Title
A Geospatial Analysis on Empirical Distribution with Applications to Atmospheric Infrared Sounder Level 3 Data

Author
Liu, Xian

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

A Geospatial Analysis on Empirical Distribution with Applications to Atmospheric Infrared Sounder Level 3 Data

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Xian Liu

2012

Abstract of the Thesis

A Geospatial Analysis on Empirical Distribution with Applications to Atmospheric Infrared Sounder Level 3 Data

by

Xian Liu

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Mark S. Handcock, Chair

The atmospheric Infrared Sounder (AIRS), a sensor aboard NASA's Aqua satellite, has been collecting temperatures, water vapor mass-mixing ratios, and cloud fractions at various atmosphere pressure levels since its launch in mid-2002. The AIRS level 3 Quantization (L3Q) product summarizes atmospheric data in each $5^\circ \times 5^\circ$ latitude-longitude grid box during a time period by a set of representative vectors and their associated weights, which can be treated as an empirical distribution. A study on the empirical distribution of AIRS L3Q data might be useful in terms of summarizing the local climate pattern. This study provides insights on how statistical methods may extract more information from the AIRS L3Q data than from the simple sample average summary in each grid box.

The thesis of Xian Liu is approved.

Hongquan Xu

Rick Paik Schoenberg

Amy Braverman

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2012

To my parents . . .

Who always be there to support me

No matter what.

Table of Contents

1 Introduction **1**

2 Data **3**

2.1 Structure and Format 3

2.2 Entropy-constrained Vector Quantization 7

3 Geospatial Analysis **10**

3.1 Variogram 10

3.2 Kriging 12

3.3 Spatial Regression 13

4 Mixture Distributions **14**

4.1 Kernel Smoothing 14

4.2 Descriptive Statistics 15

4.3 Algorithm 17

4.4 Skewness 20

5 Geospatial Lattice Analysis by Month **29**

5.1 Temperature Distribution 2011-07 29

5.2 Temperature Distribution 2011-10 32

5.3 Temperature Distribution 2012-01 36

5.4 Temperature Distribution 2012-04 38

5.5 Humidity: Water Vapor Mixing Ratio 42

5.6 Correlation between Temperature Skewness and Water Vapor Skewness 45

6 Scientific Study 47

6.1 Water Vapor and Temperature 47

6.2 South East Asia 48

7 Conclusions 53

References 55

List of Figures

2.1 Partitioned Grid Boxes on the Earth 4

2.2 Data Matrix Format 7

2.3 Cluster Number Histograms of Four Months 9

2.4 Cluster Number Maps of Four Months 9

4.1 Aggregate Temperature Mean for Each Month 17

4.2 Aggregate Temperature Standard Deviation for Each Month . . . 18

4.3 Mixture Distribution of Temperature at Pressure Level 1000hPa . 20

4.4 Skewness Demonstration 21

4.5 Skewness in 500 Simulations 23

4.6 Temperature Distributions in 9 Different Cells around the Equator 24

4.7 Temperature Distribution Skewness at Each Pressure Level, 2011-07 25

4.8 Temperature Distribution Skewness at Each Pressure Level, 2011-10 26

4.9 Temperature Distribution Skewness at Each Pressure Level, 2012-01 27

4.10 Temperature Distribution Skewness at Each Pressure Level, 2012-04 28

5.1 Spatial Trend of Temperature Skewness 2011-07 30

5.2 Semi-variogram Estimates 2011-07 30

5.3 Variogram (Temperature Skewness at 1000hPa 2011-07) 31

5.4 Evaluate Kriging surface 2011-07 32

5.5 Spatial Trend of Temperature Skewness 2011-10 33

5.6 Semi-variogram Estimates 2011-10 34

5.7 Variogram (2011-10 Temperature Skewness at 1000hPa) 35

5.8 Evaluate Kriging surface 2011-10 35

5.9 Spatial Trend of Temperature Skewness 2012-01 36
5.10 Semi-variogram Estimates 2012-01 37
5.11 Variogram (Temperature Skewness at 1000hPa 2012-01) 37
5.12 Evaluate Kriging surface 2012-01 38
5.13 Spatial Trend of Temperature Skewness 39
5.14 Semi-variogram Estimates 2012-04 40
5.15 Variogram (2012-04 Temperature Skewness at 1000hPa) 41
5.16 Evaluate Kriging Surface 2012-04 41
5.17 The skewness of water vapor distribution in 4 months 42
5.18 Semi-variogram Estimates of Water Vapor Skewness for Each Month 43
5.19 Variogram of Water Vapor for Each Month 44
5.20 Global Linear Regression for Each Month 45
5.21 Regional Linear Regression for Each Month 46

6.1 Surface Temperature against Water Vapor, Physical Value 48
6.2 South East Asia 49
6.3 Skewness in Southeast Aisa 50
6.4 Polynomial Fit of Skewness V.S. Pressure Level 51
6.5 Maps of Polynomial Regression Coefficients 52

List of Tables

2.1 Data Fields 6

4.1 Summery Statistics of Skewness from 500 Iterations 22

5.1 Parameters in Spherical Models 44

6.1 Average Polynomial Regression Coefficients 50

CHAPTER 1

Introduction

Climate change has been a pressing environmental issue of the scientific community for decades. In recent years, scientists working on climate change have seen the explosion of the volume and complexity of available data from remote sensing satellites of the National Aeronautics and Space Administration (NASA). To reduce the massive data size in a traditional way, data are first partitioned into subsets by spatial location and time period. Each subset is defined as a latitude-longitude grid box. Then summary statistics, such as sample average and standard deviation, are used to represent each grid, and further analysis is applied to these summaries. This approach works reasonably well when the data distributions in different subsets are Gaussian and have similar shapes. However, this may not be true for most of the time, and any information other than the first two moments of the empirical distribution is lost in this initial data reduction step.

A new approach beyond the averages and standard deviations is required. A data reduction method based on K-mean clustering was adopted by NASA Infrared Sounder (AIRS) project and described in [1]. This clustering based approach has the potential in keeping the distribution configuration beyond sample means and standard deviation. Although the clustering based data summary preserves much more distributional information than sample averages, statistical analysis for this new type of data summary is not widely explored. Hence, a statistical analysis of this AIRS L3Q product is introduced in this thesis. Various

methods are used in the analysis, including the study of distribution skewness, kernel smoothing, variograms, kriging and spatial regression.

In climate models, it is often assumed that temperature and water vapor have a fixed Gaussian distribution, when in fact these distributions can be more variable over space and time. Thus, if the parameterization is allowed to be more flexible, a more realistic simulation of the atmosphere may be obtained. The detail of how large-scale atmospheric system work is still not quite clear for climate scientists [15]. Finding out their distributional characteristics are a step towards understanding their underlying mechanisms. Therefore, all the spatial analysis developed in this thesis is based on the distribution skewness. By doing so, one can inform the real behavior of temperature and water vapor distributions in the atmosphere.

In climate science, a classic equation describing the relation between water vapor pressure and concentration is defined as:

$$p = \frac{nRT}{V} \tag{1.1}$$

p is the pressure in Pa, V is the volume in cubic meters, T is the temperature in degrees Kelvin , n is the quantity of gas expressed in molar mass , R is the gas constant: 8.31Joules/mol/m^3. To convert the water vapour pressure to concentration in kg/m^3: $\frac{(Kg/0.018)}{V} = \frac{p}{RT}$, where p is the actual water vapor pressure [12]. The relation between water vapor and temperature is essential in climate science. It is well known that the physical values of temperature and water vapor have an exponential relation. In this thesis, I discuss the relation between temperature and water vapor in distributional characteristics, and this is discussed more concretely in Chapter 6. In addition, we are also interested in how temperature distribution changes as altitude varies. It helps climate scientists understand how convection affects the temperature distribution.

CHAPTER 2

Data

AIRS Level 3 Quantization (L3Q) products summarize climate data in each $5^\circ \times 5^\circ$ latitude-longitude grid box collected during a time period (a month or a season) by a list of representative vectors and their frequencies. Therefore, this multivariate empirical distribution in each grid box represents a record of the regional weather event during the corresponding period of time. This clustering based approach (ECVQ) [2] has the potential ability of keeping the distribution configuration beyond sample means and standard deviations. L3Q data is recorded in the format of HDF, which is not commonly used besides NASA's data product. In this Chapter, I briefly introduce the HDF data set and the ECVQ algorithm that is used to preserve the distributional characteristics of the raw data.

2.1 Structure and Format

In L3Q data set, the data are summarized by a set of representative vectors (mean of each cluster: $m_1, \ldots, m_L$) and their associated weights (observation frequency: $w_1, \ldots, w_L$), where L represents the cluster number. In this way, the data in the i_{th} grid are represented by $S_i \equiv (m_{i1}, \ldots, m_{iL}; w_{i1}, \ldots, w_{iL})$. While dramatically reducing the data size, the data summary S_i, like a multivariate empirical distribution, still keeps most information of the joint distribution of the data vectors in each grid cell. In addition, outliers in the distribution are retained.

The datasets in this application are the AIRS monthly L3Q products collected in July 2011, October 2011, January 2012 and April 2012. The data sets are accessible at `http://mirador.gsfc.nasa.gov/` [1]. Due to the difficulty of collecting data at some extreme areas, such as Tibet, Antarctica, etc, observations are missing at those areas in L3Q data sets. The complete data set contains the information in 2592 (72×36) $5^\circ \times 5^\circ$ grid boxes. However, there are some mising values in the 4 data sets in this context. There are 2369, 2367, 2369 and 2368 grid boxes with valid empirical distributions in the four different months, 2011-07, 2011-10, 2012-01 and 2012-04, respectively.

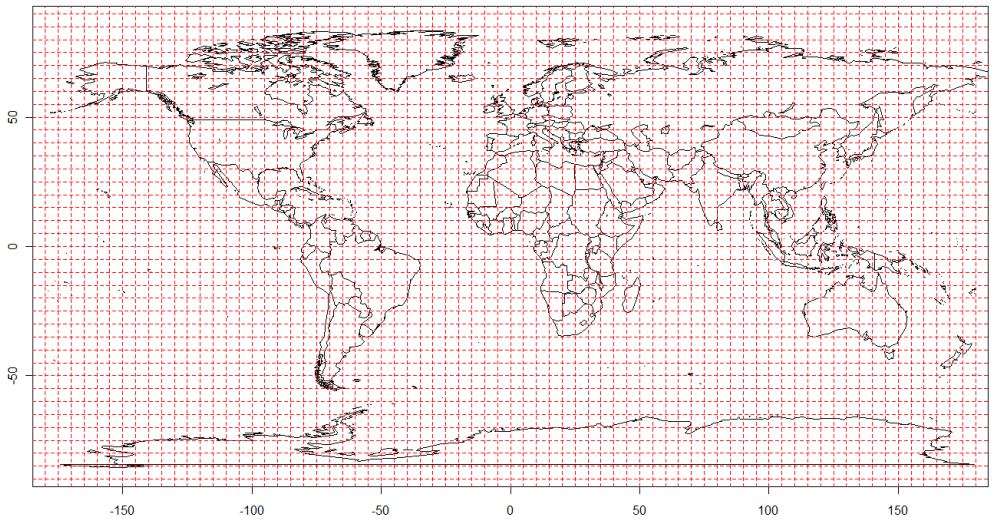

Figure 2.1: Partitioned Grid Boxes on the Earth

AIRS Level-3 quantization files are written in HDF-EOS4 format. HDF-EOS4 format is an extension of the HDF4 format to meet the requirement of the Earth Observing System (EOS) data products. These extensions facilitate the creation of Grid, Point and Swath data structures, in the case of AIRS Level-3 quantization products, they are of the grid structure [3]. When working with HDF-EOS files, one is not concerned with exactly how the data are stored physically, rather

one interacts with the data file by knowing the identifiers (filename, swath/grid names, parameter names, attribute names etc) and through a set of application programming interface (APIs) methods. Five categories of methods, access , basic I/O and inquiry are relevant for reading the data. Each AIRS Level-3 quantization product file contains a grid whose name is L3Quant. This grid is made up of four major structures: projections, dimensions, data fields and attributes.

Data fields in a grid data set are rectilinear arrays of two or more dimensions (a.k.a extra dimensions). They are related to each other by the common geolocation. In other words, a single set of geolocation information is used for all data fields within one grid data set. **Table 2.1** is the output from Matlab in this structure.

Scientific interest in these data lies in the spatial and temporal evolution of multivariate relationships among temperature and water vapor variables at multiple altitudes. Therefore, each grid cell is populated with discrete multivariate distribution estimates summarizing the most relevant of these variables: temperature and water vapor at the 11 pressure levels nearest to the surface, cloud fractions within these same levels, and three diagnostic quantities: the fraction of the footprints in the cell that cover land, the fraction of the footprints in the cell acquired during the day, and the fraction of the footprints in the cell deemed to be of less than fully acceptable quality.

Due to the complexity, HDF is a good way to store data, but not an easy format to perform analysis on. Thus, before we conduct any analysis, data cleaning of the original data file is necessary. For each month, I reconstruct the HDF data set and record them in 36 lists. Each list represents 72 matrices in each grid box at one latitude circle ($5^\circ \times 360^\circ$ rectangle). The data matrix has the

Content	Dimension
LatCenter:	[72x36 single]
LonCenter:	[72x36 single]
SouthLatBound:	[72x36 single]
NorthLatBound:	[72x36 single]
WestLonBound:	[72x36 single]
EastLonBound:	[72x36 single]
NumClusters:	[72x36 int16]
NumObsInCluster:	[72x36x100 int16]
ClusterMeanSquaredError:	[72x36x100 single]
ClusterDistortion:	[72x36x100 single]
Entropy:	[72x36x200 single]
GridCellMeanSquaredError:	[72x36x200 single]
NormalizedValues:	[4-D single]
PhysicalValues:	[4-D single]
PentadComposition:	[4-D int16]

Table 2.1: Data Fields

dimension of $n\times$ 37. The first dimension n means the cluster number in each grid box, which is provided by the ECVQ algorithm. The second dimension, 37, represents different variables as columns in each matrix. The first 35 columns are the physical values of temperature, water vapor profiles, cloud fraction profiles and 3 indicators. The last 2 columns are cluster weights and sum square error of each cluster respectively. We now reformat the data into 36 $\times$ 72=2592 matrices, with each matrix containing the data for the corresponding $5^\circ \times 5^\circ$ grid box.

Figure 2.2 shows the format of the data matrix in this thesis. tair is atmospheric temperature, h2ommr is water vapor content, and cldfrc is cloud fraction.

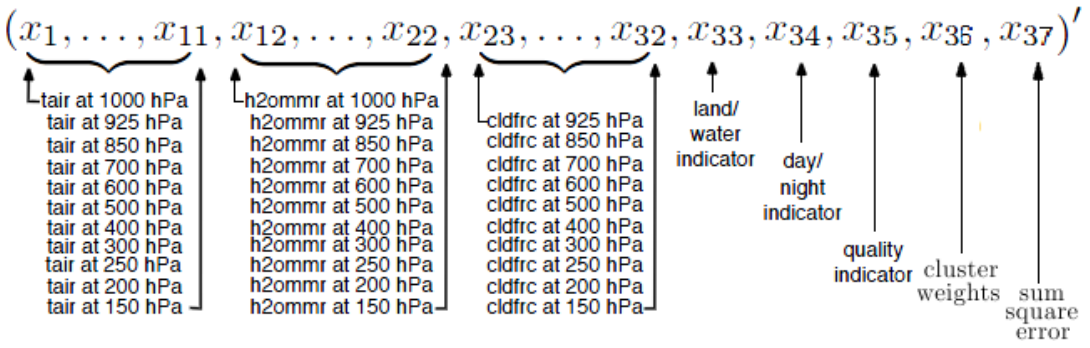

Figure 2.2: Data Matrix Format

Note that x_{33}, x_{34} and x_{35} are variables taking the value between 0 and 1. x_{36} is the vector containing the observation frequency of each cluster, which is used as weights of each cluster. x_{37} is the sum of square error over all pressure levels and all variables in each cluster. The reason for using only the 11 pressure levels nearest to the surface is that retrievals above that are not considered as reliable. In this thesis, data management is done in Matlab and Microsoft Excel, while the analyses are carried out in R.

2.2 Entropy-constrained Vector Quantization

Instead of populating each grid cell with the mean, standard deviations and counts of variables, the new approach applied by JPL is to hierarchically reduce the data in a way that preserves distributional characteristics across subsets formed by stratifying on meaningful variables. A quantization algorithm (ECVQ) developed by JPL is applied in L3Q data to derive small collections of representative vectors and their associated weights [1]. The representatives and their weights can be thought of as comparing a discrete probability distribution which can be used as proxies for the original data in subsequent analyses.

The reduction method applied in this data set is ECVQ, Entropy-constrained

Vector Quantization. ECVQ was originally introduced in order to estimate distortion rate functions. It is an iterative descent algorithm that minimizes

$$L(\alpha_K, \lambda) = \frac{1}{N} \sum_{n=1}^{N} \left[||x_n - \beta(\alpha_k(x_n))||^2 + \lambda(-log \frac{N_{\alpha K(x_n)}}{N}) \right]$$

over $\alpha_k(\cdot)$. The subscript K in $\alpha_k(\cdot)$ makes explicit that there are K possible groups to which x_n can be assigned, $\beta(\cdot)$ is the centroid function, which is the expected value in this study, and λ expresses the value of one unit of entropy reduction relative to a one unit rise in distortion [1]. The above equation is the Lagrangian formulation of a constrained optimization problem that minimizes distortion subject to a constraint on entropy.

Using ECVQ for data reduction requires selecting values for K and λ. A complete description of how to select K for the AIRS application is given in [2]. For the remainder of this discussion consider K to be some appropriate value that is much less than N, and the maximum value of K is 100 in all the data sets in use. The histograms of the cluster number in each grid box of the four months data sets are depicted in **Figure 2.3**. As shown in the histograms, it is noticeable that the distributions of the cluster number are strongly right skewed. Most of the grid boxes have the cluster number less than 40, while only around 20 grid boxes have the cluster number greater than 40.

The maps in **Figure 2.4** show us how the cluster number distributed geographically. By looking at this map, one can find the area with no observations, thus we know where to impute missing values. The grid cells with more than 40 clusters are forced to have 40 clusters in these maps to gain a better color scale, because less than 1% of the grid boxes have more than 40 clusters and they are all considered as outliers.

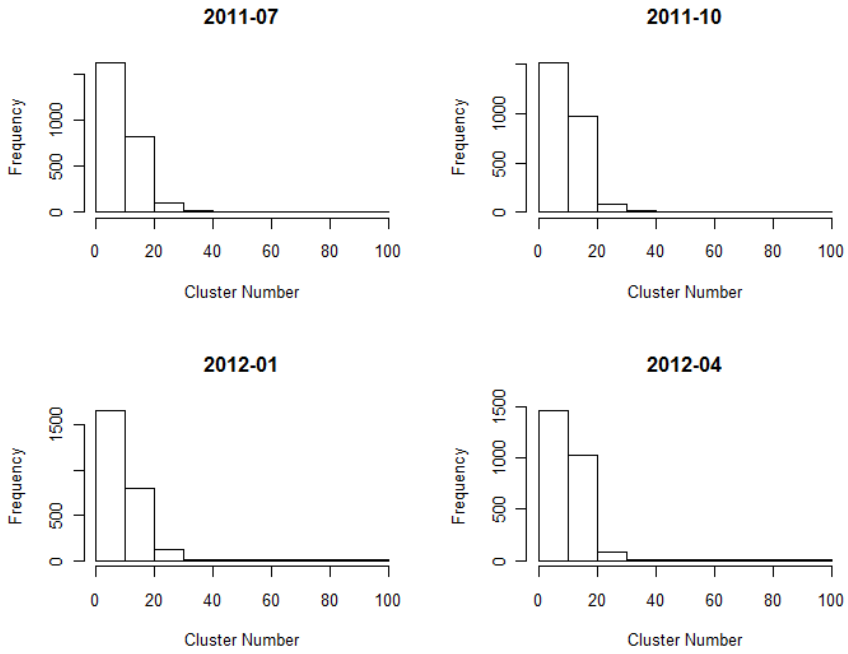

Figure 2.3: Cluster Number Histograms of Four Months

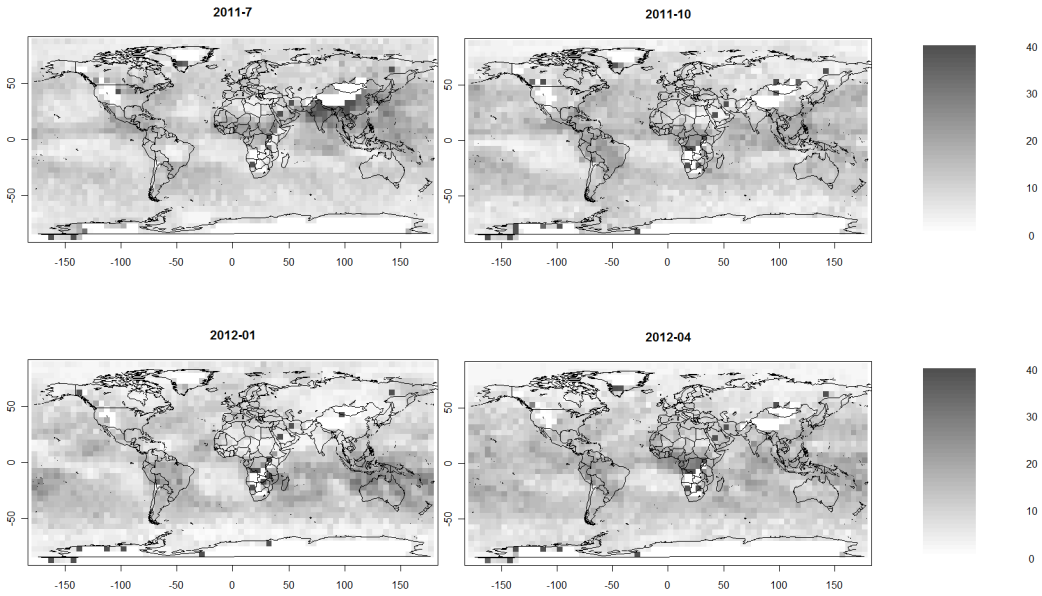

Figure 2.4: Cluster Number Maps of Four Months

CHAPTER 3

Geospatial Analysis

In this chapter, I briefly introduce several geospatial analysis techniques: variograms, kriging and spatial regression. These techniques are used in the applications of L3Q data in latter chapter. This chapter helps reader review some necessary techniques, hence one could understand the latter analysis better.

3.1 Variogram

In a geospatial lattice data set, there is a certain value in each pixel. For that, the latter can be treated as a geostatistical variable, $Z(s)$, used to establish variograms which in turn give us a better idea of the relation between different pixels in the data.

Suppose

$$2\gamma(h) = Var\left[z(s+h) - z(s)\right] \tag{3.1}$$

Typically, $\gamma(h)$ is called semi-variogram, and it increases with increasing $|h|$. The classical estimator of the variogram is:

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{N(h)}\left[z(s_i) - z(s_j)\right]^2, \tag{3.2}$$

where the sum is over all pairs of location, such that $s_i - s_j = h$. $N(h)$ is the set of all pairs, and $|N(h)|$ is the number of observed pairs.

The quantity $2\gamma(h)$, which is a function only of the increment $s_i - s_j$, is known as a variogram [4]. It is very crucial in geostatistics. The variogram symbolizes the squared difference between variables that are lag h-apart and varies in a way that depends only on the size of h. Subsequently, $\gamma(h)$ is known as a semivariogram. The variogram can be treated as a parameter of the random process $Z(\cdot)$.

Various variogram models are presented in [5]. Consider three basic isotropic models given here in terms of the semi-variogram: spherical, exponential, rational quadratic.

The spherical model is,

$$\hat{\gamma}(h) = \begin{cases} 0, & \text{for} \quad h = 0 \\ c_0 + c_s[\frac{3|h|}{2a_s} - \frac{|h|^3}{2a_s^3}], & \text{for} \quad 0 < |h| < a_s \\ c_0 + c_s, & \text{for} \quad |h| > a_s \end{cases} \tag{3.3}$$

$\theta = (c_0, c_s, a_s)'$, where $c_0 \geq 0, \quad c_s \geq 0, \quad a_s \geq 0.$

The exponential model is,

$$\hat{\gamma}(h) = \begin{cases} 0, & \text{for} \quad h = 0 \\ c_0 + c_e\{1 - exp(-||h||/a_e)\}, & \text{for} \quad |h| \neq 0 \end{cases} \tag{3.4}$$

$\theta = (c_0, c_e, a_e)'$, where $c_0 \geq 0, \quad c_e \geq 0, \quad a_e \geq 0.$

The rational quadratic model is,

$$\hat{\gamma}(h) = \begin{cases} 0, & \text{for} \quad h = 0 \\ c_0 + c_r||h||^2/(1 + ||h||^2/a_r), & \text{for} \quad |h| \neq 0 \end{cases} \tag{3.5}$$

$\theta = (c_0, c_r, a_r)'$, where $c_0 \geq 0, \quad c_r \geq 0, \quad a_r \geq 0.$

Based on the variogram, we would be able to keep track of the behaviors and relations between pairs of data in the remote sensing lattice as a function of separation distance.

3.2 Kriging

Kriging is a group of geostatistical techniques to interpolate the value of a random field at an unobserved location from observations of its value at nearby locations [6]. It belongs to the family of linear least squares estimation algorithms. The aim of kriging is to estimate the value of an unknown real-valued function, f, at a point, x^*, given the values of the function at some other points, $x_1, \ldots, x_n$. A kriging estimator is said to be linear because the predicted value $\hat{f}(x^*)$ is a linear combination that may be written as

$$\hat{f}(x^*) = \sum_{i=1}^{n} \lambda_i(x^*) f(x_i), \tag{3.6}$$

where $\sum_{i=1}^{n} \lambda_i(x^*) = 1$.

The weights , $\lambda_i(x^*)$, are solutions of a system of linear equations which is obtained by minimizing MSE:

$$MSE = \left[F(x) - \sum_{i=1}^{n} \lambda_i(x^*) f(x_i) \right]^2 \tag{3.7}$$

$$= [\epsilon(x)]^2, \tag{3.8}$$

then, apply Lagrange multiplier method of constrained optimization to minimize the function,

$$F(\lambda, \mu) = MSE - 2\mu(\sum \lambda_i - 1), \tag{3.9}$$

by taking partial derivatives with respect to λ and μ and equating them to 0, where μ is the Lagrange multiplier [6].

There are two basic types of kriging, ordinary kriging and universal kriging[5]. In ordinary kriging, we assume $f(x) = \mu + \delta(x)$, where μ is constant, and δ is some intrinsical statistics process with mean 0 and some known variogram $2\gamma(h)$. If μ is known, it becomes to simple kriging.

If we assume μ is not constant, but changes with x, it becomes the universal kriging. Now create some functions, $g_1(x), g_2(x) \ldots, g_p(x)$. Suppose $\mu(x) = \sum_{j=1}^{p} \beta_j g_j(x)$, where β_j needs to be estimated. Therefore,

$$f(x) = \sum_{j=1}^{p} \beta_j g_j(x) + \delta(x) \tag{3.10}$$

where $\delta(x)$ is some intrinsic statistics process with mean 0 and some known variogram $2\gamma(h)$. We still need to minimize MSE to find the weights λ_i

3.3 Spatial Regression

When a value observed in one location depends on the values observed at neighboring locations, there is spatial dependence. Spatial regression models indicate relationships between variables and their neighboring values. The model includes explanatory variables as the values of error terms, x and y values in surrounding regions [14].

Spatial regression method captures spatial dependence in regression analysis. It provides information on spatial relationships among the variables involved and the spatial trend. Therefore, based on the spatial regression, we would be able to find the spatial trend of the climate distributional characteristics.

CHAPTER 4

Mixture Distributions

In climate models, it is often assumed that temperature and water vapor have a fixed Gaussian distribution, when in fact these distributions can be more variable over space and time. In this Chapter, I develop an algorithm to construct the mixture distributions from L3Q data, then use kernel smoothing to smooth the distributions in each grid box. The mixture distribution is a more realistic simulation of the atmosphere. In this thesis, I use skewness as the distributional characteristics and perform analysis on it. I obtain the skewness of temperature and water vapor distributions and populate each grid box with it. Hence, I obtain a skewness geospatial lattice data set in this chapter, which is used for geospatial analysis in the latter chapter.

4.1 Kernel Smoothing

In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. By applying KDE, we would uncover structural features in the AIRS Level 3 data while a parametric approach might not be available to achieve. Let $(x_1, x_2, \ldots, x_n)$ be an iid sample drawn from some distribution with an unknown density f. Its kernel density estimator is,

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4.1}$$

where K(·) is the kernel, a symmetric but not necessarily positive function that integrates to one , and h $>$ 0 is a smoothing parameter called the bandwidth. A kernel with subscript h is called the scaled kernel and defined as $K_h(x) = 1/h \cdot K(x/h)$ [7]. Intuitively one wants to choose h as small as the data allow, however there is always a trade-off between the bias of the estimator and its variance. A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, normal, and others. Due to its convenient mathematical properties, the normal kernel is often used K(x) = $\phi(x)$, where ϕ is the standard normal density function.

Kernel density estimates are closely related to histograms, but can be endowed with properties such as smoothness or continuity by using a suitable kernel. For the histogram, first the horizontal axis is divided into sub-intervals or bins which cover the range of the data. Whenever a data point falls inside this interval, we place a box of a height. If more than one data point falls inside the same bin, we stack the boxes on top of each other. For the kernel density estimate, we place a normal kernel on each of the data points x_i. The kernels are summed to make the kernel density estimate. The smoothness of the kernel density estimate is evident compared to the discreteness of the histogram, as kernel density estimates converge faster to the true underlying density for continuous random variables [7].

4.2 Descriptive Statistics

In this study, we focus on the first 22 components of the physical values in L3Q data-11 atmospheric temperatures and 11 water vapor mass-mixing ratios corresponding to the bottom 11 pressure levels, which are included in the first 22

columns of each data matrix. With the weight of each cluster recorded in the 36th column of the matrix, we could find the weighted mean and standard deviation across all the pressure levels in each grid box. These statistics give us an basic idea about the temperature and water vapor distribution globally.

Denote each element in the first 22 columns of the matrix as x_{ij}, weight of each cluster as w_i, cluster number as n, altitude level as m, where i indicates row number and j indicates column number, $i = 1, 2, \dots$, $j = 1, 2, \dots, 11$. Then the mean μ of all altitudes is calculated as ,

$$\mu = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} w_i \cdot x_{ij}}{m \sum_{i=1}^{n} w_i} \tag{4.2}$$

and the standard deviation σ is calculated as,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \mu)^2 \cdot w_i}{m \sum_{i=1}^{n} w_i - 1}} \tag{4.3}$$

By applying these formulas, we could find descriptive statistics considering the weights of each data point, and therefore we obtain the descriptive statistics that are close to the reality.

Figure 4.1 depicts the global temperature in four seasons, each represents a month. The temperature is in the unit of Kelvin. It is obtained by filling each grid box with the cell aggregate mean. The result summarizes temperature weighted average across 11 different pressure levels in each grid box. The value is much less than the ground temperature average, since the temperature is very low at the altitude far from surface. The plots ideally fit our intuition about the temperature pattern for different seasons. **Figure 4.2** plots the standard deviation of temperature in each grid box. It measures the deviation of all the data points in one grid box. These two figures give us a view of global temperature.

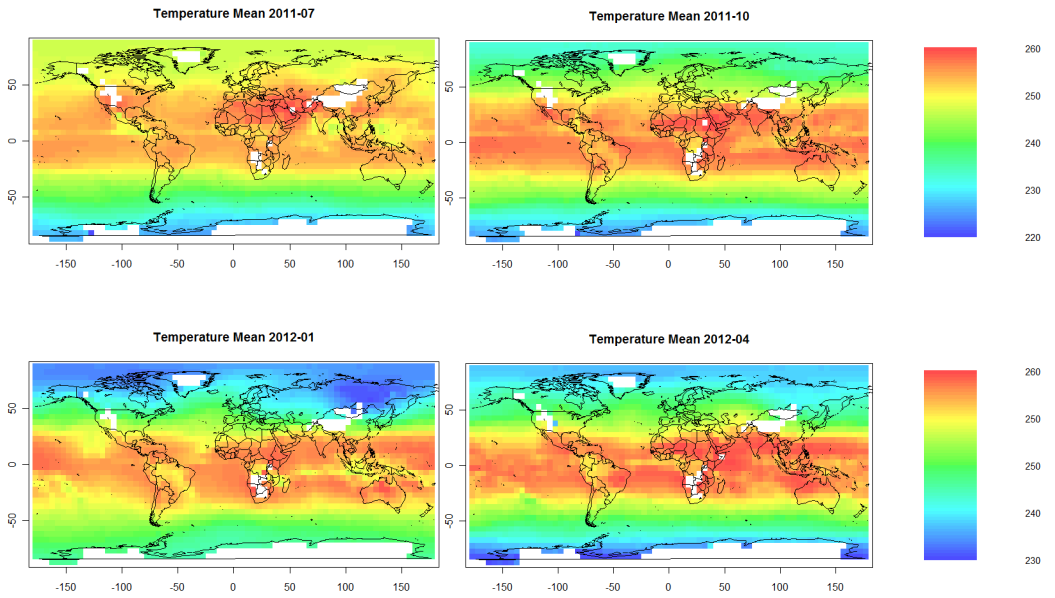

Figure 4.1: Aggregate Temperature Mean for Each Month

4.3 Algorithm

The L3Q product presents the advantage of containing more distribution information of the climate. To fully utilize it, a mixture distribution needs to be constructed. A sampling method could help achieve this. First assume each cluster as a normal distribution, thus the mixture distribution in each grid cell is formed by a mixture of n normal distributions, where n = cluster number. ECVQ gives us the weight of each cluster, thus each cluster is selected at the chance of $p_i = w_i / \sum_{i=1}^{n} w_i$. After the cluster selected, draw a random sample from $N(\mu, \sigma)$, where μ is the mean at one specific pressure level, and σ is calculated from the sum of square error at column 37. Repeat the above process N times, and compile all of the samples as one histogram. Perform the kernel smoothing on the histogram, then obtain the mixture distribution of temperature or water vapor at one specific pressure level.

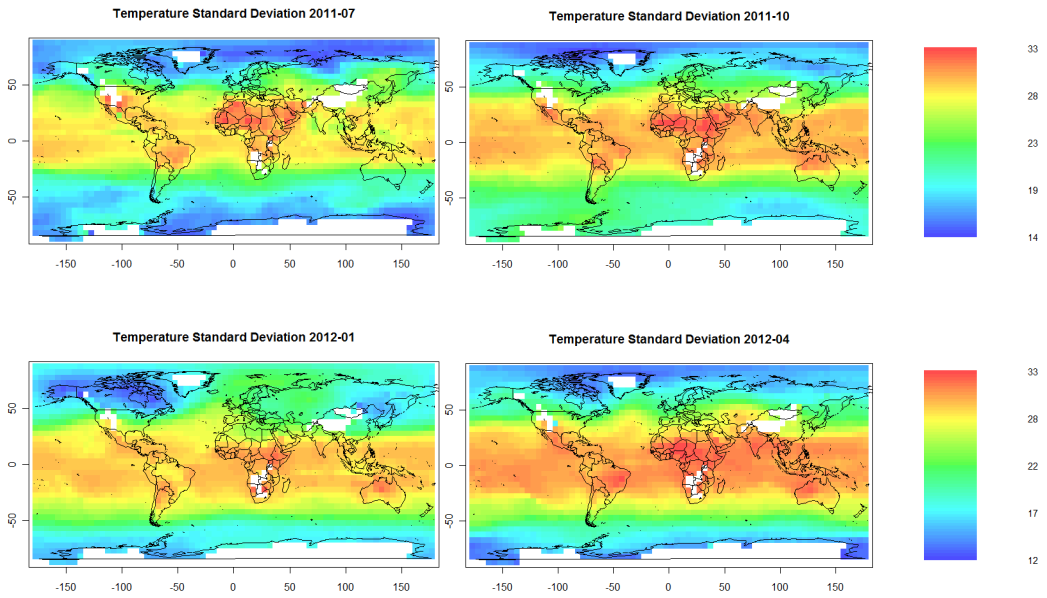

Figure 4.2: Aggregate Temperature Standard Deviation for Each Month

For each cluster in a grid cell, we assume it is normally distributed with the mean as the physical value and the standard deviation as the value from the sum square error column. L3Q included the sum square error in each cluster, but it is not a good representation of standard deviation for each component in the cluster, because it is the sum of all the mean square error of all 32 physical value components. Hence, we need to divide it by 32 and take the square root of it to get a reasonable value as the standard deviation.

Once the standard deviation is calculated, we could start the sampling procedure. Suppose that there are n clusters in the grid-box. Denote the mean, standard deviation and weight as μ_{ij}, σ_i w_i (i=1, 2, 3,...n; j=1,2..22), respectively. The weighted mixture distribution for each pressure level is attained by two approaches, and they are shown as follows,

Method I

1. Generate samples from N(μ_{ij}, σ_i) w_i times;

2. Repeat step 1 for all the n clusters, and store all the samples, N of them (N = $\sum_{i=1}^{n} w_i$);

3. Compile all the samples to generate histogram and find skewness γ_j;

4. Conduct kernel smoothing on the histogram;

5. Repeat the above steps 500 times, and find the mean value and standard deviation for all the $\gamma_j s$, check the normality of $\gamma_j s$.

Method II

1. Assign probability $p_i = w_i / \sum_{i=1}^{n} w_i$ to each cluster;

2. Select cluster with probability p_i with replacement, $\sum_{i=1}^{n} w_i$ times;

3. Within each selection, generate a sample from N(μ_{ij}, σ_i);

4. Compile all the samples to generate histogram and find skewness γ_j;

5. Conduct kernel smoothing on the histogram;

6. Repeat the above steps 500 times, and find the mean value and standard deviation for all the $\gamma_j s$, check the normality of $\gamma_j s$.

The two methods above share the same idea but with different approach. The difference is that Method II assigns probability to each cluster while Method I skips this step. To check the performance of the algorithm, use both of the methods above to form mixture distributions of temperature at the pressure level 1000 hPa (ground level). In the data set of 2012-01, random pick a grid box. And I obtain the grid box with coordinate (18, 60). The coordinate means that the grid box locates at the 18th latitude row and 60th longitude column. It is the $5° \times 5°$grid box with the range of latitude ($-5° \sim 0°$) and longitude ($115° \sim 120°$).

The histograms in **Figure 4.3** depict the mixture distribution, and the green line is the kernel smoothing curve. The smoothing used a rule-of-thumb for choos-

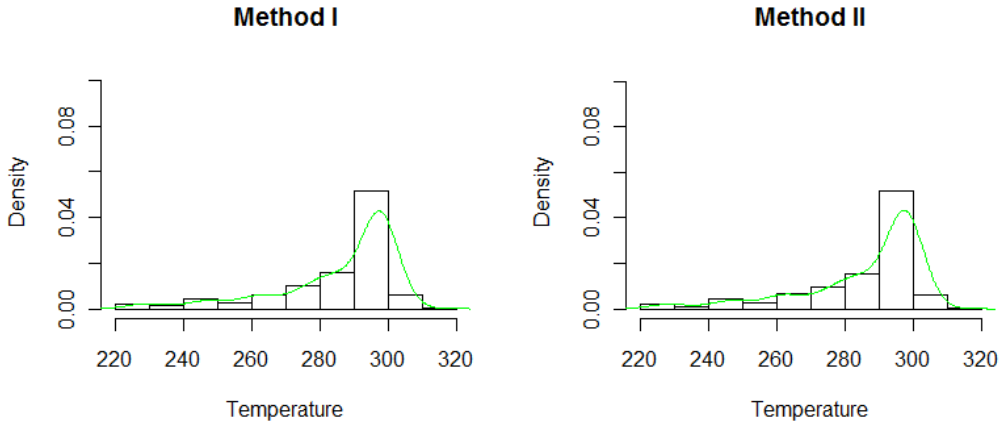

Figure 4.3: Mixture Distribution of Temperature at Pressure Level 1000hPa

ing the bandwidth of a Gaussian kernel density estimator. It defaults to 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power [7].

A strong left skew is discovered from Figure 3.1. The mixture distribution is no longer a normal distribution. It implies a strongly skewed distribution. It is much closer to the real situation than just assuming normal distribution. Skewness is an important feature of the distribution, thus it is necessary to study more on it.

4.4 Skewness

The skewness of a distribution is very important in climate science, because the tail in the distribution of temperature and water vapor control the formation of cloudiness. It informs the real behavior of temperature and water vapor distributions in the atmosphere [15].

In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. The skewness value can be positive or negative, or even undefined. Qualitatively, a negative skew indicates that the tail on the left side of the probability density function is longer than the right side and the bulk of the values lie to the right of the mean. A positive skew indicates that the tail on the right side is longer than the left side and the bulk of the values lie to the left of the mean [8]. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically but not necessarily implying a symmetric distribution. If the distribution is symmetric then the mean is equal to the median and the distribution is close to zero skewness. However, that the converse is not true in general. Zero skewness does not imply that the mean is equal to the median. Consider the distribution on **Figure 4.4**. They provide a visual means for determining which of the two kinds of skewness a distribution has.

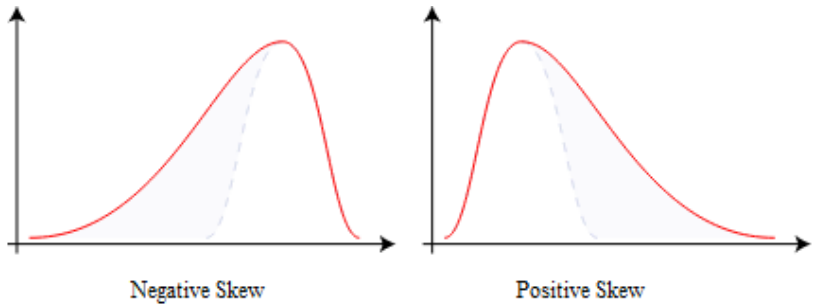

Figure 4.4: Skewness Demonstration

The skewness of a random variable X is the third standardized moment, denoted γ and defined as

$$\gamma = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right] = \frac{E[x^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3} = \frac{\mu_3}{\sigma^3} \tag{4.4}$$

where μ_3 is the third moment about the mean μ, σ is the standard deviation.

For a sample of n values the sample skewness is

$$g = \frac{m_3}{m_2^{3/2}} \frac{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} \tag{4.5}$$

where $\bar{x}$ is the sample mean, m_3 is the sample third central moment, and m_2 is the sample variance [8].

Now we need to acquire the skewness from the mixture distribution I built up. Repeat the sampling methods 500 times, and record the skewness. Then we can compare the performance of the two sampling approaches.

	μ	σ	System Time
Method 1	-1.558336	0.007736046	3.36 s
Method 2	-1.555197	0.04222394	40.46 s

Table 4.1: Summery Statistics of Skewness from 500 Iterations

The two approaches give a very close mean skewness, but the standard deviation of Method I is smaller than that of Method II. In addition, Method II takes much longer time. It is more efficient to use Method I. The following plots examine the normality check of 500 skewness simulations.

The skewness from both of these two methods are close to normal distribution, which means the algorithm complies with the central limit theorem, and it gives out a robust result. In fact, the distribution on left has a skewness of -0.1533611, and the right has a skewness of -0.1940276. I will use the first method to generate all the skewness, since it is much faster and gives a more concentrate output.

Figure 4.6 depicts temperature distribution at pressure level 1000 hPa in 9 randomly picked grid boxes around equator. As shown in the plots, the tempera-

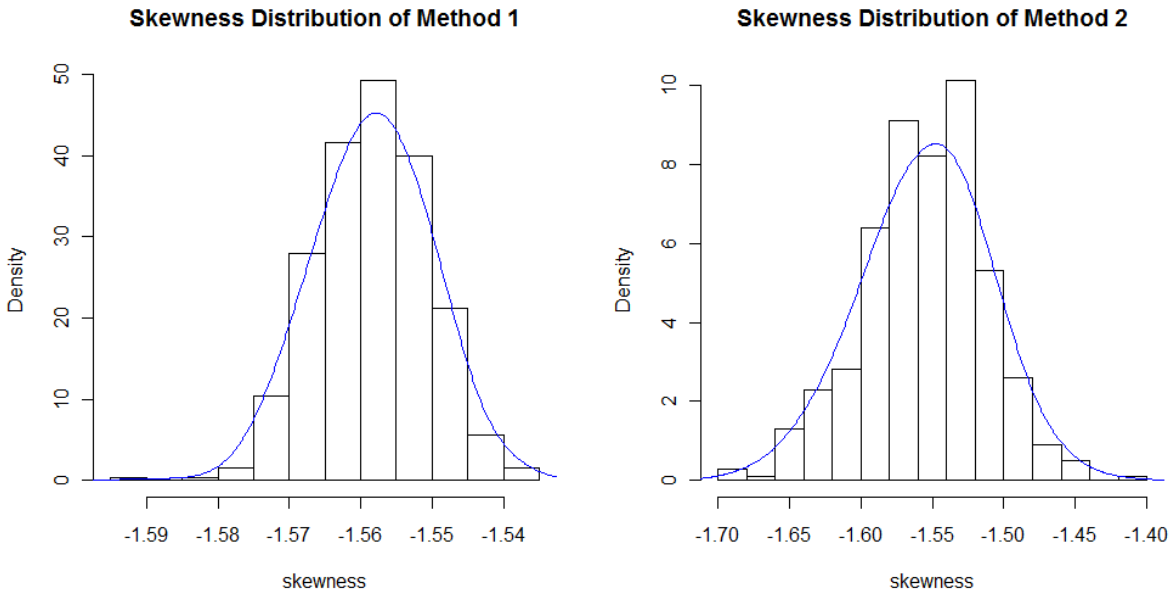

Figure 4.5: Skewness in 500 Simulations

ture distributions are strongly left skewed, which may due to the reason of being located around the equator. The convection in the tropical tropopause region may affect the temperature distribution. It leads to a left skewed temperature distribution at the tropical area [8]. The skewness in other region could possibly be positive of negative. Hence, the curiosity becomes to explore the distribution skewness of the entire earth at each pressure level.

Figure 4.5 to **Figure 4.8** depict the global temperature skewness in 2011-07, 2011-10, 2012-01 and 2012-04, one month in each season. They all share the same pattern that mostly negative skewness are discovered in the region between latitude -45 ° and -45 °. As the pressure level decreases (altitude getting higher), the negative skew area shrinks, which means the skewness getting larger in the vertical direction from bottom to top at this region. It may worth some work to find the relationship between skewness and altitude at this region, since one could have an idea about how temperature distributions change as altitude varies

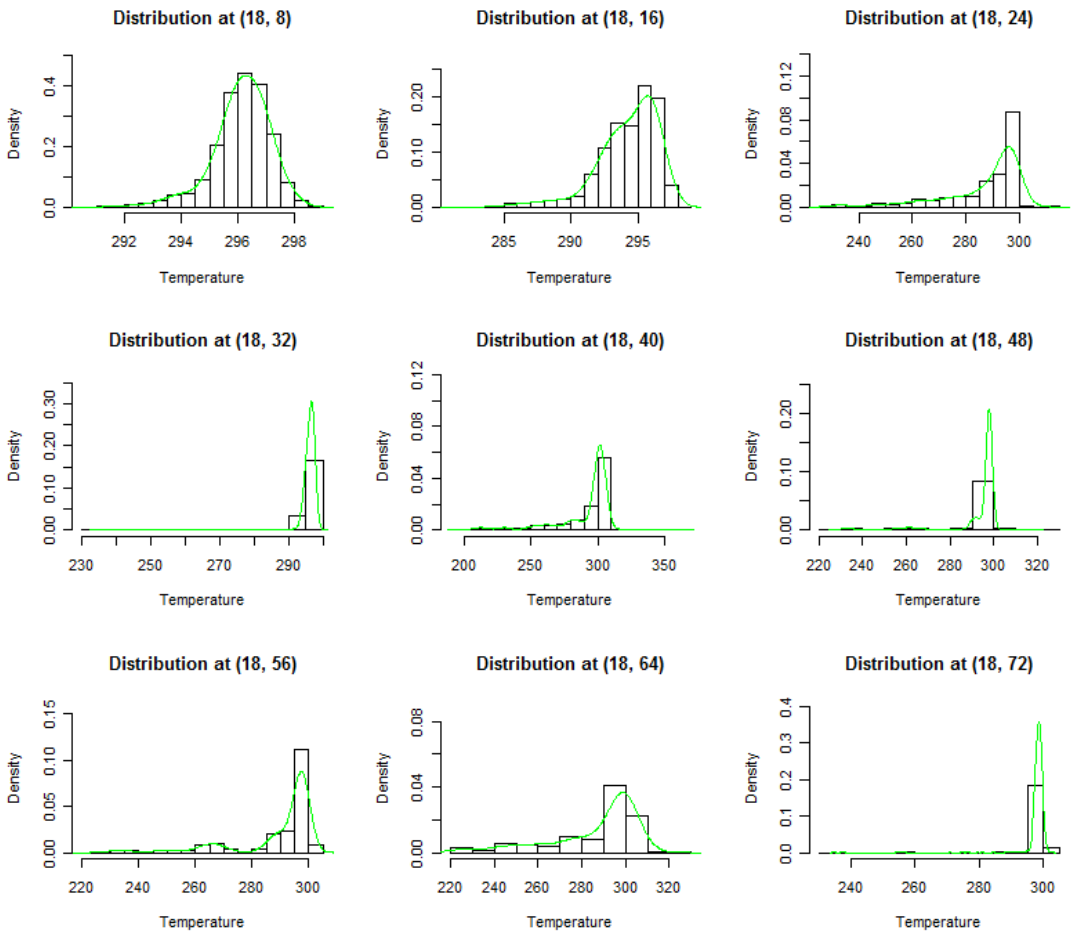

Figure 4.6: Temperature Distributions in 9 Different Cells around the Equator

and study how convection affects the temperature distribution in the tropical tropopause region. In addition, the negative skewness region tends to be larger and more dense in spring (2012-04), and more scattered in the summer (2011-07).

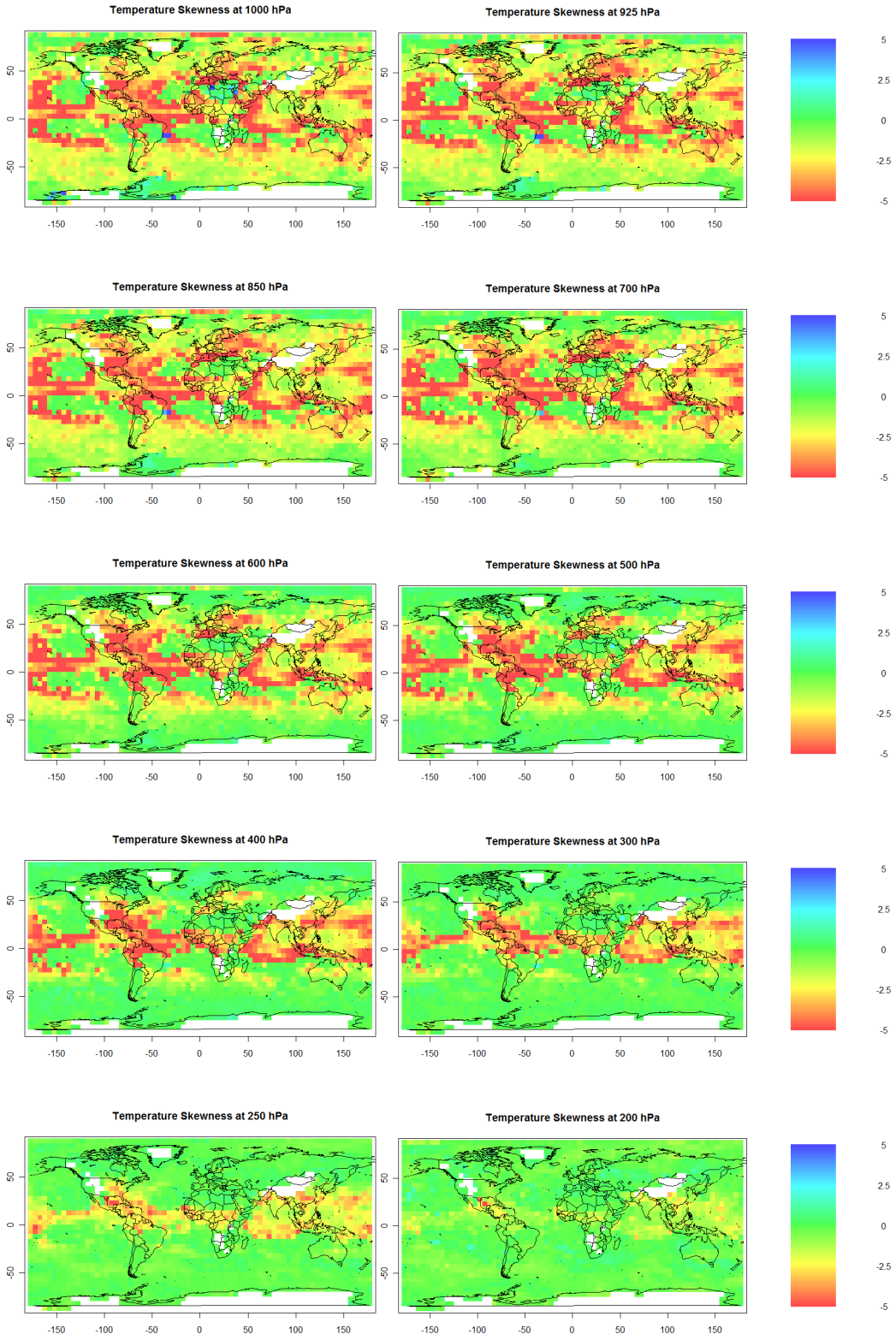

Figure 4.7: Temperature Distribution Skewness at Each Pressure Level, 2011-07

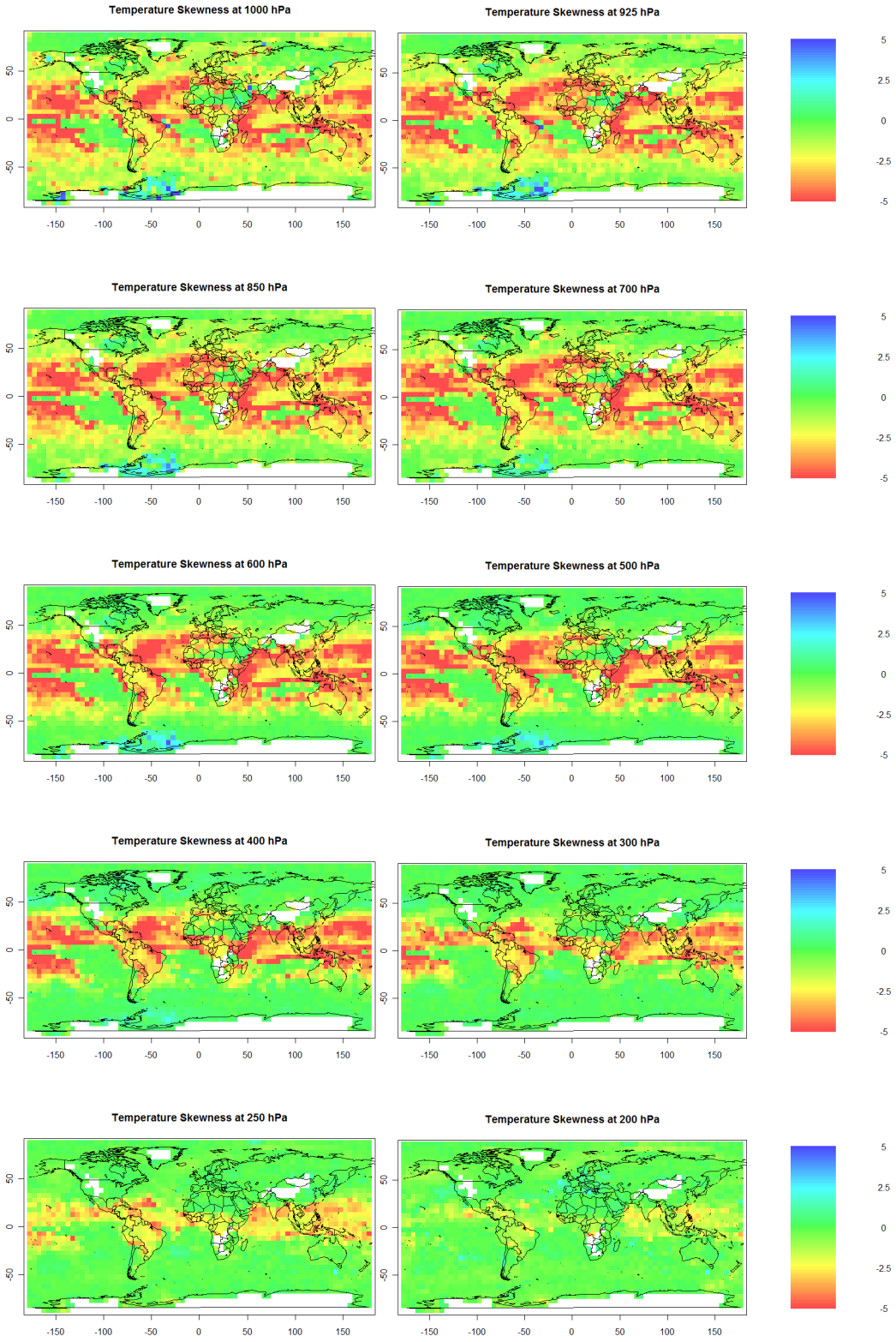

Figure 4.8: Temperature Distribution Skewness at Each Pressure Level, 2011-10

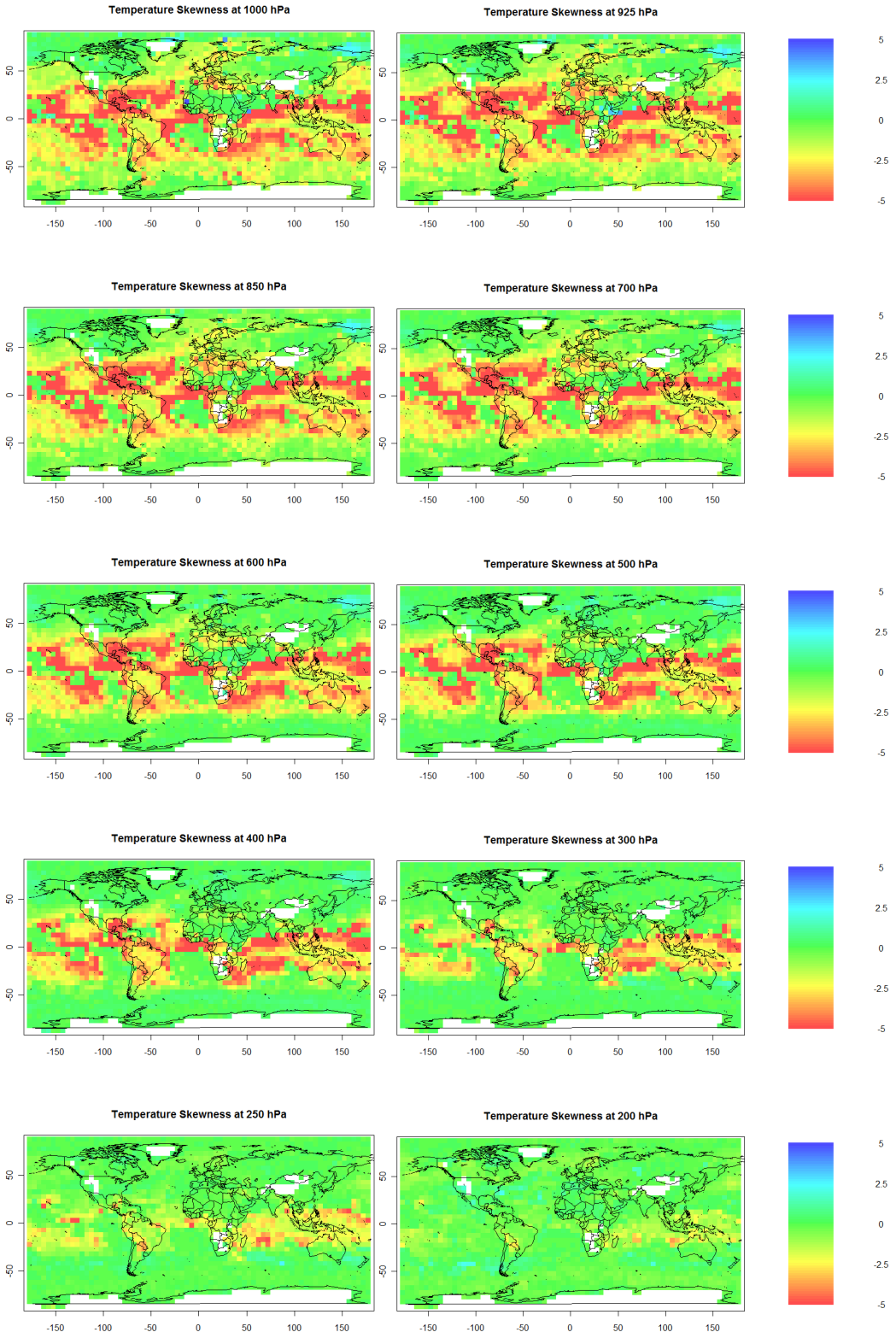

Figure 4.9: Temperature Distribution Skewness at Each Pressure Level, 2012-01

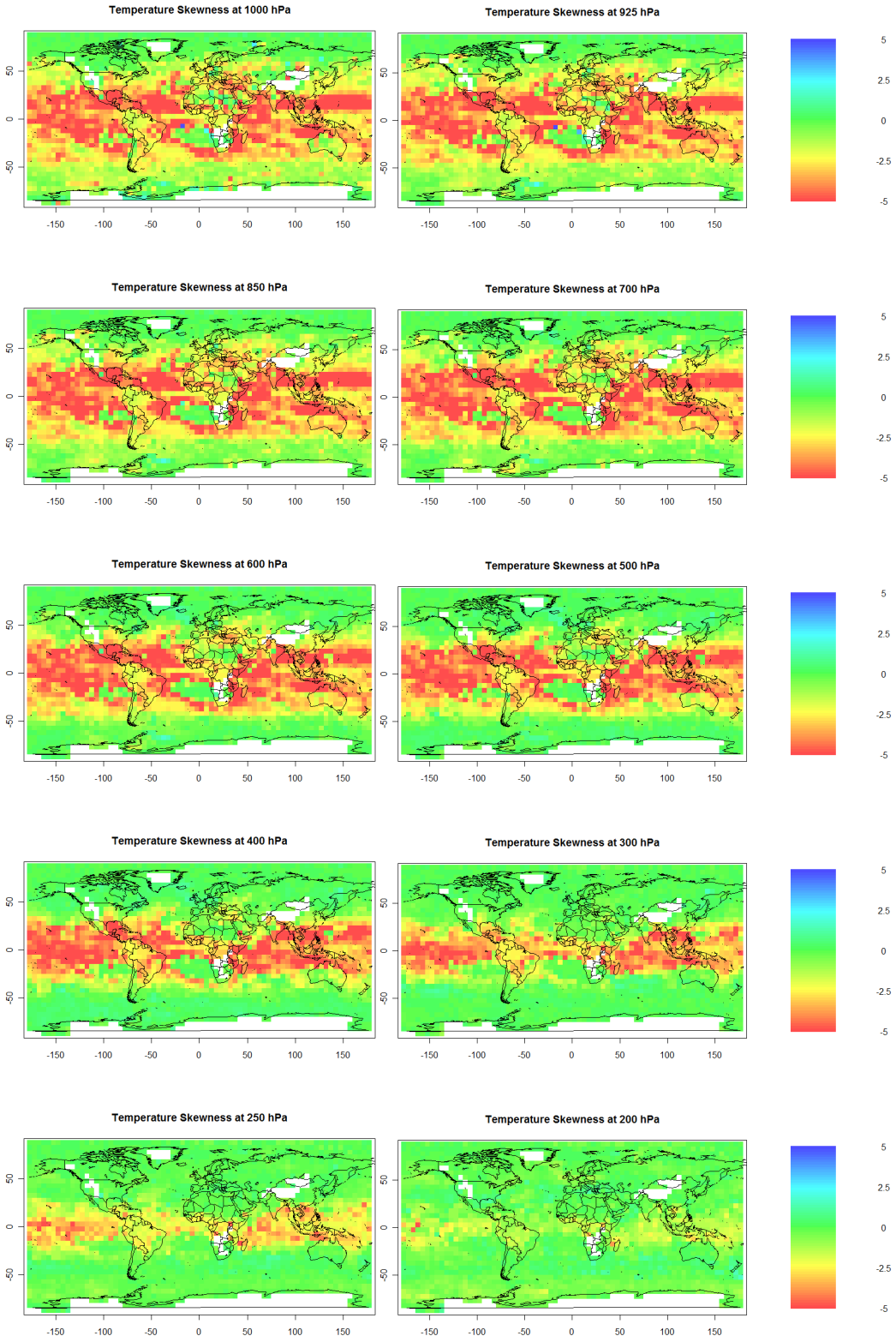

Figure 4.10: Temperature Distribution Skewness at Each Pressure Level, 2012-04

CHAPTER 5

Geospatial Lattice Analysis by Month

The skewness maps were summarized and shown in the previous chapter. In this chapter, I use skewness as lattice data to drive spatial analysis. Through the spatial analysis, one can understand how skewness correlates to each other spatially with different lags as well as having an idea about the spatial trend of skewness and imputing the missing skewness with kriging . Since the four months are in different seasons, I discuss each month one by one. The analysis is driven on the temperature distribution at the pressure level of 1000hPa.

5.1 Temperature Distribution 2011-07

As shown in **Figure 4.7**, area between latitude $-50°$ and $+50°$ possess strong negative skewness , and very few grid boxes possess positive skewness. Interestingly, the strong negative skewness constitutes several ring shapes with skewness close to zero at the center of the ring. The strongest skewness are observed on oceans. They are located within east Pacific and west Atlantic. To visualize the spatial trend of the skewness, it is helpful to look at the spatial regression.

The Spatial trend plot is shown in **Figure 5.1** on the next page. The spatial tend of skewness give us an idea on how distribution changes globally. The blue areas represent regions with positive skewness, and red color corresponds to negative skewness. The plot shows a decreasing trend diagonally from southeast to

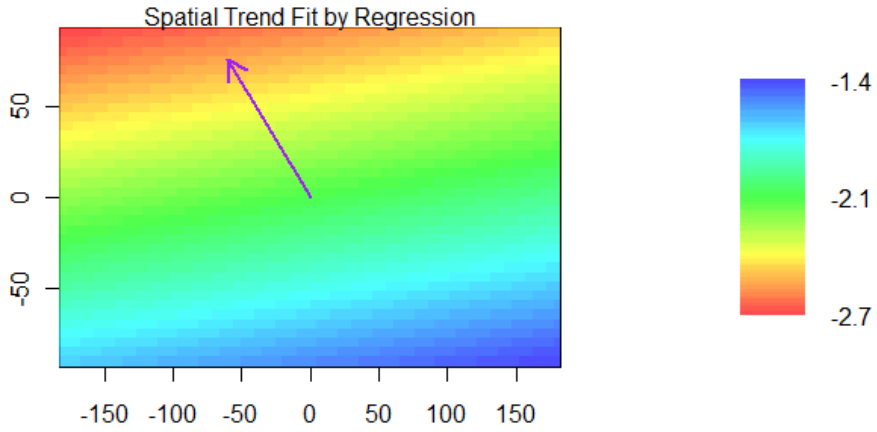

Figure 5.1: Spatial Trend of Temperature Skewness 2011-07

northwest. It means that the skewness tends to be more left skewed from the east of south hemisphere to the west of north hemisphere.

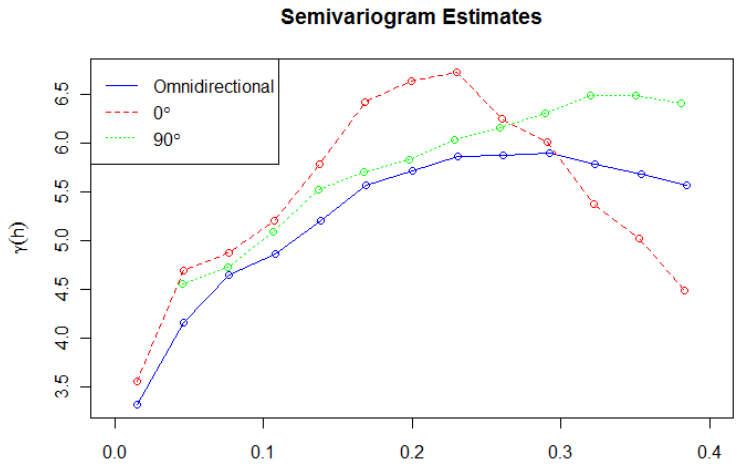

Figure 5.2: Semi-variogram Estimates 2011-07

To understand how the skewness spatially correlated to each other, we need to take a look at the variogram analysis. The semi-variogram plots shown in **Figure 5.2** depicts semivariogram estimates in N-S direction, E-W direction and omnidirection. The skewness at different pixels is correlated with each other up to relative distance of 0.25 units in the omnidirection, and up to 0.3 at the N-S direc-

tion. The semi-variogram in E-W direction acts quite differently. The parabolic behavior shows a highly regular spatial variability in E-W direction.

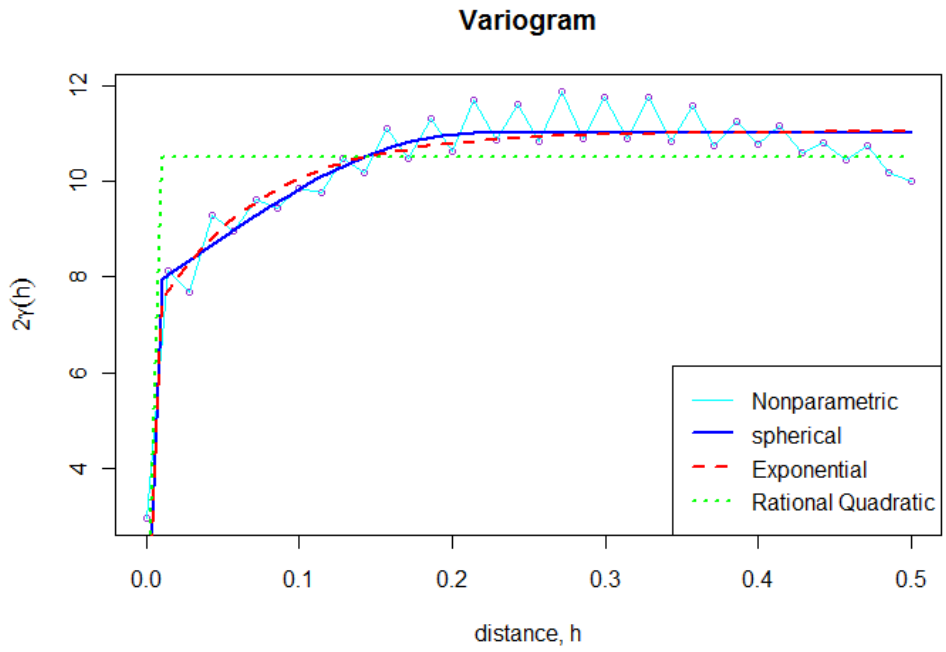

Figure 5.3: Variogram (Temperature Skewness at 1000hPa 2011-07)

As shown in **Figure 5.3**, it appears from the variogram that the skewness of temperature at pressure level 1000 hPa is somewhat correlated with one another up to 0.2 of relative distance of separation. After 0.25, the observations are independent of one another. The variogram plot based on the temperature skewness in each pixel shows that the variogram follows closely to the exponential model. In the case of the exponential model, the function to fit is:

$$\hat{\gamma}(h) = \begin{cases} 0, & \text{for} \quad h = 0 \\ c_0 + c_e\{1 - exp(-||h||/a_e)\}, & \text{for} \quad |h| \neq 0 \end{cases} \tag{5.1}$$

where $\theta = (c_0, c_e, a_e)'$, $c_0 = 6.96858$, $c_e = 4.070524$, $a_e = 0.07014682$. These values are generated from **optim()** function in R, which is a general-purpose opti-

mization based on NelderMead, quasi-Newton and conjugate-gradient algorithms.

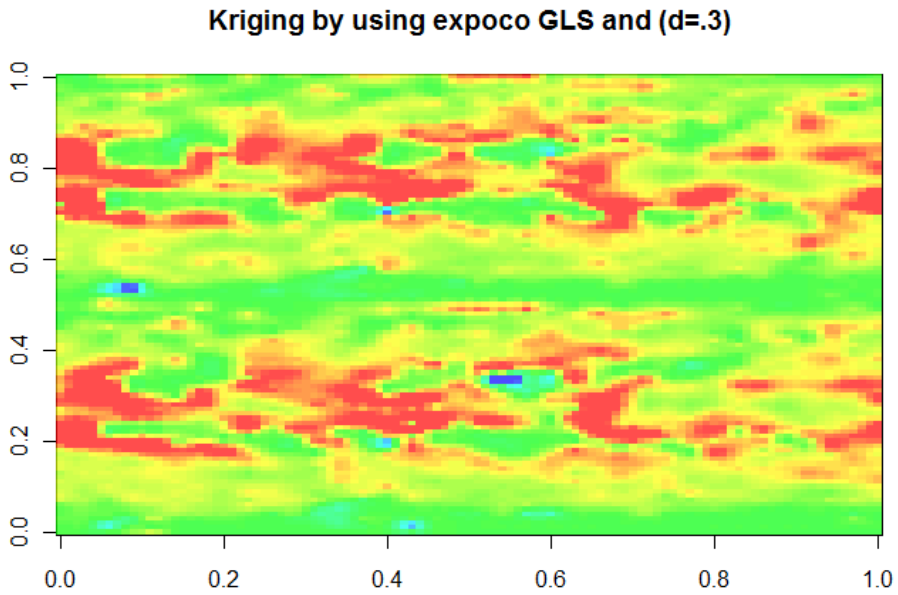

Figure 5.4: Evaluate Kriging surface 2011-07

The plot in **Figure 5.4** shows a kriging estimate. It uses exponential covariance model with d=0.3 and a degree of 4 polynomial surface to the data by general least square. With the kriging estimation, it fills out the blank area where the original data set has no observations. Therefore, we could estimate the distributional characteristic of temperature for the region without observations.

5.2 Temperature Distribution 2011-10

In **Figure 4.8**, it shows the temperature skewness at pressure level 1000hPa. As shown in the plot, the negative skewness area tends to be smaller than that of

July, and the strong negative skewness seems to have a larger density at some region, such as southeast Pacific Ocean. The entire negative skewness region also shifts down to the south hemisphere as well. However, the skewness at higher latitude still stays close to zero. One could also observe some positive skewness at Antarctica. The difference of the skewness between the two months is very likely caused by season change and may inspire the interest of climate scientists. By merely looking at the skewnss map on its own, it is hard to find the spatial trend and the correlation between the lattices. Hence, it would be helpful to look at the spatial regression plot as well as the semivariogram plots.

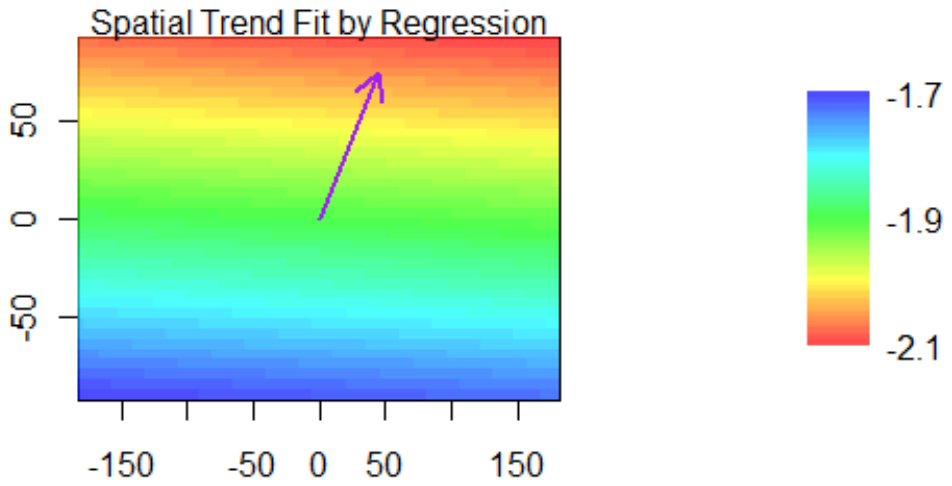

Figure 5.5: Spatial Trend of Temperature Skewness 2011-10

The Spatial Trend plot is shown above. The blue areas represent regions with positive skewness, and the red color corresponds to negative skewness. The plot shows a decreasing trend slightly diagonally from bottom to top. It means that the distribution tends to be more left skewed from the south hemisphere to the north hemisphere. The trend is different from that in the summer.

The semi-variogram plot in **Figure 5.6** depicts semivariogram estimates in ver-

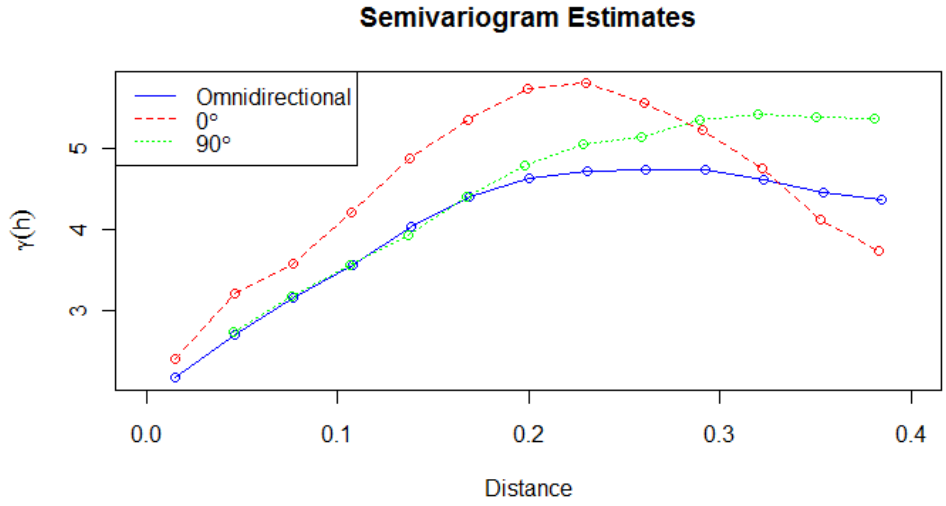

Figure 5.6: Semi-variogram Estimates 2011-10

tical direction, horizontal direction, and omnidirection. The skewness at different pixel is correlated with each other up to relative distance of 0.2 at all directions, and up to 0.3 at the N-S direction. the The semi-variogram in horizontal direction acts quite differently. The parabolic behavior of its characteristic shows a highly regular spatial variability in E-W direction. The semivariogram is similar to that of July with some minor difference on lags.

The variogram generated in **Figure 5.7** based on the temperature skewness in each pixel shows that the variogram of skewness follows closely to the spherical model. In the case of the spherical model, the function to fit is:

$$\hat{\gamma}(h) = \begin{cases} 0, & \text{for} \quad h = 0 \\ c_0 + c_s[\frac{3|h|}{2a_s} - \frac{|h|^3}{2a_s^3}], & \text{for} \quad 0 < |h| < a_s \\ c_0 + c_s, & \text{for} \quad |h| > a_s \end{cases} \tag{5.2}$$

where $\theta = (c_0, c_s, a_s)'$, $c_0 = 4.6263106$, $c_s = 4.6412260$, $a_s = 0.2231205$. It appears from the variogram that the skewness of temperature at pressure level 1000 hPa is somewhat correlated with one another up to 0.2 of relative distance of separation. After 0.2, the observations are independent of one another.

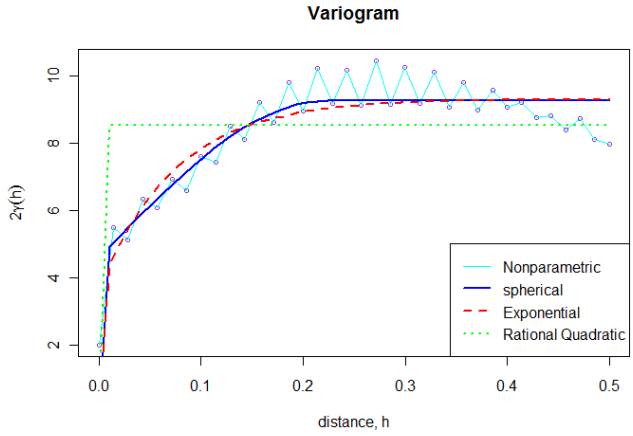

Figure 5.7: Variogram (2011-10 Temperature Skewness at 1000hPa)

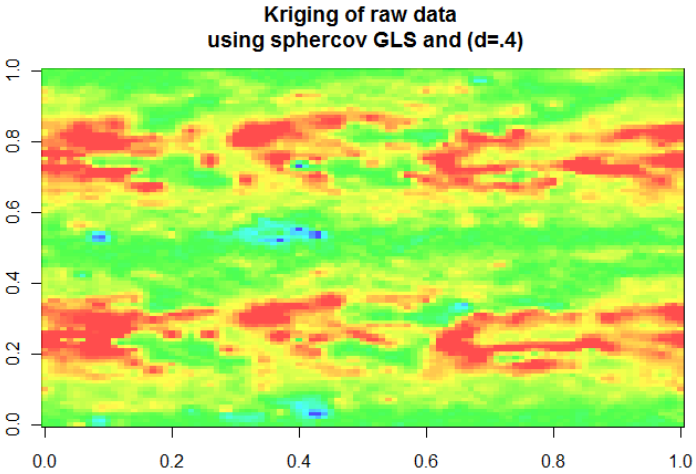

Figure 5.8: Evaluate Kriging surface 2011-10

The kriging estimation in **Figure 5.8** fills out the blank area where the original data set has no observations. We could estimate the distributional characteristic of temperature for the region with missing values.This kriging estimation uses exponential covariance model with d=0.4 and a degree of 4 polynomial surface to the data by general least square.

5.3 Temperature Distribution 2012-01

As shown in **Figure 4.9**, the negative skewness area keeps shrinking and shifting to the south in January. It actually has the smallest negative skewness area in January among the four months. We could find some positive skewness in Bering Strait, though not very strong. By merely looking at the skewnss map on its own, it is hard to find the spatial trend and the correlation between the lattices. Hence, it would be helpful to look at the spatial regression plot as well as the semivariogram plots.

.

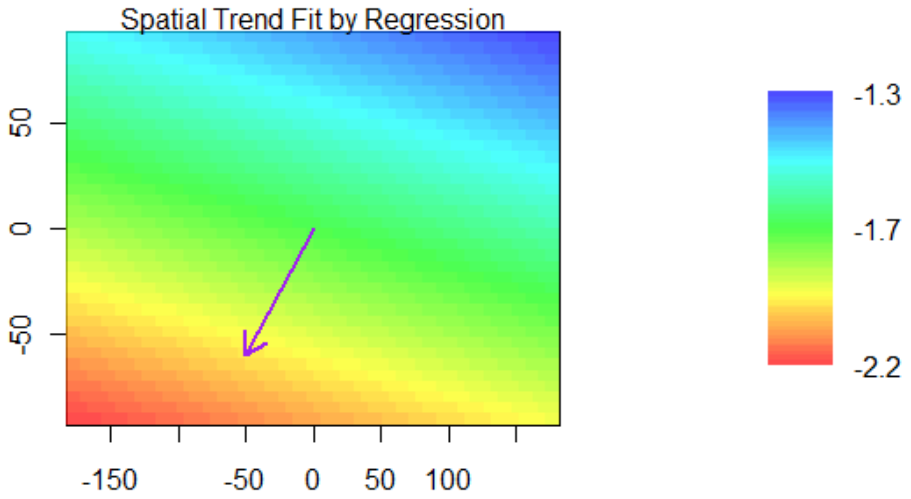

Figure 5.9: Spatial Trend of Temperature Skewness 2012-01

The Spatial Trend plot is in **Figure 5.9**. The blue areas represent regions with positive skewness, and red color corresponds to negative skewness. The plot shows a decreasing trend diagonally from northeast to southwest. It means that the distribution tends to be more left skewed from the right corner of north hemisphere to the left corner of north hemisphere. The trend in January is the opposite of that in July. This might be explained by the opposite seasons - summer and winter.

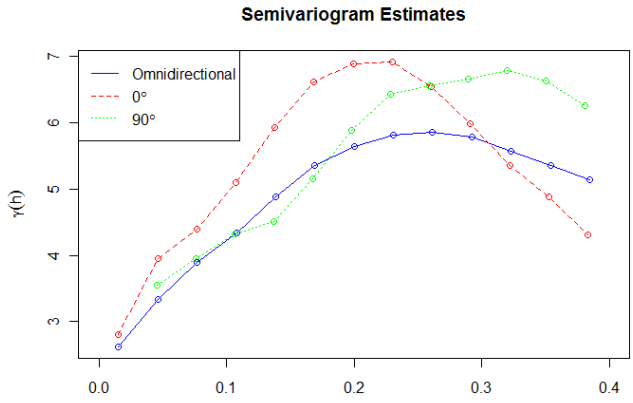

Figure 5.10: Semi-variogram Estimates 2012-01

To understand how the skewness correlated to each other, we need to refer the variogram analysis. The semi-variogram plot shown above depicts semivariogram estimates in N-S direction, E-W direction and omnidirection. The skewness at different pixels is correlated with each other up to relative distance of 0.25 units in the N-S direction. The semi-variogram in E-W direction and all directions have the parabolic shapes. The parabolic behavior shows a highly regular spatial variability in E-W direction and Omnidirection.

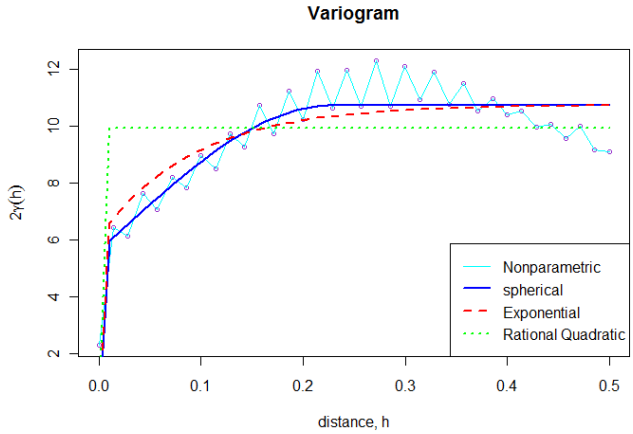

Figure 5.11: Variogram (Temperature Skewness at 1000hPa 2012-01)

In **Figure 5.11**, it appears from the variogram that the skewness is somewhat correlated with one another up to 0.2 of relative distance of separation. After 0.2, the observations are independent of one another. The variogram plot based on the temperature skewness in each pixel shows that the variogram follows closely to the spherical model, where $\theta = (c_0, c_s, a_s)'$, $c_0 = 5.6097624$, $c_s = 5.1258593$, $a_s = 0.2297672$.

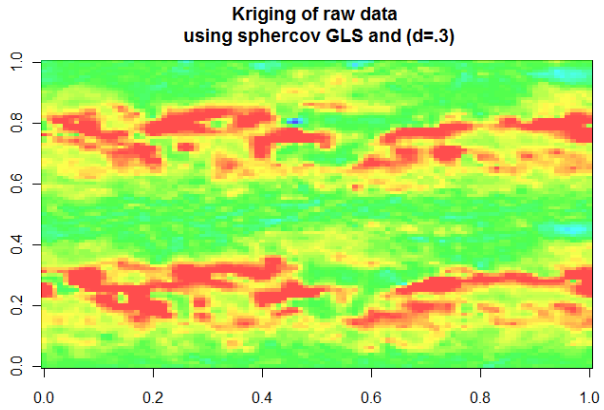

Figure 5.12: Evaluate Kriging surface 2012-01

The plot in **Figure 5.12** shows a kriging estimate by using spherical covariance model with d=0.3, and degree of 4 polynomial surface to the data by general least square. With the kriging estimation, it fills out the blank area where the original data set has no observations.

5.4 Temperature Distribution 2012-04

The temperature skewness map at pressure level 1000 hPa was given in **Figure 4.10**. As shown in that plot, it is noticeable that most of the region between latitude $-45°$ and $+45°$ have a temperature distribution strongly skewed to the left,

while the other region possess a distribution with neutral skewness. The negative skewness area in April is the largest among the four months. The region at south Atlantic between South America and Africa always has a neutral skewness, while its neighbor areas are all negatively skewed. By merely looking at the plot on its own, it is hard to find the spatial trend and the correlation between the lattices. Hence, it would be helpful to look at the spatial regression plot as well as the semivariogram plots.

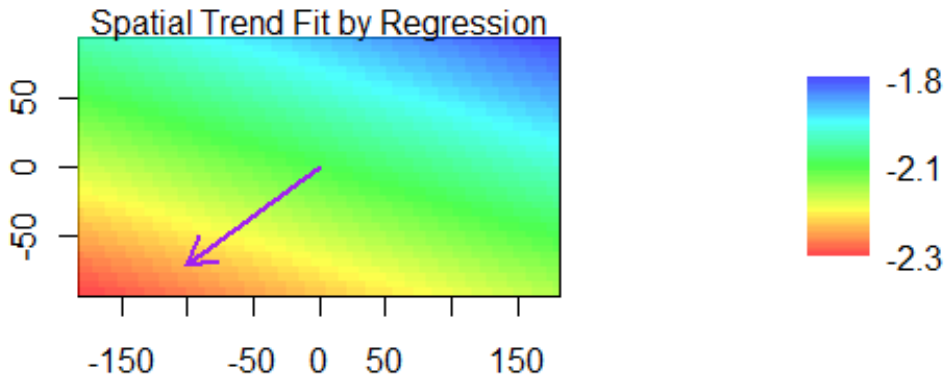

Figure 5.13: Spatial Trend of Temperature Skewness

The Spatial Trend plot is shown above. The blue area represents regions with positive skewness, and the red color corresponds to negative skewness. The plot shows a decreasing trend diagonally from top-right to bottom-left. At this point, one can decompose the spatial trend in each month. In the first half of the year-winter and spring, the trends both contain the direction component from north to south. And in the second half of the year-summer and autumn, they both contain the direction component from south to north.

The semi-variogram plot in **Figure 5.14** depicts semivariogram estimates in N-S direction, E-W direction and omnidirection. The skewness at different pixels is correlated with each other up to relative distance of 0.23 units in both N-S di-

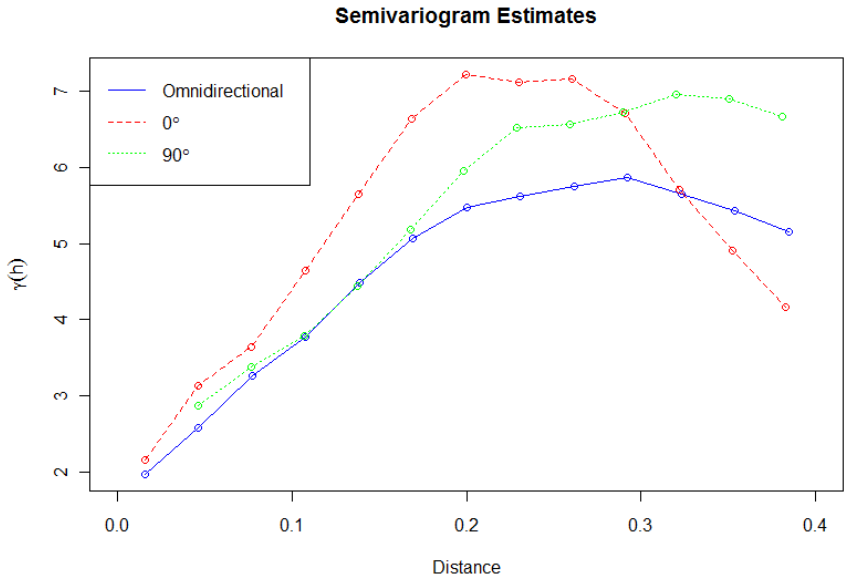

Figure 5.14: Semi-variogram Estimates 2012-04

rection and omnidirection. These two estimates have the logarithmic shape. The semi-variogram in E-W direction acts quite differently. The parabolic behavior shows a highly regular spatial variability in horizontal direction.

The variogram in **Figure 5.15** generated based on the temperature skewness in each pixel shows that the variogram of skewness follows closely to the spherical model. In the case of the spherical model, the function to fit is:

$$\hat{\gamma}(h) = \begin{cases} 0, & \text{for} \quad h = 0 \\ c_0 + c_s[\frac{3|h|}{2a_s} - \frac{|h|^3}{2a_s^3}], & \text{for} \quad 0 < |h| < a_s \\ c_0 + c_s, & \text{for} \quad |h| > a_s \end{cases} \tag{5.3}$$

$\theta = (c_0, c_s, a_s)'$, where $c_0 = 3.816408$, $c_s = 6.716823$, $a_s = 0.242199$. These values are generated from **optim()** function in R. It appears from the variogram that the skewness of temperature at pressure level 1000 hPa is somewhat correlated with one another up to 0.25 of relative distance of separation. After 0.25,

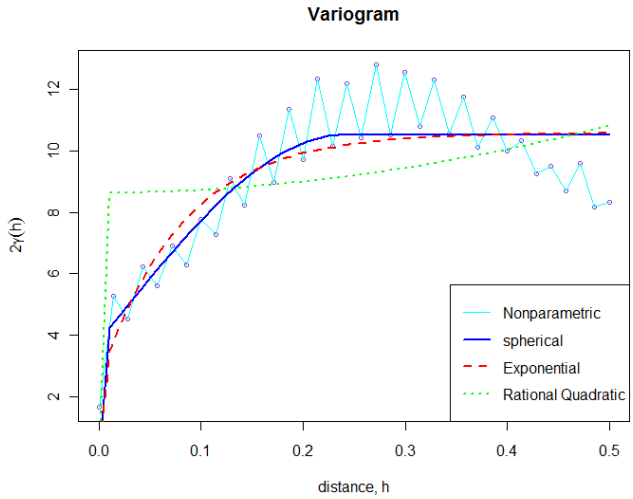

Figure 5.15: Variogram (2012-04 Temperature Skewness at 1000hPa)

the observations are independent of one another.

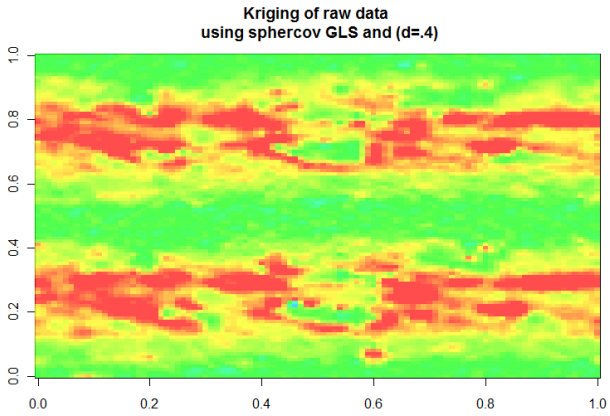

Figure 5.16: Evaluate Kriging Surface 2012-04

The plot in **Figure 5.16** shows a kriging estimate by using spherical covariance model with d=0.4, and degree of 4 polynomial surface to the data by general least square. With the kriging estimation, it fills out the blank area where the original data set has no observations.

5.5 Humidity: Water Vapor Mixing Ratio

In Chapter 3, the data matrix was formed with 37 dimensions as shown in **Figure 2.2**. The first 22 components are the 11 atmospheric temperatures and 11 water vapor mixing ratios corresponding to the bottom 11 pressure levels. Water vapor mixing ratio is defined as mass of water vapor per unit mass of dry air [10]. By following the same technique we did for temperature, use Method I to generate mixture distribution of water vapor and then acquire the distribution skewness of it. I study the water vapor mixing ratio at the lowest pressure level for the 4 months.

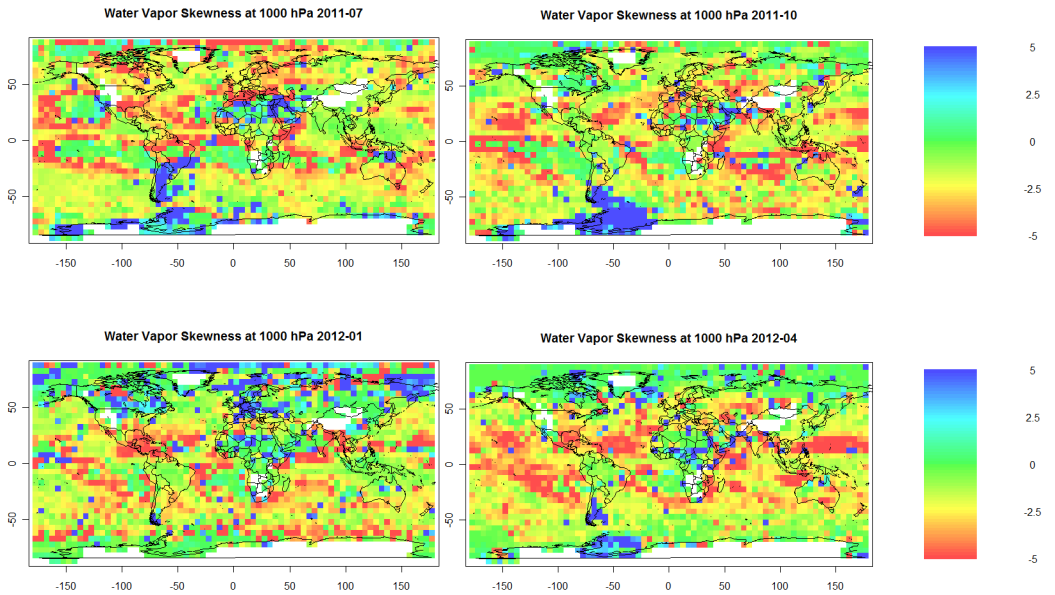

Figure 5.17: The skewness of water vapor distribution in 4 months

As shown in **Figure 5.17**, water vapor skewness is quite different from that of temperature. There is more noise on each of the four maps, especially in January.

However, the region (between latitude $-45°$ and $+45°$) containing strong negative skewness is consistent with that of temperature, and they somehow share the same pattern. Hence, it would be meaningful to test the relation between temperature skewness and water vapor skewness at this region. In addition, we observe a strong positive skewness at the region between South America and Antarctica in July, October and April. In January, most of the positive skewness is found in the north hemisphere, and it is clustered at Berlin Strait, Europe and Canada.

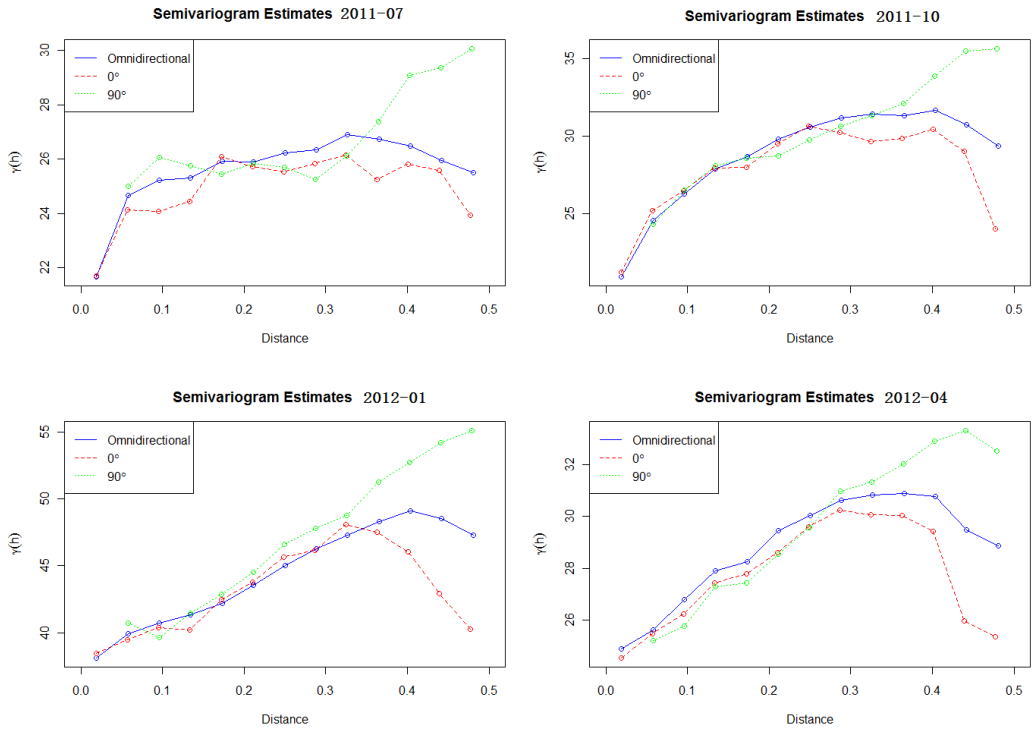

Figure 5.18: Semi-variogram Estimates of Water Vapor Skewness for Each Month

As shown in **Figure 5.18**, the four plots depict semivariogram estimates of water vapor in each of the four month. In July, the skewness at different pixel is correlated to each other up to a relative distance of 0.2. The semivariogram estimates in other months depict a highly variability in skewness, especially in January. It explained the exceeding noisy on the skewness map of January. Estimates in N-S and W-E directions on all the plots reveal a highly regular spatial

variability.

The variogram of water vapor skewness in four months, as shown in **Figure 5.19**, appears to follow spherical model, and they share the similar shape. The skewness is highly correlated with each other for separation distance of less than 0.2 of the relative distance. In the case of the spherical model, the parameters to fit the spherical models are:

Month	c_0	c_e	a_e
2011-07	44.8930994	8.1432768	0.1608551
2011-10	43.0377484	14.5147029	0.1962805
2012-01	77.1184634	9.7272200	0.2871647
2012-04	50.5961879	6.3430784	0.2361531

Table 5.1: Parameters in Spherical Models

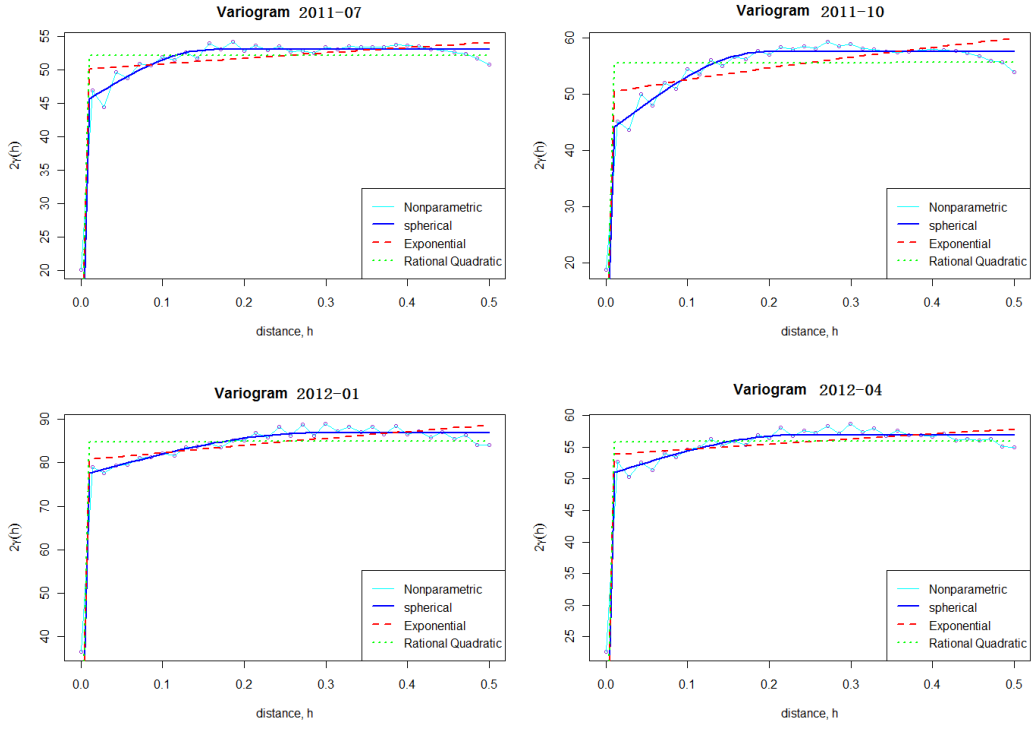

Figure 5.19: Variogram of Water Vapor for Each Month

5.6 Correlation between Temperature Skewness and Water Vapor Skewness

As shown in the skewness maps **Figure 4.7** and **Figure 5.16**, the skewness of temperature and water vapor seems to be positive correlated, and the correlation tends to be higher at the region between latitude $-45°$ and $+45°$.

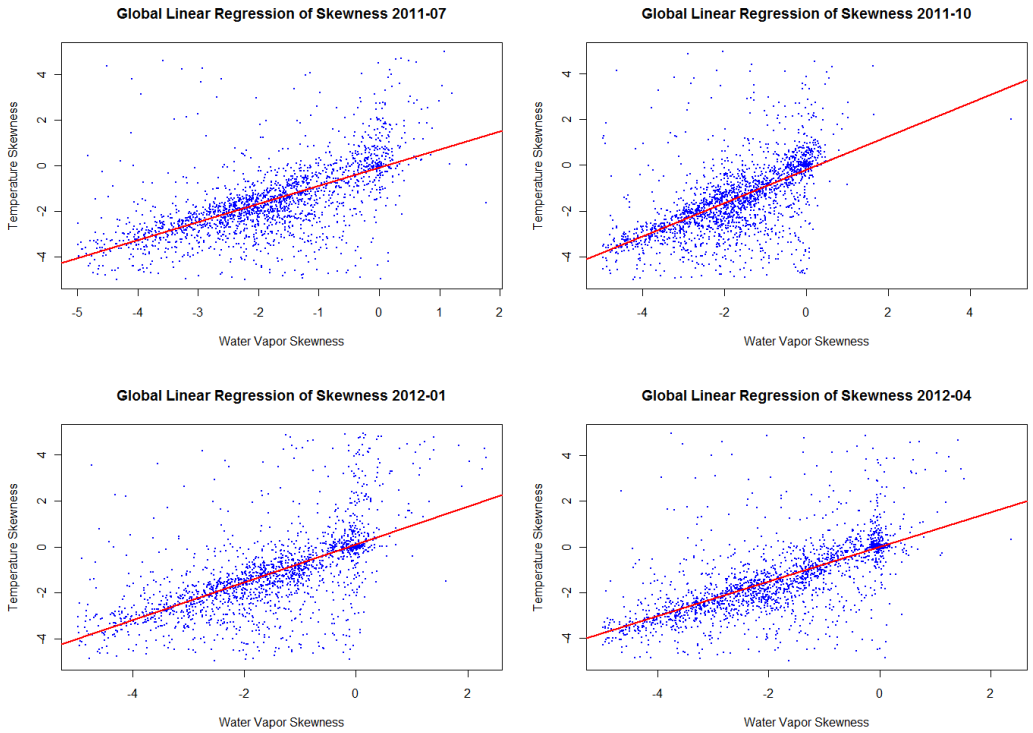

Figure 5.20: Global Linear Regression for Each Month

From **Figure 5.20**, we could see that the least square linear regression fits the temperature and water vapor skewness globally well. It depicts a positive correlation between temperature and water vapor skewness. The R^2 for each of the month regression model is, 0.4138, 0.3886, 0.4247, 0.4809 respectively, and the slopes are 0.703451, 0.757459, 0.717439 , 0.69453.

It was mentioned earlier that the region between latitude $-45°$ and $+45°$ containing strong negative skewness is consistent with that of temperature, and they somehow share the same pattern. We should expect a higher linear correlation based on the global skewness maps. **Figure 5.21** confirmed this intuition by showing a stronger linear relation. The R^2s are 0.4348, 0.3961, 0.4871, 0.4919, and the slopes are 0.6428, 0.6987, 0.6524,0.6011. The higher correlations indicate that water vapor and temperature tend to have more similar distribution at this region.

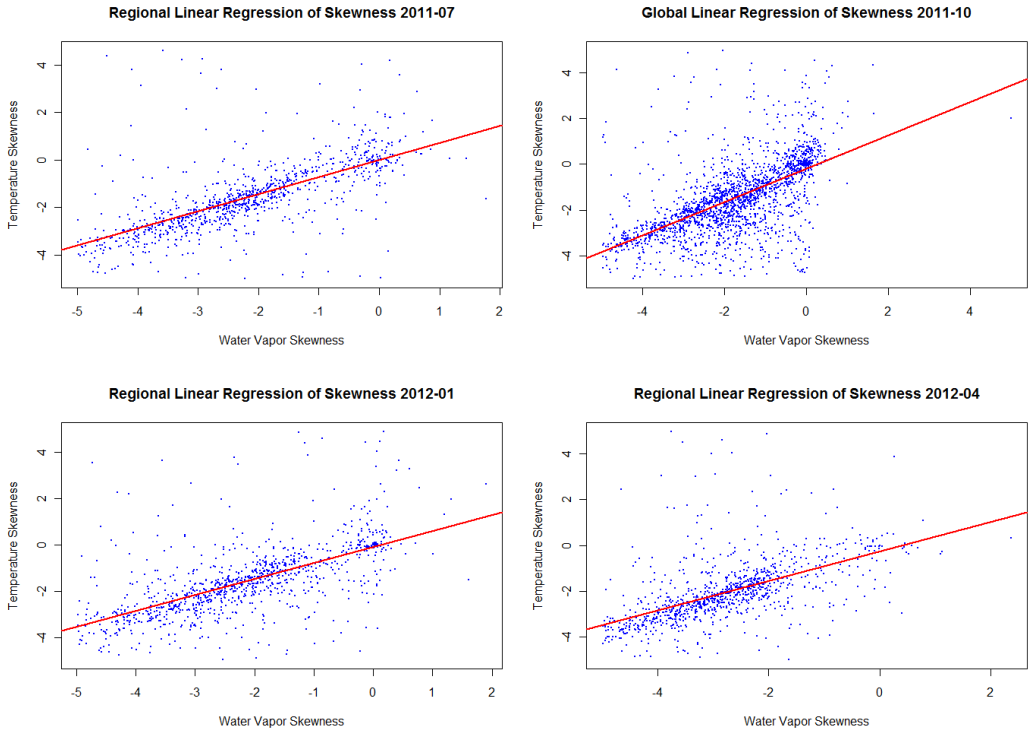

Figure 5.21: Regional Linear Regression for Each Month

CHAPTER 6

Scientific Study

In this chapter, I discuss the relation between temperature and water vapor in physical values and compare it to the relation between temperature and water vapor distribution skewness. In addition, I study the relation between temperature skewness and different altitudes in Southeast Asia.

6.1 Water Vapor and Temperature

Recent climate model simulations[11] lend theoretical support to a positive feedback among surface temperature, water vapor and green house effect. One might expect the relationship between water vapor and sea level temperature to resemble an exponential form, because saturation vapor pressure increases roughly exponentially with temperature. This relationship is fairly well described by the equation of the form,

$$lnW = A + B \cdot T \tag{6.1}$$

where W stands for water vapor and T stands for temperature. A and B are coefficients that depends on geographic region.

As shown in **Figure 6.1**, compile all the surface temperature and water vapor in each month on the scatter plot. One can observe that water vapor increase exponentially with temperature, and it complies with the theoretical support in [11]. This also complies with the finding of positive correlation between the distri-

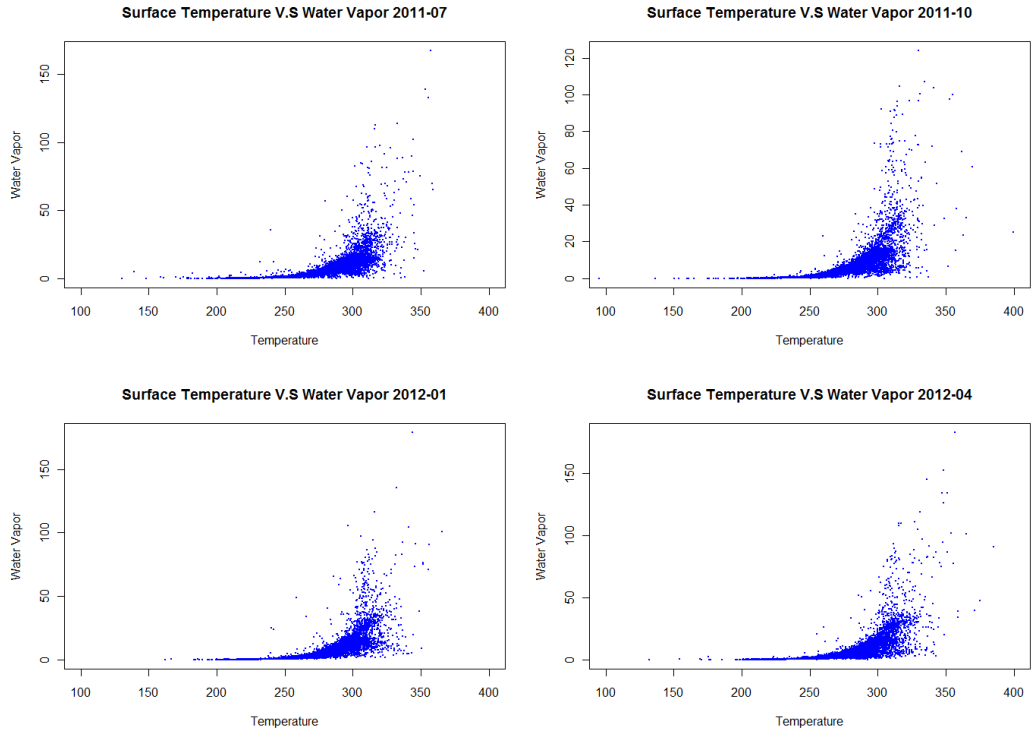

Figure 6.1: Surface Temperature against Water Vapor, Physical Value

bution skewness of surface temperature and water vapor. Unlike the exponential relation between the physical value of temperature and water vapor, the skewness of temperature and water vapor distributions tends to have a positive linear relation. The linear relation is discussed in Chapter 5.6. **Figure 5.20** and **Figure 5.21** reveal the strong linear relation of skewness both globally and regionally.

6.2 South East Asia

Figure 3.5 to **Figure 3.8** depict the global temperature skewness in 2011-07, 2011-10, 2012-01 and 2012-04. The negative skewness increases as the altitude getting higher. The skewness has a close relationship with pressure level. In this section, I use the skewness in Southeast Asia (2012-01) to study the relationship

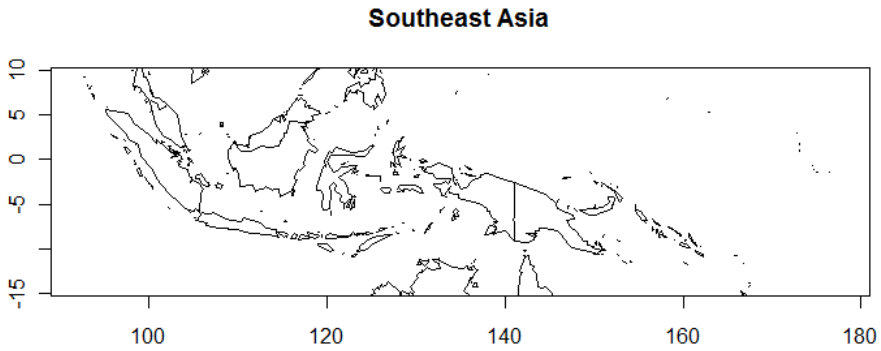

Figure 6.2: South East Asia

between temperature skewness and altitude change. The reason of choosing this area is that it is possibly one of the most vulnerable areas in the global-climate-change scenarios now being put forward by scientists [13], therefore it is benefit to learn more about the temperature distribution pattern in this region.

Southeast Asia covers the area that is bounded by Latitude $-15° \sim +10°$, and longitude $90° \sim +180°$. If reflect on the entire grid data set, it is a subset bounded by Row 16 to 20, and Column 55 to 72. There are 80 grid boxes in this region.

Figure 6.3 shows the skewness against altitude in each grid box at different latitude rows. Each line in the plot represents the trend in one grid box, and each plot contains 18 such lines. The trend of skewness against pressure level possesses a strong polynomial characteristic. Fitting the polynomial regression of skewness against altitude in each grid box will help us understand how temperature skewness changes with pressure level in Southeast Asia. In this case, a second order polynomial regression will be sufficient enough to fit the non-linear model, and it has the form,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \tag{6.2}$$

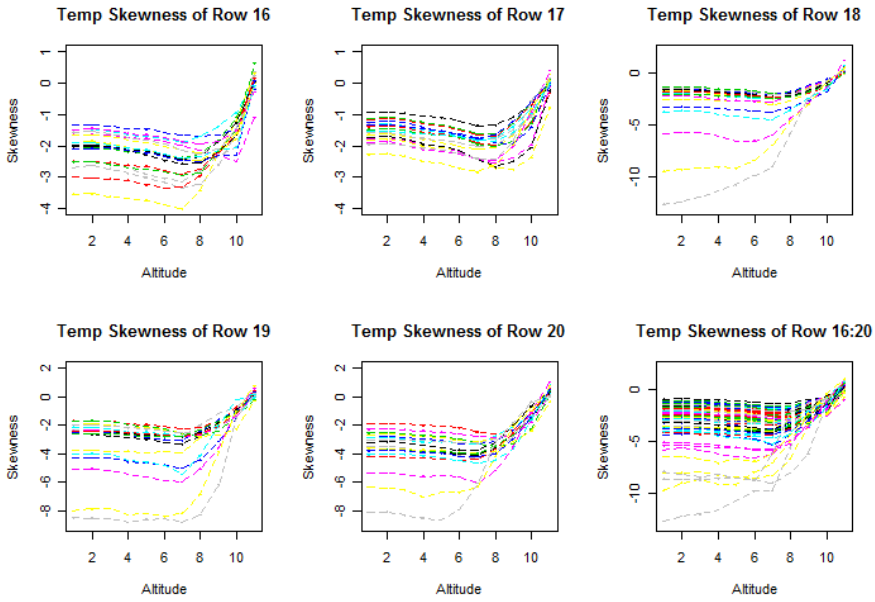

Figure 6.3: Skewness in Southeast Aisa

It gives us 240 coefficients by fitting 80 such polynomial regression models in each grid box. **Table 6.1** shows the average of the three coefficients in each latitude row.

Latitude	β_0	β_1	β_2
16	-1.23	-0.55	0.05
17	-0.76	-0.48	0.05
18	-2.34	-0.62	0.08
19	-2.08	-0.85	0.09
20	-2.56	-0.83	0.10
overall	-1.80	-0.67	0.07

Table 6.1: Average Polynomial Regression Coefficients

Figure 6.4 includes the scatter plot of skewness of different pressure levels at Latitude Row 16. As shown in this figure, the red dots represent the skewness at

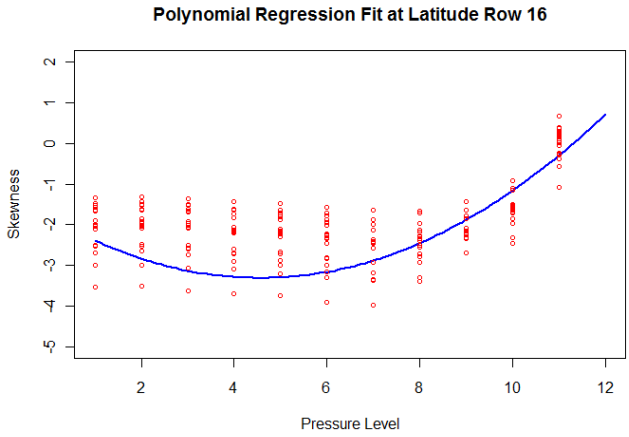

Figure 6.4: Polynomial Fit of Skewness V.S. Pressure Level

different pressure level in each grid box. The blue curve is the fitted line with the average coefficients in latitude Row 16. It depicts a polynomial trend of skewness against altitude in the region bounded by latitude $-15° \sim -10°$, and longitude $90° \sim +180°$. **Figure 6.5** is the map showing how the coefficients change across all the grid boxes. The region with low intercept (β_0) value tends to have strong negative skewness at the pressure level close to surface.

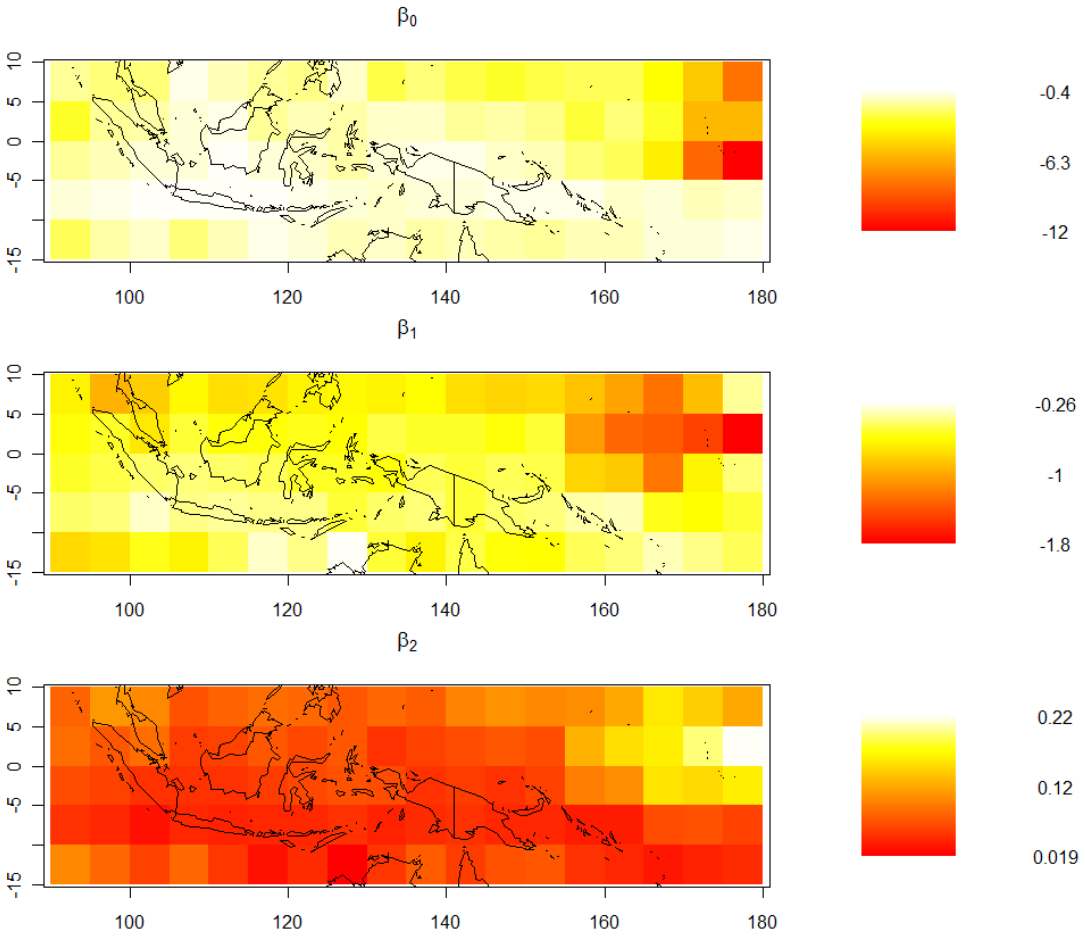

Figure 6.5: Maps of Polynomial Regression Coefficients

CHAPTER 7

Conclusions

In this thesis, we study potential statistical methods to analyze an unconventional type of data, AIRS Level 3 data. I use a sampling method to first construct mixture distributions and then study the skewness derived from each distribution. After populating each grid box with distribution skewness, we perform the lattice geospatial analysis on it. The following are some highlights of our findings.

1. The global temperature distribution skewness in different months shares the same pattern with most of the tropical area skewed to left, and most of the other regions have skewness close to 0. This complies with the tropical convection theory in climate science.

2. In October, January and April, the variograms generated for the skewness of temperature distribution in each pixel follows the Spherical model. In July, it follows the exponential model. In general, the skewness at different pixels is correlated with each other up to a relative distance around 0.25.

3. Water Vapor skewness has larger noise and more positive skewness compared to temperature skewness. The variograms of water vapor skewness follows spherical model in all of the 4 months.

4. Unlike the exponential relationship between the physical values of temperature and water vapor, the linear model is a good fit for the relationship between temperature and water vapor distribution skewness, especially in the tropical region.

5. In general, temperature skewness increases as the pressure level decreases. In the region of Southeast Asia, the trend between skewness and pressure level follows a strong polynomial characteristic.

Through the analysis on AIRS L3Q data, we demonstrate that skewness contains useful information for analysis about global and regional short-term climate. However, climate data set usually contains a few outlying data values. The regular moment skewness will be highly influenced by these values. If the L Moments is taken, it will be far less sensitive to the outlier. In addition, 4 months of data is not sufficient to study climate change. However, the fact is that our world is currently experiencing issues caused by global warming and it continues to threaten mankind. Hence, it will be useful to perform a tempo-spatial analysis over the data of a longer period of time. We will pursue this direction in our future research.

References

[1] Braverman, A. , Eric J. F., Brian H. K., (2011), *Massive Data Set Analysis for NASA's Atmospheric Infrared Sounder*, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.

[2] Johnson, M. H., Ladner, R. E., and Riskin, E. A. (2000), *Fast Nearest Neighbor Search for ECVQ and Other Modified Distortion Measures* , IEEE Transactions on Image Processing, 1435-1437.

[3] Edward T. O. (2007), *Quantization Product Quick Start*, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.

[4] Marzban C. , Sandgathe S. (2008), *Verification with Variogram*, Department of Statistics, University of Washington, Seattle, Washington.

[5] Cressie, N. (1990), *Statistics for Spatial Data*, John Wiley & Sons, Inc., ©1993

[6] Oliver, M.A. (2001), *Kriging: a method of interpolation for geographical information systems*, The University of Birmingham, Birmingham, UK.

[7] Wand, M., Jones, M. (1995), *Monographs on Statistics and Applied Probability*, Chapman & Hall, ©1995.

[8] Miller, I., Miller, M. (2004), *Mathematical Statistics with Applications*, Pearson Education, Inc., ©2004.

[9] Gettelman, A., Salby, M., Sassi, F. (2002), *Distribution and influence of convection in the tropical tropopause region*, National Center for Atmospheric Research, Boulder, Colorado.

[10] *Water Vapor in the Atmosphere*, `http://www.geog.ucsb.edu/~joel/g266_s10/lecture_notes/chapt03/oh10_3_01/oh10_3_01.html`, July 2012.

[11] Hall, A., Manabe, S. (2000), *Effect of water vapor feedback on internal and anthropogenic variations of the global hydrologic cycle*, Atmospheric and

Oceanic Sciences Program, Princeton University, Princeton, New Jerse.

[12] *Atmospheric Thermodynamics*, http://atoc.colorado.edu/~cassano/atoc5050/Lecture_Notes/wh_ch3_part3.pdf, July 2012.

[13] Yuen, B., Kong, L. (2009), Climate Change and Urban Planning in Southeast Asia, *Cities and Climate Change*, 2009 Vol. 2.

[14] Ward, M., Gleditsch, K. (2007), *An Introduction to Spatial Regression Models in the Social Sciences* , Department of Political Science University of Washington Seattle, Washington.

[15] Su, H., Jiang, J. (2008), *Observed Vertical Structure of Tropical Oceanic Clouds Sorted in Large-Scale Regimes*, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.