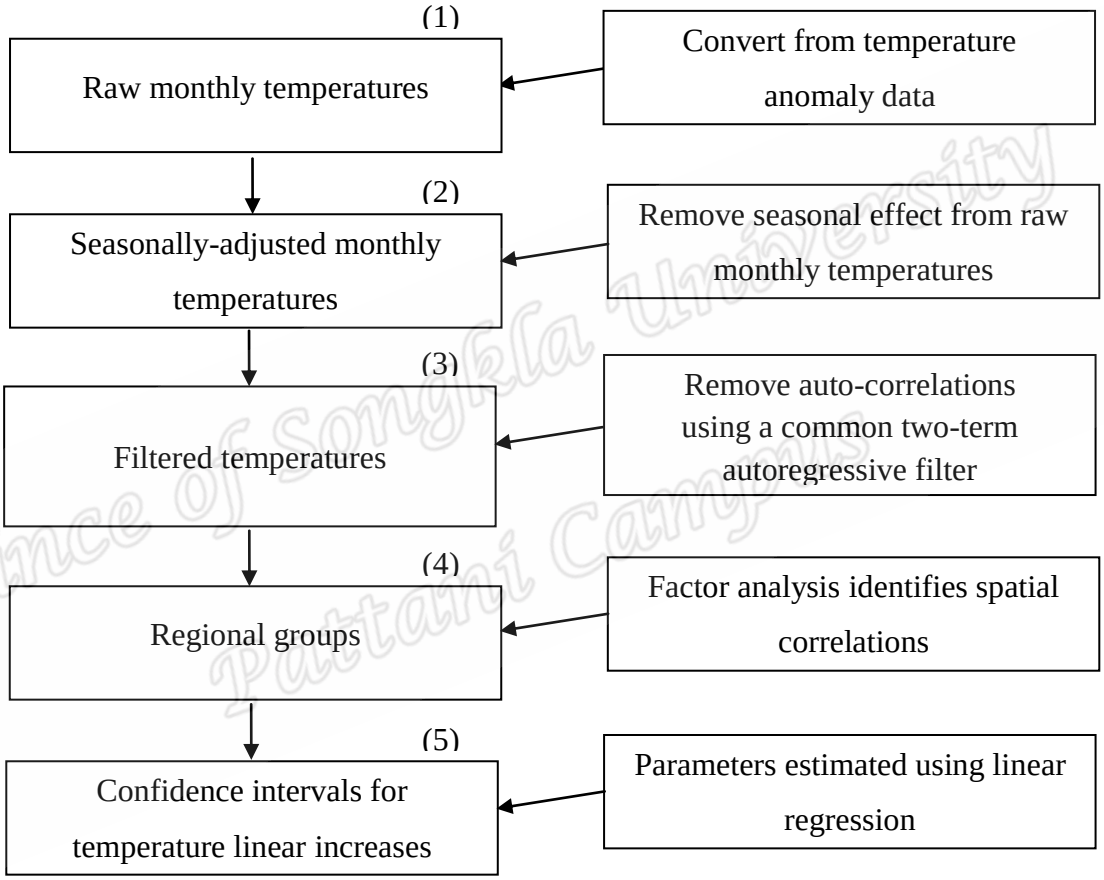

Figure 2.3 The diagram of climate change in Southeast Asia

This figure shows how to investigate the trend of temperature change in the Southeast Asia as the following; (1) The anomaly monthly data was converted to raw temperature by adding back the monthly average temperature from 1961-1990 in each grid-box. (2) The seasonal monthly variation (raw temperatures) in each grid-box was adjusted by subtracting the monthly averages (seasonal effect) and then adding back the overall mean temperature. (3) Next we removed auto-correlations by using a

second order autoregressive process, yielding the filtered temperatures. (4) Factor analysis was then applied to the filtered temperatures to identify spatial correlations giving adjoining grid-boxes combined into regional groups. (5) Linear regression models were used to fit parameters for each factor. These parameters were used to explain the increase of temperatures with confidence intervals. Next is the diagram of solar radiation absorption in Australia.

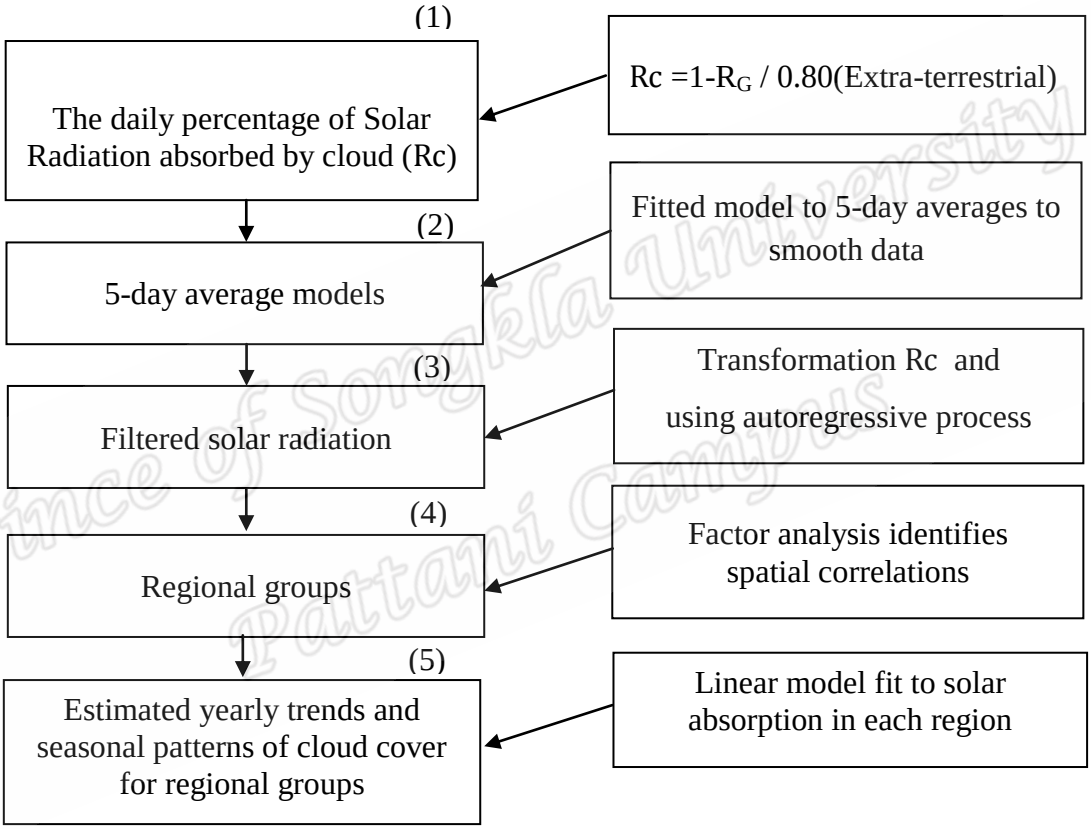

Figure 2.4 The diagram of solar radiation absorption in Australia

This study diagram shows how to account for solar radiation energy absorption in Australia as the following; (1) The amount of solar energy reaching the lower atmosphere is a constant fraction (P_0) of the extra-terrestrial solar radiation (R_E) in mega joules/square metre. In this paper, $P_0 = 0.80$, the constant value is observed from maximum solar radiation in each station. The daily percentage of solar radiation absorbed by clouds (Rc) was calculated by subtracting daily solar (R_G) from 0.80 the

extra-terrestrial. (2) To reduce the serial correlation between daily observed radiation levels, these data are aggregated into 5- day averages. Then we fitted linear models to average values of Rc. (3) The residuals from fitted models assessed by quantile plots were not normally distributed. Thus Rc was transformed (normality achieved by square root transformation of positive percentages), and removed auto-correlations (first order autoregressive process). The filtered solar radiation was the result. (4) Factor analysis was used to analyse filtered solar radiation and separated stations into regional groups. (5) The yearly trends and seasonal patterns for the regional groups were estimated.

2.4 Graphical methods

Time series plot (Cryer, 1986) is the most frequently used form of graphic design. It is the first step in any time series analysis by plotting of the observed values of a series on the vertical axis versus time on the horizontal axis. In this study, average monthly temperatures were plotted on the vertical axis versus months over the time period we studied on the horizontal axis.

Auto-correlation function plot is a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing auto-correlations for data values at varying time lags. If random, such auto-correlations should be near zero for any and all time-lag separations. For the k order auto-correlation, the lag is k time unit. The k order auto-correlation is the correlation coefficient of the first $n-k$ observations, $y_t, t = 1, 2, \ldots, n-k$ and the next $n-k$ observations, $x_t, t = k+1, k+2, \ldots, n$. The formula for auto-correlation coefficient at lag k (r_k) is of the form

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t-\bar{y})(y_{t+k}-\bar{y})}{\sum_{i=1}^{n}(y_t-\bar{y})^2} \quad , \tag{2.1}$$

where $\bar{y}$ is the mean of observations.

Bubble chart is a technique in which a set of numeric quantities is represented by closely packed circles whose areas are proportional to the quantities. In this study bubble charts were used to display the correlation matrix of the filtered temperatures and can be compared to the correlation in terms of their size.

2.5 Statistical methods

These statistical methods were used both in climate change and solar radiation energy studies.

Linear regression analysis and correlation

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. A linear regression line of variable Y on variable X has the form

$$Y = \beta_0 + \beta_1 X + \in \quad , \tag{2.2}$$

where β_0 and β_1 are the parameters of the model and $\in$ are error variable. We can use the information provided by observation to give us estimates b_0 and b_1 of β_0 and β_1 ; thus we can write

$$\hat{Y} = b_0 + b_1 X \quad , \tag{2.3}$$

where $\hat{Y}$ denotes the predicted value of Y for a given X, when b_0 and b_1 are determined. Equation (2.3) could then be used as a predictive equation. For linear

regression analysis, a valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables. In a sample size n, the formula to account for the correlation coefficient between (x_1, y_1), (x_2,y_2),…..,(x_n,y_n) is of the form

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\{\sum_{i=1}^{n}(x_i-\bar{x})^2\}^{1/2}\{\sum_{i=1}^{n}(y_i-\bar{y})^2\}^{1/2}} \quad , \tag{2.4}$$

Time series analysis with autoregressive processes

A time series is a collection of observations made sequentially in time. Methods of analyzing time series constitute an important area of statistics (Chatfield, 1996). Most time series patterns can be described in terms of two basic classes of components: trend and seasonality. Much time series data can be adequately approximated by a linear function. If there is a clear monotonous nonlinear component, the data first needs to be transformed to remove the nonlinearity. Usually a logarithmic, exponential, or polynomial function can be used. Most time series data consist of elements that are serially dependent, thus we can describe the correlation of these observed data from auto-correlation coefficients. Autoregressive (AR) models were also used to account for the auto-correlations among the residuals from the fitted linear models (Venables and Ripley, 2002). AR(p) indicates an autoregressive model of a p th-order and the AR(p) model is defined by the equation

$$Z_t = \sum_{i=1}^{p} \theta_i Z_{(t-i)} + a_t \quad , \tag{2.5}$$

where $Z_t = y_t - \hat{y}_t$ is the residual value at the time t, $t-i$ is the order of the past time, $\theta_1, \theta_2, \ldots, \theta_p$ are the parameters of model, and a_t is the value that is not

explained by the past values. The series is assumed to be stationary, a_t and $Z_{(t-i)}$ are independent.

Factor analysis

Factor analysis is a mathematical model which attempts to explain the correlation between a large set of variables in term of a small number of underlying factors. It addresses the problem of analyzing the structure of the interrelationships (correlations) among a large number of variables by defining a set of common underlying dimensions, known as factors (Hair *et al*., 2009). Data reduction can be achieved by calculating scores for each underlying dimension and substituting them for the original variables. Factor analysis is an interdependence technique in which all variables are simultaneously considered, each related to all others, and still employing the concept of multivariate, the linear composite of variables.

Factor analysis was applied to identify correlations between outcome variables by maximizing the likelihood of covariance matrix and minimizing the correlations between the factors for a specified number of factors. In this study the factor model reduces spatial correlations. The factor model formulation with *m* common factors, involving a weighted sum of factors to the data, is of the form

$$y_{ij} = \mu_j + \sum_{k=1}^{m} \lambda_j^{(k)} \emptyset_i^{(k)}, \qquad (2.6)$$

where y_{ij} are variables in factor j, time i, μ_j are the constant, $\lambda_j^{(k)}$ are termed loading factors and $\emptyset_i^{(k)}$ are the common factors. We used the covariance matrix of estimated slopes in the regression model to fit the factor model.

Multivariate linear regression analysis

Multivariate linear regression is the extension of multiple linear regression to allow for several correlated outcome variables. Multivariate regression estimates the same coefficients as one would obtain using separate univariate regression models. In addition, multivariate regression, being a joint estimator, also estimates the between-equation covariance. Suppose that data are available for *n* observations, and the response variables are arranged into a *matrix* whose columns are *p* outcome variables and rows correspond to the *n* observations. The model (Mardia *et al*., 1980) is defined in matrix form as

$$\mathbf{Y}_{(n\times m)} = \mathbf{X}_{(n\times q)}\,\mathbf{B}_{(q\times m)} + \mathbf{E}_{(n\times m)}, \qquad (2.7)$$

In this formulation $\mathbf{Y}_{(n\times m)}$ is an observed matrix of *p* response variables on each of the *n* observations, $\mathbf{X}_{(n\times q)}$ is the matrix of *q* predictors (including a vector of 1s) in columns and *n* observations in rows, $\mathbf{B}_{(q\times m)}$ contains the regression coefficients (including the intercept terms), and $\mathbf{E}_{(n\times m)}$ is a matrix of unobserved random errors with mean zero and common covariance matrix $\mathbf{\Sigma}$. Thus, the error terms associated with different response variables may be correlated. The estimated model in matrix form is given by

$$\widehat{\boldsymbol{Y}}_{(n\times m)} = \mathbf{X}_{(n\times q)}\,\widehat{\boldsymbol{B}}_{(q\times m)} \quad , \qquad (2.8)$$

For our study $n = 432$, m is the number of grid-boxes in each factor, q is the number of the estimated parmeter, $\widehat{\boldsymbol{Y}}_{(n\times m)}$ is a matrix of the estimated filtered temperatures, $\mathbf{X}_{(n\times q)}$ is a matrix of month elapsed (months measured in decades), $\widehat{\boldsymbol{B}}_{(q\times m)}$ is a matrix of the estimated parameters.

The multivariate linear regression method also gives the covariances between the estimated parameters or variance-covariance matrix. This model has the additional advantage in that it takes into account correlations between data in different factors. In this study a multivariate linear regression model is used in each factor. The linear model is also used to predict temperature changes over the future.