HYDROLOGICAL VISUALIZATION AND ANALYSIS SYSTEM AND DROUGHT RELATED FEATURE SELECTION BASED ON SECTIONAL CORRELATION MEASUREMENT

by

PIRAPORN JANGYODSUK

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2012

To my mother Nipa and my father Wichai Kurathong
who sacrifice themselves to made me who I am.

ACKNOWLEDGEMENTS

I would like to thank my supervising professor, Dr. Jean Gao, for constantly motivating and encouraging me, and also for her invaluable advice during two years of my master studies. I wish to thank Dr. Ramez Elmasri and Dr. Dong-Jun Seo for their interest in my research and for taking time to serve in my master thesis committee.

I would also like to extend my appreciation to Dr. Dong-Jun Seo and the Office of Hydrologic Development in the National Oceanic and Atmospheric Administration (OHD-NOAA) for providing invaluable hydrological data. I am grateful to Dr. Jean Gao, Vassilis Athitsos and Dr. Heng Huang for all the great classes and materials that help me finish my master thesis. I am especially grateful to Kim Dongchul and Ashis Kumer Biswas for providing and organizing a new server. I wish also to thank all my lab mates for their interest in my research and for the helpful discussions and invaluable comments. Again, I wish to thank Dr. Jean Gao for taking the time to review and comment this manuscript.

I am deeply grateful to all the teachers, both in Thailand and in the Unites States. I would like to thank Dr. Siriporn Ong-rungreung for encouraging to pursue graduate studies in the United States.

Finally, I would like to express my deep gratitude to my husband who supports me, both in academics and in life. I am extremely fortunate to be so blessed. I am also extremely grateful to my family in Thailand for their sacrifice and encouragement.

July 19, 2012

ABSTRACT

HYDROLOGICAL VISUALIZATION AND ANALYSIS SYSTEM AND DROUGHT RELATED FEATURE SELECTION BASED ON SECTIONAL CORRELATION MEASUREMENT

PIRAPORN JANGYODSUK, M.S.

The University of Texas at Arlington, 2012

Supervising Professor: Jean Gao

Because of the larger data storage and the faster computational power, computer can process and store much finer resolution data. Aside from data analysis, data visualization is also an important task to understand the data. In this work, the Hydrological Visualization and Analysis System is developed to help both hydrologists and local people view and examine the high resolution hydrological data. Then, this data is analyzed to determine which variables are related to the change of drought condition in Arlington, Texas. A new correlation measurement method called sectional correlation is proposed and used as an objective function of the drought related feature selection. The proposed sectional correlation algorithm has good performance in terms of computational efficiency and the accuracy. The result error is quite low, about 2% of the range of value.

TABLE OF CONTENTS

Appendix

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Computer has been used to help analyze and visualize the data from many fields such as Medical, Biology, Geology, Hydrology, etc. Since the computer storage has become larger and larger, and the computational power is much faster than in the early age, the data can be computed and kept in finer resolution. Recently, a new 30-year historical hydrological data set has been produced from the Office of Hydrologic Development in the National Oceanic and Atmospheric Administration (OHD-NOAA). This data set has 32 hydrological variables in finer spatial and temporal resolution which has advantage over other data sets in terms of detail and accuracy.

This work uses the available data to provide two contributions. First, the Hydrological Visualization and Analysis System (or HyVAS) has been developed to visualize this data set and made available to hydrologists, local organizations, and people in the West Gulf River Forecast Center (WGRFC) area. This visualization is designed to be interactive so users can easily manipulate the visualized data as interested. Many pros and cons from other visualization applications are considered before this system is decided to be a web-application.

Because this data set is a spatio-temporal data. The visualization can be done in two ways, space (area) and time. The spatial visualizations are mostly done by plotting a map (2D). In some variables such as soil moisture, and soil temperature, the data is sub-categorized into four layers. Thus, another dimension, which is the depth, also needs to be visualized. For these variables, the data can be displayed in 3D.

Temporal visualization is simply done by plotting time series onto a graph. This graph is also flexible to manipulate and interact for easier data understanding.

The second contribution is the data analysis which is focused on drought condition. This analysis is which hydrological variables are related to the change of drought condition in Arlington, Texas. The forward feature selection is applied to select the variables. Then, the measurement of how good the selected features are must be defined. The correlation measurement is chosen to do this job but some drawbacks in popular methods are found so these methods cannot be used. Thus, a new correlation measurement algorithm called sectional correlation is proposed and is used as the objective function of the drought related feature selection.

1.2 Thesis Overview

The rest of this thesis is organized into two chapters.

Chapter 2 provides the information of the Hydrological Visualization and Analysis System (HyVAS). It starts by describing the hydrological data used in this work. Then, the HyVAS is introduced. Its architecture, data-flow, and functions of the five visualization tools are explained. It also shows how the 2D and 3D visualization are computed and displayed on a web browser.

Chapter 3 focuses on analysis of the correlation between the hydrological data and five drought indices. First, two current problems in sectional correlation measurement and in noise detection in time series motif discovery are identified. It covers some widely used correlation measurement algorithms and their drawbacks. After that, it introduces a new algorithm for sectional correlation measurement. The new feature selection method based on the proposed sectional correlation is presented. In the last section, the experimental results and explanations for sectional correlation and the drought related feature selection is mentioned.

CHAPTER 2

HYDROLOGICAL VISUALIZATION AND ANALYSIS SYSTEM

2.1 Introduction

Visualization is an important tool to find spatial and temporal relationship in numerical data. The purpose of this web-application, the Hydrological Visualization and Analysis System (HyVAS), is visualizing and also analyzing the historical hydrological data from the Office of Hydrologic Development in the National Oceanic and Atmospheric Administration (OHD-NOAA). This web-application is hoped to help the hydrological community view the data and find the relationship, trend, hidden pattern, or any useful information in this data.

2.2 Related Works

There are many applications for hydrological data visualization, such as HydroDesktop [1], ArcGIS [2], and GISHydro [3].

The major advantage of the HydroDesktop is that it has been developed since 2009 and more recent data is available via the CUAHSI Hydrologic Information System. Its purposes are downloading, visualizing, and analyzing hydrologic and climate data from the CUAHSI-HIS. The drawback is its OS and environment dependent. It must be run on Windows XP or Windows 7. For other operating system, this application must be run on a Windows emulator. Furthermore, the .NET Framework is required. This installation process might be frustrating.

The ArcGIS is commercial software from Esri. There are many advantages over other applications. It also has been developed for a long time. The current version is 10.1. Many useful functions have been added. It's widely used around the world. Many organizations also provide data in ArcGIS format or *.shp*. The main drawback is the license fee. Thus, the local users cannot afford this application for leisure use. Another disadvantage is that its function cannot be modified to serve the users' interest.

The GISHydro has two versions, the web version and the software version. Both of them are free to use. They have been developed since 1997. The purpose of the GISHydro is to help engineers analyzing watershed in Maryland area. The disadvantage of the software version is its dependent on ArcView or ArcGIS platform. For the web version, it also needs a plug-in called Citrix before accessing to the GISHydro web server.

The HyVAS is developed to eliminate all the applications' flaws mentioned above. The first advantage of HyVAS over others is that because it's a web-based application, it doesn't need an installation. Also, it is platform independent so that it can be used on any OS. In addition, no license fee is needed. The only recommendation is to use it in Google Chrome to get all the functions provided.

Another advantage of the HyVAS is its data set. This data set has finer spatial and temporal resolution than others which provides more accurate information for users who interest in small areas like in state, county, or city level. This can help the local hydrological organizations view and manage their water supplies.

Finally, this system provides both spatial and temporal visualization while most of others provide only spatial functions. Besides, the HyVAS provides 3D-visualization that is lacked in many other web-based hydrological applications. The 3D Trend of Soil Moisture Range Tool in the HyVAS provides 3D-maps with soil depth which will help users see trend or relationship of the data between soil layers.

2.3 Methodology

The HyVAS has three tools for spatial visualization and two tools for temporal visualization. The method for spatial visualization is rendering a colored map based on the value in each geographical position. For the temporal visualization, a time series created from all 30-years data at a certain geographical position is displayed. In this section, the data set will be introduced first. The overall architecture will be discussed, then each of the five visualization tools will be brought up.

2.3.1 Data Set

The hydrological data used in this work is kept in HRAP (Hydrologic Rainfall Analysis Project) coordinate system [4] which is another coordinate system used within the National Weather Service. It can be converted to and from the geographic coordinate system (latitude and longitude).

$$HRAP_X = \left(\frac{Earth\ Radius}{Grid\ Length} \times (1+\sin(sdlat)) \times \frac{\cos(lat)}{(1+\sin(lat))} \times \sin(lon)\right) + 401 \tag{2.1}$$

$$HRAP_Y = \left(\frac{Earth\ Radius}{Grid\ Length} \times (1+\sin(sdlat)) \times \frac{\cos(lat)}{(1+\sin(lat))} \times \cos(lon)\right) + 1601 \tag{2.2}$$

where *lat* is latitude in radians, *lon* is longitude in radians, and *sdlat* is the standard latitude in radians.

$$gi = \left(\frac{Earth\ Radius}{Grid\ Length} \times (1+\sin(sdlat))\right)^2 \tag{2.3}$$

$$rr = (HRAP_X - 401)^2 + (HRAP_Y - 1601)^2 \tag{2.4}$$

$$latitude = \arcsin\left(\frac{gi - rr}{gi + rr}\right) \times rad2deg \tag{2.5}$$

$$longitude = 270 + sdlon - \arctan\left(\frac{HRAP_Y - 1601}{HRAP_X - 401}\right) \times rad2deg \qquad (2.6)$$

where $sdlon$ is the standard longitude, $sdlat$ is the standard latitude, and $rad2deg$ is the conversion from radians to degrees.

This data set has 32 variables, as shown in Table 2.1, in high spatial resolution (4×4 km^2. per cell) and also in high temporal resolution (every 6 hours from January 2, 1979 to December 31, 2008). This spatial coverage is all the states in the US except Alaska and Hawaii.

Table 2.1: 32 Variables of the OHD-NOAA's hydrological data set

	Variable	Layer	Description
1	accmax	1	Maximum water equivalent since snow began to accumulate, mm
2	adimpc	1	Additional impervious area water content, mm
3	evap	1	Actual evapotranspiration, mm per dt
4	liqw	1	Liquid water storage, mm
5	lzfpc	1	Lower zone primary free water content, mm
6	lzfsc	1	Lower zone supplemental water content, mm
7	lztwc	1	Lower zone tension water content, mm
8	pevap	1	Potential evapotranspiration, mm per dt
9	rain	1	Rainfall forcing, mm per dt
10	rmlt	1	Rain plus melt dept, mm
11	runoff	1	Surface flow component, mm per dt
12-15	smliq	1-4	Unfrozen volumetric soil moisture at Noah defined layers
16	sndpt	1	Snow depth, mm
17	snow	1	Snowfall forcing, mm per dt
18	snowfrac	1	Snow cover fraction, dimensionless
19-22	soilm	1-4	Total volumetric soil moisture at Noah defined layers
23-26	soilt	1-4	Soil temperature at Noah defined layers
27	subflow	1	Subsurface flow component, mm per dt
28	swe	1	Snow water equivalent, mm
29	tem	1	Air temperature forcing, C
30	twe	1	Total water equivalent, mm
31	uzfwc	1	Upper zone free water content, mm
32	uztwc	1	Upper zone tension water content, mm

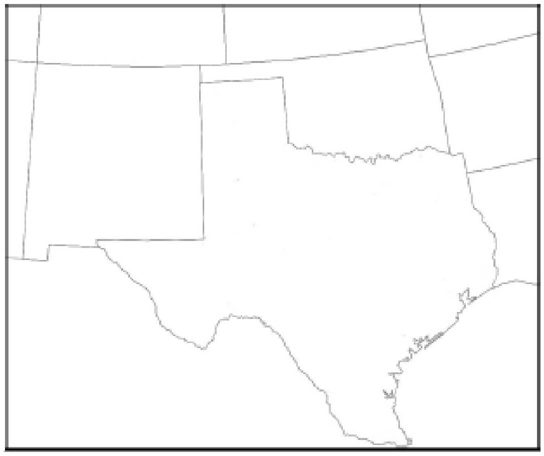


Figure 2.1: West Gulf River Forecast Center (WGRFC) Area

Although the data covers all the US, this system provides the spatial visualization only in the West Gulf River Forecast Center (WGRFC) area for better and faster rendering result. The WGRFC (see Figure 2.1) includes Texas, Oklahoma, New Mexico and some part of Arizona, Utah, Colorado, Kansas, Missouri, Arkansas, and Louisiana. From all 32 variables, this current system focuses only on the soil moisture (soilm) because this variable is important in monitoring and possibly predicting the drought condition. Anyway, other variables can be added to the system later.

2.3.2 Overall Architecture

In order to keep the system simple and not redundant, all five tools in the HyVAS share the same architecture which covers the database, data-flow, and web-interface.

2.3.2.1 Database

The data is kept in *.mat* format which can be read and written by MATLAB. Thus, this system uses MATLAB as its database. The reason that common relational databases, such as MySQL, MS SQL, and Oracle, are not chosen is that one of the tools, which is the 3D Trend of Soil Moisture Range Tool, needs to process Region Connecting Algorithm which can be done faster in MATLAB. Another reason is to safe time retrieving data from a common database and sending data to MATLAB by keeping the data in the format ready for MATLAB to process. For other four tools, a common database can be used but it will cause redundant to the system. To make the system's database simple and non-redundant, MATLAB is chosen for this task.

2.3.2.2 Data-Flow

The common technologies used in data flow of all tools are HTML, JavaScript, Ajax, PHP[5], and MATLAB. The data flow (see Figure 2.2) starts when the web-application receives input data from the user. It sends a request for data via Ajax to a PHP page which triggers the MATLAB using a *system* command. After that, the MATLAB reads the requested data and writes it to a CSV file. Then, the PHP page reads the CSV file and sends the requested data back to the web-interface. Finally, the web-interface keeps that data locally for graphic visualization rendering by JavaScript.

2.3.2.3 Web-Interface

The five tools in the HyVAS can also be categorized based on the graphic rendering modes into 2D and 3D Graphic Visualization Tools.

1. 2D Graphic Visualization Tools

   The tools in this category are Average Soil Moisture Map Tool, Full Time Series Tool, and Time Series Comparison Tool. The visualization part of these tools, which are 2D map and graph, are rendered by HTML5-Canvas elements ([6], [7]). For rendering a map, each HRAP cell is represented by a colored rectangle.

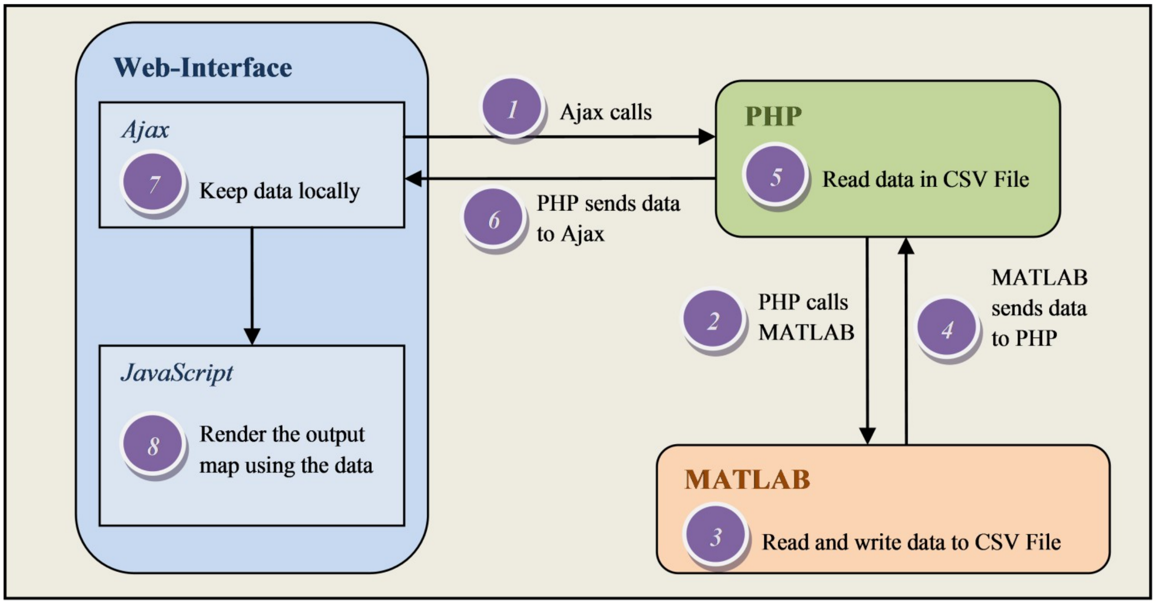


Figure 2.2: Data-flow of the HyVAS

The color map is displayed beside the map to provide interpretation of the map. In a graph rendering, two axis and title are drawn first. Then, the temporal data is sorted chronologically and plotted on the pre-rendered graph. These tools have the same set of visualization-aid functions as following,

- Zoom: The map/graph can be zoomed in and out in eight levels. JavaScript is used to re-render the map/graph to the bigger or smaller size while the display size is fixed. When a graph is zoomed, its two axis are also expanded and shrunk accordingly.
- Pan: When the map/graph is zoomed in, it can be pan to any area of interest but limited within the WGRFC area. When the map/graph is zoomed-in, the rendered map/graph size is bigger than the display size. Then, the display box can be moved around by JavaScript to show different parts of the map/graph. When a graph is panned, its two axes are also move accordingly. To move the display box, the JavaScript changes the

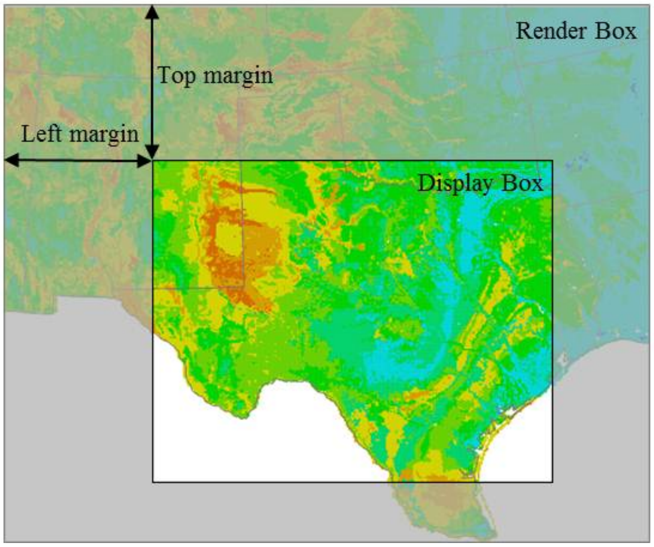


Figure 2.3: Pan by changing top and left margin

left and top margin of the display box as in Figure 2.3. This move process is done without graphic re-rendering so it is fast and smooth.

- Information Tool-tip: If this function is enabled, the relevant information of that map point or time series will be displayed on a tool-tip. Because the data shown in the tool-tip has already been computed to show in the Information box, JavaScript only activates the tool-tip to appear on the screen.

2. 3D Graphic Visualization Tools

The tools in this category are 4-Layer-Soil Moisture Maps Tool, and 3D Soil Moisture Trend Tool. These tools' visualization parts, which are 3D maps and stacks of four maps, are rendered by a WebGL library called *three.js* ([8], [9]). In a map rendering, each HRAP cell is represented by four vertices and a rectangle face. First, four vertices are put onto 3D coordinate. Then, the rectangle face with a color that represented that cell value range is put on those vertices. The common set of visualization-aid functions are as following,

- Zoom: The map can be zoomed in and out in by the zoom-in and zoom-out buttons, and also by dragging mouse up or down. JavaScript is used to move the 3D map closer/further to the user so that it appears bigger/smaller. Because re-rendering is not required, this function is very efficient and smooth.
- Rotate: The map can be rotated in all three axis by six buttons (two buttons for each axis) and also by dragging mouse. JavaScript is used to control the 3D map rotation. Also, no re-rendering is needed.
- Area Selection: This allows users to select the area of their interest. Only the selected area will be visualized. This function helps reducing the data retrieving time and the graphic rendering time when user is interested in some specific area.

2.3.3 Average Soil Moisture Map Tool

The purpose of this tool is spatially visualizing average soil moisture value in the WGRFC area by coloring each geographical position (HRAP cell) based on its soil moisture value. The range of soil moisture value and its color is displayed in Figure 2.4. For example, if the soil moisture value at HRAP cell (100,100) is 0.2315, the color of that cell is bright yellow. The average soil moisture value is calculated from

$$\overline{soilm} = \frac{\sum\limits_{i=1}^{4} soilm_i \times depth_i}{\sum\limits_{i=1}^{4} depth_i}, \tag{2.7}$$

where $\overline{soilm}$ is the average soil moisture value, $soilm_i$ is the soil moisture at level i, and $depth_i$ is the depth (cm.) of that soil layer (see Table 2.2). This tool helps users compare and find trend of soil moisture between regions of the WGRFC area.

Table 2.2: Depth of soil at each layer

Layer	Depth (cm.)
1	0 - 10
2	10 - 40
3	40 - 100
4	100 - 200

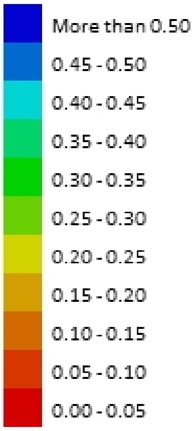
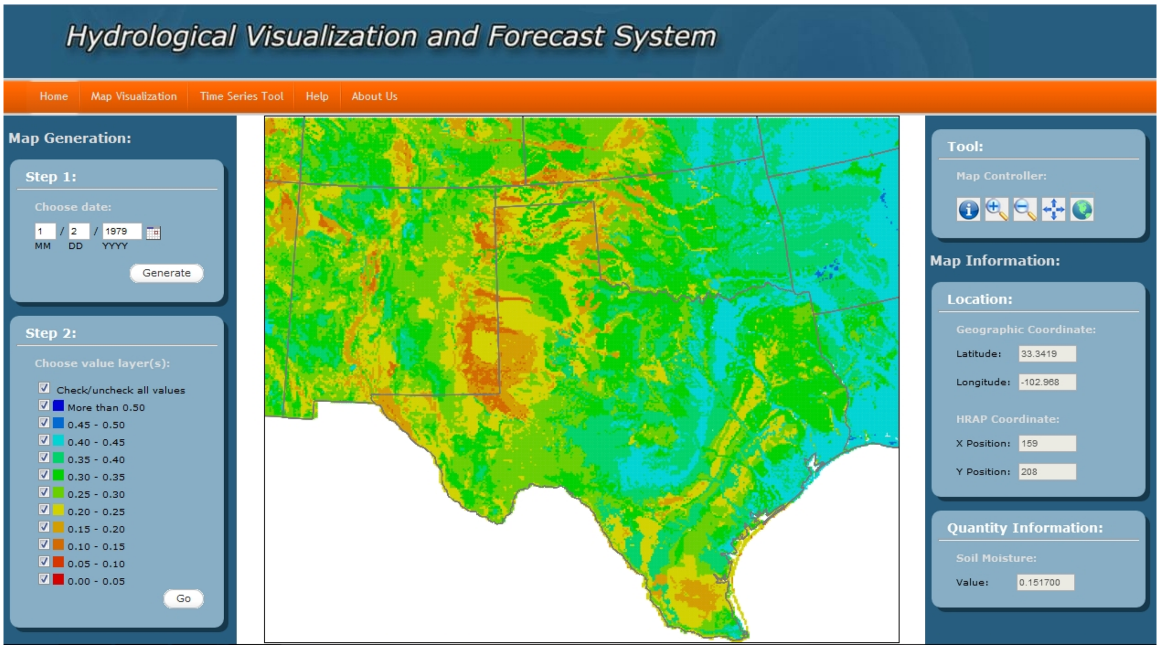


Figure 2.4: Ranges of soil moisture value and their represented color



Map Generation:

Step 1:

Choose date:

1 / 2 / 1979
MM DD YYYY

Step 2:

Choose value layer(s):

☑ Check/uncheck all values
☑ More than 0.50
☑ 0.45 - 0.50
☑ 0.40 - 0.45
☑ 0.35 - 0.40
☑ 0.30 - 0.35
☑ 0.25 - 0.30
☑ 0.20 - 0.25
☑ 0.15 - 0.20
☑ 0.10 - 0.15
☑ 0.05 - 0.10
☑ 0.00 - 0.05

Tool:

Map Controller:

Map Information:

Location:

Geographic Coordinate:

Latitude: 33.3419

Longitude: -102.968

HRAP Coordinate:

X Position: 159

Y Position: 208

Quantity Information:

Soil Moisture:

Value: 0.151700

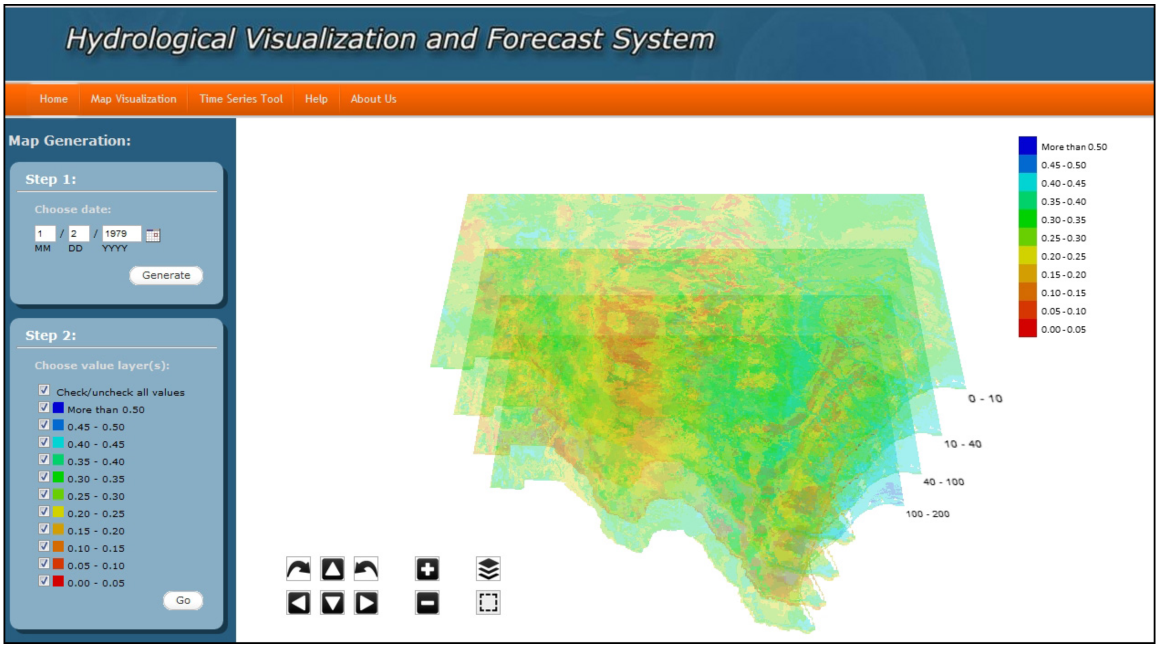
Figure 2.5: Average Soil Moisture Map of January 2, 1979

To use this tool, the user chooses the date of interest in the Step 1 box (Figure 2.5). Then, the colored average soil moisture map is displayed at the center of the web-application. After that, the user can use visualization-aid tools in the Tools box to enable information tool-tip, zoom in, zoom out, pan, and view full map as described in the Web-Interface section. To provide the numerical information and help the user easily pinpoint any location of interest, when the mouse moves over any point in the map, the geographic coordinate, the HRAP coordinate, and the soil moisture value of that point is shown in the Map Information box. In addition, when the user's only interested in some specific ranges of value, the ranges can be chosen or discarded by checking or unchecking color map checked boxes in the Step 2 box (Figure 2.5).

2.3.4 4-Layer Soil Moisture Maps Tool

This tool is for spatially visualizing all four layers of soil moisture value. Four maps are rendered at the same time. They are stacked on top of the other based on the order of layer. The $soilm_1$ map is on top, then the $soilm_2$ map and so on. The bottom one is the $soilm_4$ map. The user needs to choose one map to visualize at a time. The benefit is that when the user wants to change to other layer, it can be done easily without re-retrieving data and re-rendering map. This tool helps users compare and find trend of soil moisture between regions of the WGRFC area and also between layers.

The usage starts in the same process as the Average Soil Moisture Map Tool by choosing the date of interest. Then, four colored maps representing four layers of soil are displayed in transparent at the center of the web-application. The user can choose the layer to visualize by clicking on that layer. As a result, that chosen layer map will become a solid colored map and the others will be disappeared. After that, the user can use visualization-aid tools at the bottom left of the maps to rotate in



Map Generation:

Step 1:

Choose date:

1 / 2 / 1979
MM DD YYYY

Step 2:

Choose value layer(s):

- ☑ Check/uncheck all values
- ☑ More than 0.50
- ☑ 0.45 - 0.50
- ☑ 0.40 - 0.45
- ☑ 0.35 - 0.40
- ☑ 0.30 - 0.35
- ☑ 0.25 - 0.30
- ☑ 0.20 - 0.25
- ☑ 0.15 - 0.20
- ☑ 0.10 - 0.15
- ☑ 0.05 - 0.10
- ☑ 0.00 - 0.05

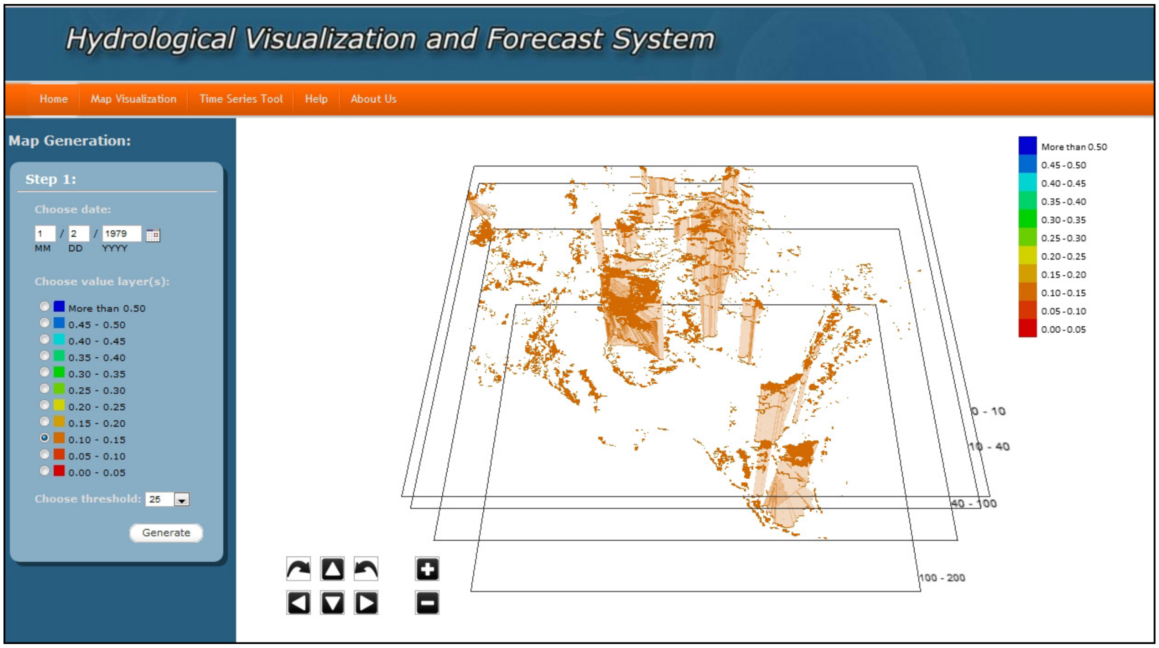
Figure 2.6: 4-Layer of Soil Moisture Maps on January 2, 1979

X, Y, or Z axis, zoom in, zoom out, re-select visualized layer, and re-select the area of interest. This tool also provides the ability to use mouse dragging to zoom and rotate the maps. In addition, when the user can choose or discard any range of soil moisture value by checking or unchecking color map checked boxes in the Step 2 box (Figure 2.6).

2.3.5 3D Trend of Soil Moisture Range Tool

The motivation of this tool is for analyzing and visualizing the trend or relationship of a specific soil moisture range between layers. The maps are also stacked in the same manner as in the 4-Layer of Soil Moisture Maps Tool but, in this tool, only one range of soil moisture is displayed at a time. This tool helps users compare and find trend of soil moisture between layers.

To use this tool, the user needs to choose the date and the soil moisture range of interest. Then, four colored maps representing four layers of soil are displayed at the



Map Generation:

Step 1:

Choose date: 1 / 2 / 1979 (MM DD YYYY)

Choose value layer(s):

○ More than 0.50
○ 0.45 - 0.50
○ 0.40 - 0.45
○ 0.35 - 0.40
○ 0.30 - 0.35
○ 0.25 - 0.30
○ 0.20 - 0.25
○ 0.15 - 0.20
● 0.10 - 0.15
○ 0.05 - 0.10
○ 0.00 - 0.05

Choose threshold: 25

Figure 2.7: 3D Trend of the 0.10 - 0.15 Soil Moisture on January 2, 1979

center of the web-application but these maps show only one color that representing the selected range (See Figure 2.7). The available visualization-aid tools are the same as those in the 4-Layer Soil Moisture Maps Tool.

Region Connecting Algorithm

In order to visualize the relationship or trend of a range of soil moisture value between adjacent layers, rectangle sides are put to connect the top layer edges to the bottom layer edges. This process creates a prism-like object which two bases (regions) are not required to have the same shape or area. See example in Figure 2.8.

The region connecting algorithm (in Table 2.3) works as following,

1. First, regions must be extracted from each layer.
    - The cells that have value out of the selected range are not considered so that the 3D map is not too busy and is easy to understand.

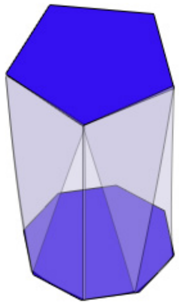


Figure 2.8: Example of connected pentagon region to octagon region

- Region label is assigned to each cell. The cells in the same region have the same number. This labeling process helps the algorithm to later identify which pair of regions are overlapped to each other.
- To reduce the processing and rendering time, some small regions which size is less than 25 cells, are discarded.

2. Then, for every pair of connected layer, connect the regions from the top layer to the bottom.
   - Find top and bottom regions that are overlap to each other more than 30% of the smaller region. For example, the top layer region has 100 cells and the bottom has 50 cells. If the overlap size is 15 cells, these two regions are connected. If the overlap size is less than 15 cells, this connection is discarded.
   - Extract edges from those regions. Edge is represented by a list of points. Thus, the starting point of each edge must be determined.
   - A greedy algorithm is applied to match points from two edges from the top layer and the bottom layer. Although the greedy algorithm might not yield the best match, it is fast enough for web-application and the result is not needed to be optimal.

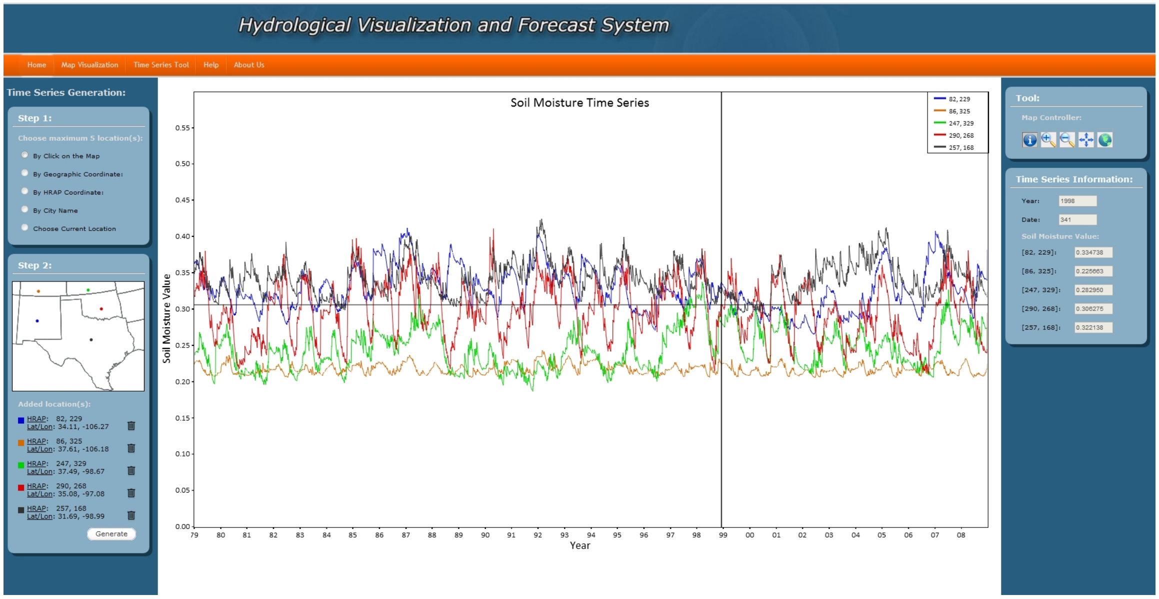
- Put a side, i.e. a rectangle face, connecting top layer region to the bottom layer region. In this step, the connected regions are presented as a prism-like object.

Table 2.3: The Summarized Region Connecting Algorithm

```
For each soil moisture layer (4 layers) {
    Cut out the HRAP cells that are out of the user's specified range
    Put the adjacent within-range-cells to the same region
    Assign a number to each region
    Drop regions that have less than 25 cells
}
For each pair of connected layers (3 pairs) {
    For each region on the top layer {
        Find overlap regions in the bottom layer
        For each overlapping region {
            Extract edge cells of those regions from both layers
            Determine a starting cell in both layers
            Use greedy algorithm to match top layer's edges to the bottom
            Put a rectangle face to connect the top to the bottom layer
        }
    }
}
```

2.3.6 Full Time Series Tool

The purpose of this tool is to temporally visualize the average soil moisture value for up to five specific HRAP cells at a time. This tool retrieves the chronologically sorted 30 years-data, and plot a time series. This should help users visualize and find the monthly/seasonal/annual trend hidden in the historical data. In addition, it helps visualize some statistic properties, such as mean and variant of the data at the specified HRAP cell.



Time Series Generation:

Step 1:

Choose maximum 5 location(s):

○ By Click on the Map

○ By Geographic Coordinate:

○ By HRAP Coordinate:

○ By City Name

○ Choose Current Location

Step 2:

Added location(s):

🟦	HRAP: 82, 229 Lat/Lon: 34.11, -106.27
🟧	HRAP: 86, 325 Lat/Lon: 37.61, -106.18
🟩	HRAP: 247, 329 Lat/Lon: 37.49, -98.67
🟥	HRAP: 290, 268 Lat/Lon: 35.08, -97.08
⬛	HRAP: 257, 168 Lat/Lon: 31.69, -98.99

Tool:

Map Controller:

Time Series Information:

Year:	1998
Date:	341

Soil Moisture Value:

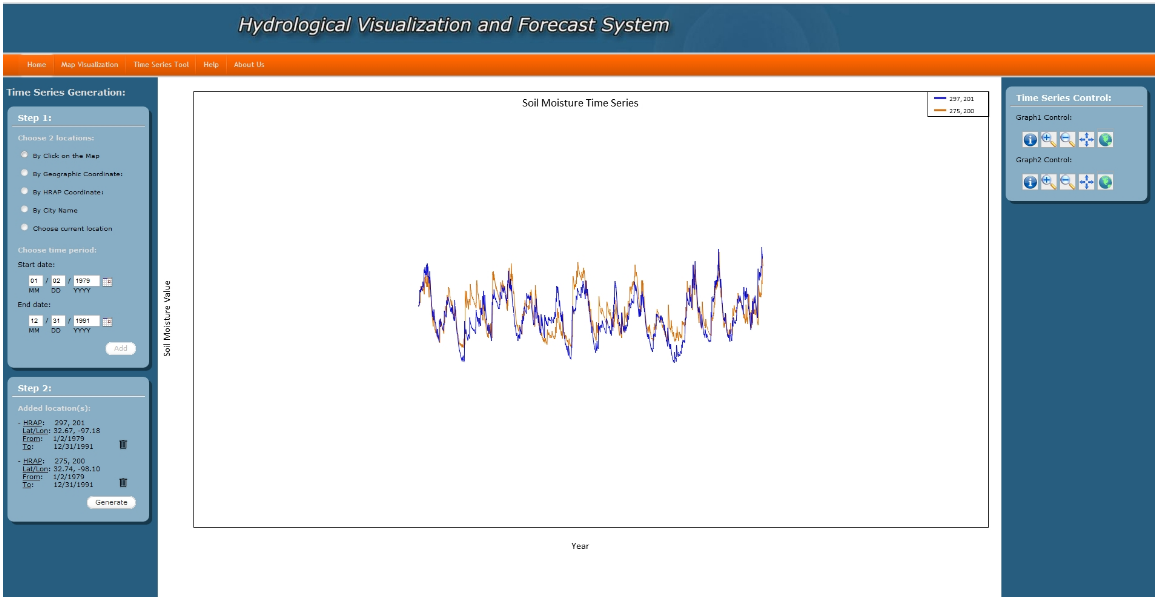
[82, 229]:	0.334738
[86, 325]:	0.225663
[247, 329]:	0.282950
[290, 268]:	0.306275
[257, 168]:	0.322138

Figure 2.9: Full Time Series Visualization Tool

The usage of this tool is a bit complicate because of many location selection methods. In the Step 1 box (Figure 2.9), there are five ways to choose an HRAP cell as following,

- By clicking directly on the input map
- By specifying a geographic coordinate (latitude and longitude)
- By specifying an HRAP coordinate (X-axis and Y-axis position)
- By selecting a city name from the database
- By using a Google API called *Geolocation* to get the user's current geographic coordinate

After choosing locations, the mini-map and the coordinate in both systems are shown in the Step 2 box (Figure 2.9) to summarize the information of the selected locations. The user can also remove the selected locations by clicking the bin button. If the total number of selected locations is less than five, more location(s) can be added. The available visualization-aid tools are enabling information tool-tip, zoom-



Start date: 01 / 02 / 1979 MM DD YYYY

End date: 12 / 31 / 1991 MM DD YYYY

- HRAP: 297, 201
  Lat/Lon: 32.67, -97.18
  From: 1/2/1979
  To: 12/31/1991
- HRAP: 275, 200
  Lat/Lon: 32.74, -98.10
  From: 1/2/1979
  To: 12/31/1991

Figure 2.10: Time Series Comparison Tool

ing in, zooming out, panning, and viewing full graph. When the information tool-tip function is enabled, a horizontal line and a vertical line, which are crossing at the mouse pointer, appear on the graph to help users better view the graph information. Furthermore, when the mouse moves over any point in the graph, the numerical information, i.e. the average soil moisture value, date and year, are provided in the Time Series Information box.

2.3.7 Time Series Comparison Tool

This tool is for visually comparing the similarity of two average soil moisture time series/sub-sequences. These two time series/sub-sequences can be from the same or different HRAP cell and date.

The usage of this tool is almost the same as that of the Full Time Series Tool. In the Step 1 box (Figure 2.10), besides the five location selection methods, the start/end date selection are provided for users to cut a sub-sequence from the 30 years-time series. The coordinate in both systems and the chosen period are displayed in the Step 2 box (Figure 2.10) to summarize the information about the displayed time series/sub-sequences. The user can also remove the selected locations by clicking the bin button and re-select a new one. This tool provides two sets of visualization-aid tools for each time series/sub-sequence. The visualization-aid tools are the same as those in the Full Time Series Tool. Each time series/sub-sequence can be zoomed in, zoomed out and move on top of the other for similarity comparison.

2.4 Future Work

The Hydrological Visualization and Analysis System is still in infant stage. There are lots of adjustments and expansions can be applied. Other 32 variables will be added. Area scope of this system can be expanded to view wider area or shrank to small level like state, county, or city. More tools and visualization-aid functions can also be added to serve both hydrologists and local users. It can also be modified for locally use in any organization. The only current limitation is that this system needs to be run on Google Chrome because it needs WebGL in 3D rendering. This limitation is hoped to be removed soon in a new version of browsers or by applying plug-ins to browsers.

2.5 Conclusion

In this work, we developed a web-based system that provides spatial and temporal visualization tools. It's aimed to help the hydrological community to visualize the new data set, OHD-NOAA data set, without a complicated process of program installation or any OS dependence. This system has many advantages over other works. It provides both spatial and temporal visualizations. It's environment independent. It's more flexible to add more functions and data set than commercial programs. Moreover, it provides both 2D and 3D visualization.

CHAPTER 3

DROUGHT RELATED FEATURE SELECTION BASED ON SECTIONAL CORRELATION MEASUREMENT

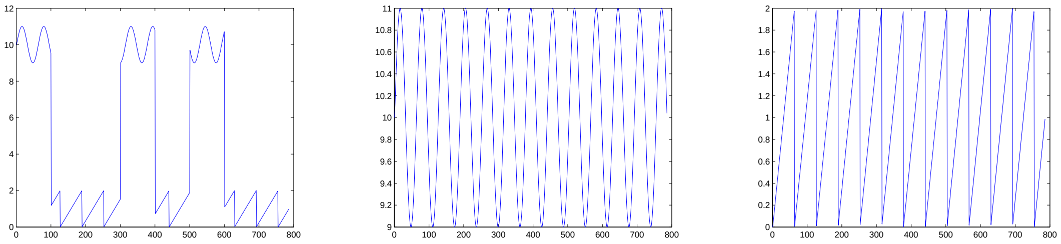
3.1 Introduction

The purposes of this work are 1) to propose a new correlation measurement algorithm based on regression called sectional correlation measurement, and 2) to find the best subset of variables which are related to the drought condition by using the sectional correlation measurement as the objective function. The second purpose also provides a contribution to the hydrological community. The hydrologists can focus only on the selected variables when they predict the future drought condition.

Usually, the correlation is measured between all pairs of two random variables. In some cases, the correlation happens only in some part/subset of the data. Many correlation measuring methods such as the Pearson correlation coefficient, Spearman's Rank, correlation coefficient, Jackknife correlation coefficient, and Semi-Partial Correlation, cannot find this type of correlation. For example, the Pearson correlation coefficient cannot find the partial correlation between time series A and B and also between time series A and C in Figure 3.1.

In time series data, correlation and similarity measurement can be seen as the same thing because their goal is to find if the two compared time series have the same shape or not. So this sectional correlation can be applied to solve problems in time series/sub-sequences comparison, and motif discovery.

This rest of the paper is organized into 4 sections. In Section 2, we introduce the data set used in the experiment and the problem description in Section 3. Section 3



(a) (b) (c)

Figure 3.1: Partial time series correlation example, (a) Time series-A, (b) Time series-B, (c) Time series-C and $\rho_{A,B}$ = 0.0991 and $\rho_{A,C}$ = -0.0574

summarizes three current problems which motivate the proposed correlation method and the related works in each problem. Then in Section 4, we describe the sectional correlation algorithm and also how to apply it to the feature selection. Section 5 provides the experimental results. The final section concludes this paper.

3.2 The OHD-NOAA Data Set

The data set used in this work is the hydrological data set from the Office of Hydrologic Development in the National Oceanic and Atmospheric Administration. This data set has 32 variables in high spatial resolution (4×4 km^2. per cell) and also in high temporal resolution (every 6 hours from January 2, 1979 to December 31, 2008). This data coverage is all the states in the US except Alaska and Hawaii.

Although the data covers all the US, this work focus only on the 10×10 cell^2 area around Arlington, Texas. We also use lower temporal resolution, which is reduced from 6-hours to daily, to reduce the computational cost. Another reason is that the HyVAS also uses this resolution to reduce data transferring and graphic rendering cost. Thus, in order to keep the system's consistency and just in case that this algorithm might be integrated to the HyVAS, this algorithm use the same resolution as the HyVAS.

Table 3.1: 32 Variables of the OHD-NOAA's hydrological data set

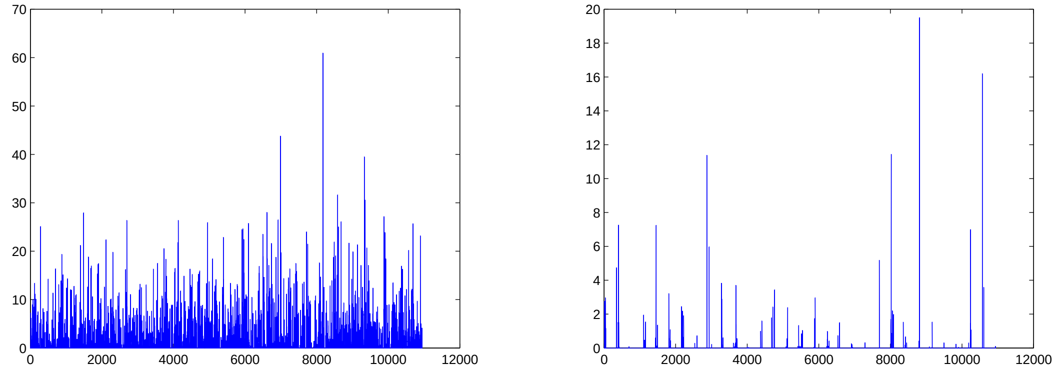
	Variable	Layer	Description
1	accmax	1	Maximum water equivalent since snow began to accumulate, mm
2	adimpc	1	Additional impervious area water content, mm
3	evap	1	Actual evapotranspiration, mm per dt
4	liqw	1	Liquid water storage, mm
5	lzfpc	1	Lower zone primary free water content, mm
6	lzfsc	1	Lower zone supplemental water content, mm
7	lztwc	1	Lower zone tension water content, mm
8	pevap	1	Potential evapotranspiration, mm per dt
9	rain	1	Rainfall forcing, mm per dt
10	rmlt	1	Rain plus melt dept, mm
11	runoff	1	Surface flow component, mm per dt
12-15	smliq	1-4	Unfrozen volumetric soil moisture at Noah defined layers
16	sndpt	1	Snow depth, mm
17	snow	1	Snowfall forcing, mm per dt
18	snowfrac	1	Snow cover fraction, dimensionless
19-22	soilm	1-4	Total volumetric soil moisture at Noah defined layers
23-26	soilt	1-4	Soil temperature at Noah defined layers
27	subflow	1	Subsurface flow component, mm per dt
28	swe	1	Snow water equivalent, mm
29	tem	1	Air temperature forcing, C
30	twe	1	Total water equivalent, mm
31	uzfwc	1	Upper zone free water content, mm
32	uztwc	1	Upper zone tension water content, mm

3.3 Current Problems

In this section, we illustrate three problems and their related works in correlation measurement and time series similarity measurement.

3.3.1 Correlation of a Random Variable to a Section of the Other Random Variable

We found that some time series in hydrology like the daily rainfall or snowfall time series, which we call event-time series in the rest of this paper, have unique characteristics as following,



(a) (b)

Figure 3.2: Examples of event-time series (a) Rain, and (b) Snow

1. Most of the data are zeros when the event does not occur, i.e. no rain or snow.

$$Mode(TS_i) = 0, \qquad \forall i \in \{1...\text{length of the time series}\} \tag{3.1}$$

   where TS is a time series

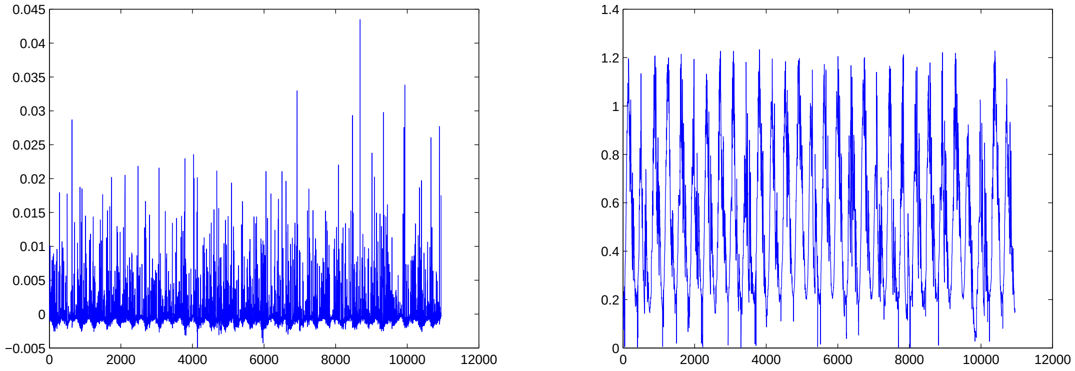
2. All values are non-negative because when event occurs, the value always more than zero. Thus the minimum value is zero.

$$Min(TS_i) = 0, \qquad \forall i \in \{1...\text{length of the time series}\} \tag{3.2}$$

   where TS is a time series

3. Each event occurs only in a short period of time comparing to the length of the time series. Each event starts when the value starts to be above zero and ends when the value becomes zero again. For example, see Figure 3.2, the average length of rainfall events are 3.11 days and the average length of snowfall events are 1.78 days.

This unique shape is worth being emphasized because one third of all variables are rainfall-like-time series. Besides, at the time that an event does not occur, the impact from other variables can be seen easier. From Figure 3.3(a) and 3.3(b), it



(a)

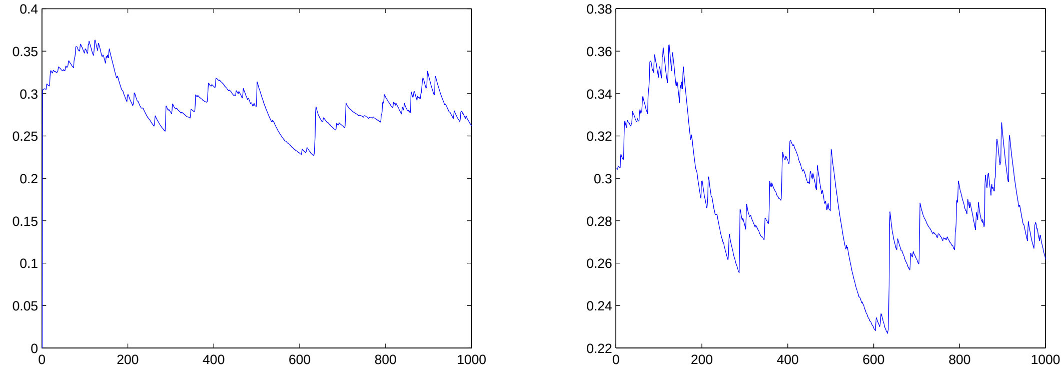
(b)

Figure 3.3: Example of partial time series correlation (a) Change of soil moisture, and (b) Evapotranspiration

can be seen that the bottom part of the change of soil moisture time series, the below 0 part, has a periodic pattern like the evapotranspiration time series. The Pearson correlation between these 2 time series is -0.0046, so it cannot capture this correlation because the impact from the event-time series is much higher.

3.3.2 Noise in Time Series or Sub-sequence Deters the Normalization Result

Also, the two most familiar similarity measurement methods are Euclidean distance-based and Dynamic Time Warping-based. The latter is more robust because it can find the best match (lowest distance between the 2 random variables) by repeating some data points. All pairs of data points in the compared time series or sub-sequences are considered. Before the similarity measurement can be done, two time series or sub-sequences must be normalized. Problem will occur if one of them has highly distant noises. These noises alter the normalization results and make these time series or sub sequences slightly difference in every pair of data points. As a result, both similarity measurement, Euclidean distance and DTW, appear to have



(a)

(b)

Figure 3.4: Example of the noise effect to the similarity measurement (a) The original time series, and (b) The time series after removing noise

high value which means these compared time series are not similar even though, these two without noises are totally similar.

To illustrate the noise problem above, we provide an example in Figure 3.4. The first point in the original time series can be considered as noise because it appears to be out of range from other data points. Then, another time series, as shown in Figure 3.4(b), is created by copying the original time series but the first point has the same value as the second point. The mean and the standard deviation of these time series are slightly different as shown in Table 3.2. After normalized both time series, the similarity between these two time series is measured. The result distance is quite high (9.8865) though, without noise, these time series appears to be similar.

Table 3.2: Mean and standard deviation comparison

Time Series	Mean	Std.
With noise	0.2898	0.0312
Without noise	0.2901	0.0298

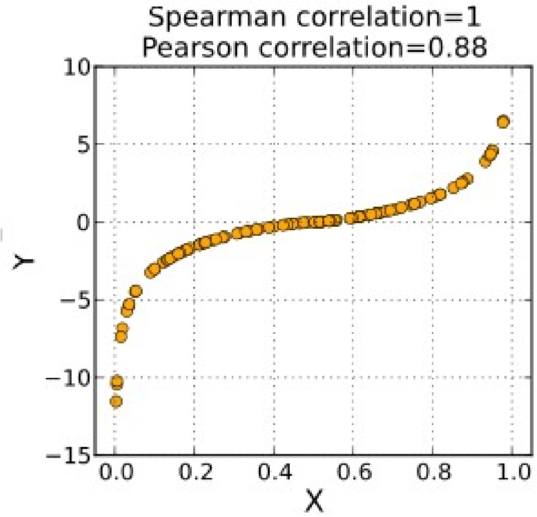


Figure 3.5: Example of the Spearman correlation and non-linear relationship

The Pearson correlation coefficient is the most popular correlation measurement because it has low computational cost and it is easy to understand. Obviously because of is simplicity, it can only detect the linear relationship between two variables. Also, it uses all data points so it cannot detect highly distant noise.

Another widely used correlation is the Spearman's rank correlation coefficient [10]. This method is not limited only to the linear relationship like the Pearson correlation coefficient. (See Figure ??) In the highly distant noise case, the Spearman correlation coefficient can detect the correlation better than the Pearson because it uses rank of data instead of real value. Although it seems better than the Pearson, it cannot determine which data point is the highly distant noise.

The other widely used rank correlation is the Kendall rank correlation coefficient (or Kendall's tau coefficient) [11]. This method can detect the sectional correlation because it measures the similarity of the data ordering. Unfortunately, it cannot detect the range of section. Also, it cannot find the highly distant noise.

The Jackknife correlation [12] seems to solve the highly distant noise by repeating the calculation of the correlation coefficient, each time removing one value, then averaging the results. This algorithm is only resistant to one highly distant noise. Another drawback is its high computational cost.

Another interesting correlation coefficient is the semi-partial correlation (also called part correlation) [13]. This correlation measurement is used to find the correlation between a variable and the other variable after the effect of the third has been removed. In the other word, we also focus on the correlation between two variables but one of them has an effect from the third variable which cannot be controlled in the experiment. For example, for students in a specific high school as an experimental group, we want to find the correlation between sex and the choice of college. The choice of college might also be affected by the students' SAT score. Thus, the SAT score variable must be removed from the choice of college variable before correlation calculation. This method is almost like what we need to find but the difference is that we focus on finding if the SAT score variable is partially (or sectional) correlated to the choice of college or not. As such, this correlation measurement cannot be applied to our work.

In summarize, we need another correlation algorithm that

- Able to find a sectional correlation, i.e. a correlation between a time series and a section of other time series
- Able to remove noise(s) from a time series for better result in similarity measurement

3.3.3 Drought Related Feature Selection

In this work, we also focus on how to find the best subset of variables that, in combination, can determine how the drought condition change in each day with minimized error. We cannot simply choose the set of variables based on its sectional correlation score. Some variables are correlated to the the same range of target time series.

Feature selection methods can be categorized into three groups, filter methods, wrapper methods, and embedded methods. In filter methods, each feature similarity or related score is computed before the learning process [14], [15]. Two popular filter scores are correlation [16], [17] and mutual information [18]. The drawback of this method is that it possible that some uncorrelated variables with help from other variable can be correlated to the target variable. Another disadvantage is it tends to choose redundant variables, i.e. it selects the two or more variables that are correlated to each other.

In wrapper methods [19], [20], the algorithm searches for possible subset first. Then each subset of features is evaluated based on the prediction accuracy. Even though the prediction performance is usually better than that of the filter methods, the major disadvantage of these methods is the computational cost is usually high. Another reason this method is not used in our work is that it is prone to over-fitting.

The last one is the embedded methods [21], [22]. In these methods, features are searched and selected in the process of model learning. Each model is then evaluated in cross-validation step. The model (feature subset and weight) with best evaluation score is then chosen. Our chosen method falls into this type. The reasons we choose this are 1) it is less prone to over-fitting, 2) it is less computational expensive than wrapper methods, and 3) we can applied our proposed sectional correlation to this method as the objective function, i.e. the function for evaluating how good a model is.

Mostly, feature selection methods are based on greedy algorithm which can be categorized into forward or backward. The difference is the number of features at the start of the process. The forward feature selection starts with an empty set of feature and one feature is added to the subset in each round. On the other hand, the backward feature selection starts with a full set of features and one feature is removed in each round, for example the SVM-RFE (Support Vector Machine based Recursive Feature Elimination Method) [23].

Branch and bound method [24] is also used to search for the best feature subset. This method guarantees the globally optimal result if the objective function or criterion satisfies monotonicity. Because of the monotonicity, the algorithm can reject some sub-optimal subsets without any actually evaluation. Unfortunately, our objective function is not monotonic which makes the results to be not optimal. Also, if the monotonic property is not hold, this algorithm still can be used but it has more computational cost than the greedy algorithm. Thus, we did not apply this method to our algorithm.

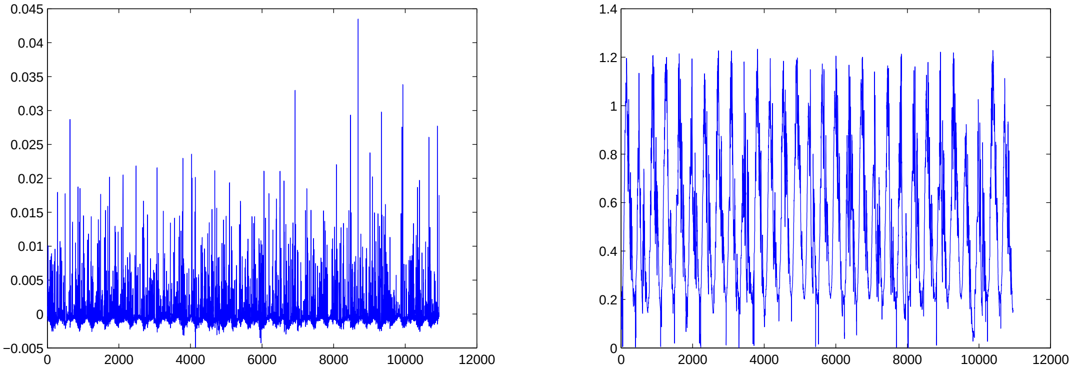
We choose the forward feature selection method because it has an advantage over the other in terms of computational cost. In backward feature selection, the larger size of subset must be processed. On the other hand, the forward feature selection starts with smaller size of subset.

In summarize, our method is the forward-embedded feature selection method. The objective function is proposed sectional correlation score.

3.4 Methodology

3.4.1 Sectional Correlation Measurement

The purpose is to find the correlation between two time series. Most of the correlation calculations reflect two-way relationship between the two compared time series while this proposed method reflects one-way relationship so we need to define



(a) (b)

Figure 3.6: Example of target and input time series used in the Methodology section (a) The change of soil moisture time series, and (b) The evapotranspiration time series

the two compared times series as the input time series and the target time series. The algorithm matches the input time series to a part/range of the target time series that causes the correlation coefficient to be the highest value. This algorithm then returns the Pearson correlation coefficient of that matched parts along with the upper and lower bound of the part/range of the target time series. We provide an example along with the algorithm description for better understanding of how this works.

The change of soil moisture time series, as shown in Figure 3.6(a), is used as the example target time series. This is created using

$$C_i = SM_i - SM_{i-1}, \tag{3.3}$$

where C is the change of soil moisture time series, SM is the soil moisture time series, and i = [1 ... length of the time series]

And the input time series is the evapotranspiration time series as shown in Figure 3.6(b). Visually, the evapotranspiration time series seems to be correlated to the bottom part of the change of soil moisture time series. This proposed algorithm will be used to find exact range and the correlation coefficient between that part of the target time series and the input time series. The algorithm processes as the following steps,

1. Two types of time series must be treated different based on its relevance.

   - For event time series, such as rainfall and snowfall, only the data points that have value more than zeros will be considered.

     $$considered\,index = iTS_i > 0, \;\; \forall i \in \{1...length\,of\,the\,time\,series\} \tag{3.4}$$

     where iTS is the input time series

   - For non-event time series, all data points should be considered.

     $$considered\,index = \forall i \in \{1...length\,of\,the\,time\,series\} \tag{3.5}$$

   The Evapotranspiration time series is a non-event time series so all data points will be used.

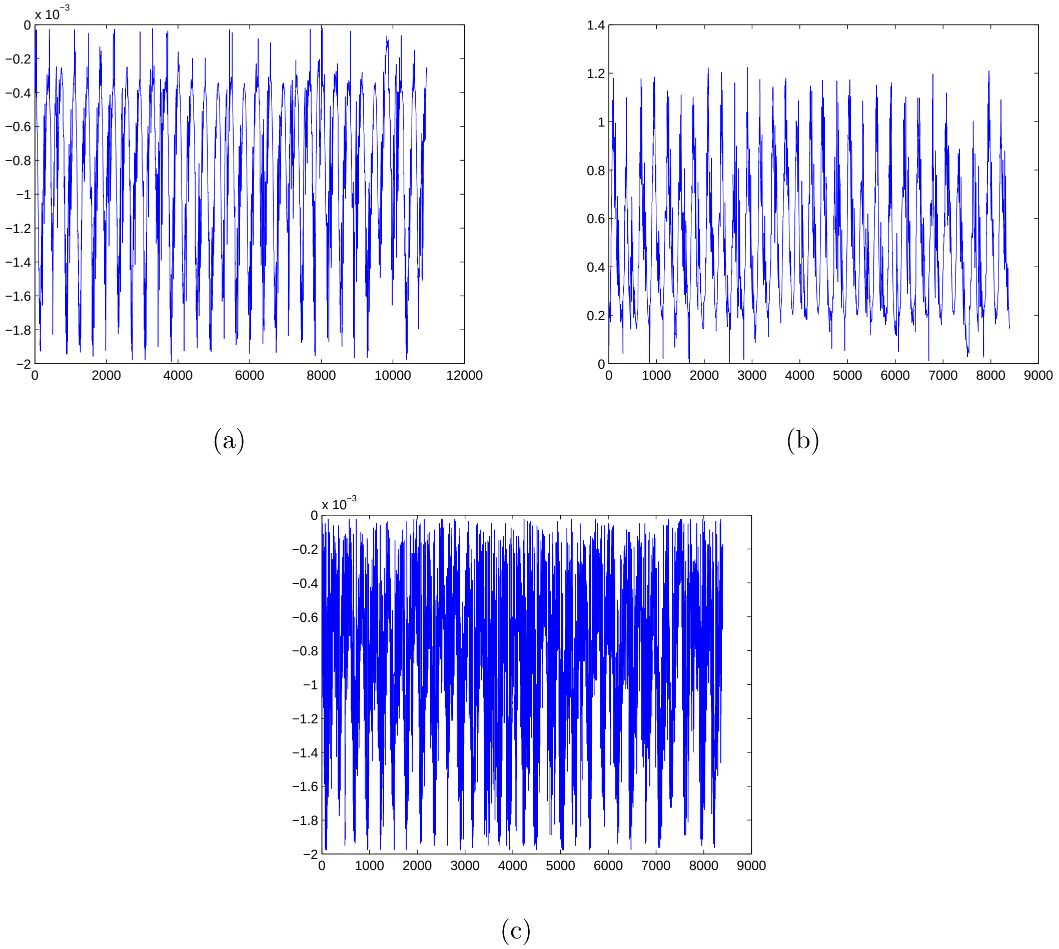
2. The Support Vector Regression is used to match the input time series to the target time series. The SVR training data is the input time series and the SVR target data is the target time series. Then, the SVR tries to predict the target time series using the input time series. Thus, the the prediction result can determine the range of the target time series that maximize the correlation coefficient.

3. The upper and lower bound of the correlated part of the target time series are calculated by

   $$upper\,bound = max(predition\,value) \tag{3.6}$$

   $$lower\,bound = min(predition\,value) \tag{3.7}$$

   The result range of the change of soil moisture value is [-0.0020 -0.000022].



(a)

(b)

(c)

Figure 3.7: The result from Support Vector Regression (a) The predict time series, (b) The correlated input time series, and (c) The correlated target time series

4. Calculate the section scale by

$$Section\ Scale = \frac{upper\ bound - lower\ bound}{max(target\ value) - min(target\ value)}, \tag{3.8}$$

5. Cut out some data points in the target time series that have value out of the correlated part.

$$correlated\ index = (TS_i \geq lower\ bound)\ \&\&\ (tTS_i \leq upper\ bound) \tag{3.9}$$

where $\forall\ i \in considered\, index$ from equation 3.4 or 3.5

and tTS is the target time series

$$correlated\, iTS = iTS_i, \ \ \forall i \in correlated\, index \tag{3.10}$$

where iTS is the input time series

$$correlated\, tTS = tTS_i, \ \ \forall i \in correlated\, index \tag{3.11}$$

where tTS is the target time series The correlation between evapotranspiration and change of soil moisture time series can be seen in Figure 3.7(b) and 3.7(c) respectively.

6. Calculate the Pearson correlation score from those data point.

$$\rho_{PCC-Section} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}, \tag{3.12}$$

where where ρ is the Pearson Coefficient Coefficient, X = $correlated\, iTS$ and Y = $correlated\, tTS$

7. Calculate the sectional correlation by

$$SectionalCorrelation = \rho_{Section} \times Section\, Scale, \tag{3.13}$$

8. Finally, the algorithm returns the Sectional Correlation, upper bound, and lower bound.

3.4.2 Feature Selection: Using Partial Correlation as the Objective Function

The purpose of this algorithm is to find a subset of variables that related to the target variables. The key idea is using the forward feature selection method which objective function is the sectional correlation score to find the set of features/variables which can be combined to match the target variable. We also use 10-fold cross-validation to avoid over-fitting problem. The algorithm can be breakdown into n steps as following,

1. Split the data into training and testing data
2. Split the training data into 10 parts. Use the first to ninth parts for training and the tenth part for validating
3. Calculate sectional correlation score of each variable to the target variable. Keep the variable with the highest score to the pool of selected variables.
4. Repeat the step 3 but add all the previous selected variables to the sectional correlation calculation. Stop the repetition when the sectional correlation score stops increasing.
5. Calculate the sectional correlation score of the chosen variables to the target variable using the validating data.
6. Repeat the step 3-5 nine times using the ninth, eighth, seventh, and so on for validating and the rest for training.
7. Choose the subset of variables that has the highest sectional correlation score.
8. Test the selected subset by using Support Vector Machine to regress the step 1 training data of the chosen variables to the target variable.
9. Use the regression model to construct a target variable from the step 1 testing data of the chosen variables.
10. Calculate the Root Mean Square Error (RMSE) between the regression result and the testing data.

3.5 Experimental Results

3.5.1 Sectional Correlation of 32 OHD-NOAA Variables to Drought Indices

The proposed algorithm is applied to find how much the 32 OHD-NOAA variables correlated to the change of other 6 variables, average soil moisture (layer 1-4) and five drought indices, which are the soil moisture anomaly (SMA), Standard Precipitation Index (SPI) 3, 6, 9, and 12 months. The calculation of the SMA and SPI can be found in [25] and [26] respectively. The algorithm tells 1) sectional correlation score, 2) the section, i.e. range of the target variable, that is the most similar to the variable, 3) the Pearson correlation coefficient between the the compare variables to the detected section of the target variable. We present the result for all the OHD-NOAA variables sorted from the highest sectional correlation score to the lowest.

As in Table 3.3 and 3.4, the result orders of variables are not the same. The reason behinds this is that non-event typed data are more similar to the section [-0.0011, 0] which is a small range compared to the range of the change of soil moisture. If the Pearson correlation coefficient between this section of the soil moisture change and the variable s really high, such as evap, pevap, and tem, the sectional correlation will be high. The result shows that this algorithm can solve the sectional correlation problem by finding that evap has high correlation to the [-0.0020, 0.000019] section of soil moisture change.

The results below obviously show that the increasing of all the drought indices is highly correlated to liqw, rain, rmlt, runoff, subflow, and uzfwc. On the other hand, the variables that related to the decreasing of drought indices are not obvious.

Table 3.3: The sectional correlation between Averaged Soil Moisture Change and the 32 variables in descending order

Variable	Sectional Corr.	Section	Pearson Corr. (Section)
rain	0.616600	-0.000787 to 0.041529	0.704512
uzfwc	0.453488	-0.000666 to 0.033339	0.644780
rmlt	0.348147	-0.000714 to 0.033507	0.491885
subflow	0.254435	-0.000842 to 0.019655	0.600197
runoff	0.034703	-0.000625 to 0.015308	0.105308
evap	-0.030911	-0.001992 to -0.000018	-0.757297
pevap	-0.012105	-0.001298 to -0.000091	-0.484697
soilt1	-0.006736	-0.001066 to -0.000008	-0.307758
soilt3	-0.006704	-0.001016 to 0.000139	-0.280572
lzfpc	-0.006499	-0.001327 to 0.000018	-0.233539
soilt2	-0.006337	-0.001043 to -0.000087	-0.320498
soilt4	-0.004883	-0.001013 to 0.001733	-0.085967
soilm1	0.004316	-0.001423 to 0.000149	0.132765
smliq4	0.004194	-0.001057 to 0.002241	0.061487
soilm4	0.004194	-0.001057 to 0.002241	0.061487
smliq1	0.003925	-0.001406 to 0.000134	0.123276
snow	0.003280	-0.000601 to 0.006008	0.023991
liqw	0.003177	-0.000622 to 0.007505	0.018900
accmax	0.003115	-0.000625 to 0.002814	0.043794
smliq3	-0.002744	-0.001067 to 0.000019	-0.122137
soilm3	-0.002710	-0.001061 to 0.000023	-0.120958
tem	-0.002525	-0.001068 to 0.005706	-0.018023
snowfrac	0.002246	-0.000628 to 0.000131	0.143049
swe	0.002180	-0.000612 to 0.003630	0.024844
twe	0.002158	-0.000604 to 0.003757	0.023925
sndpt	0.001915	-0.000606 to 0.002384	0.030961
uztwc	0.001606	-0.001050 to -0.000359	0.112392
lztwc	-0.000991	-0.000965 to -0.000300	-0.071986
lzfsc	-0.000480	-0.000987 to -0.000602	-0.060224
adimpc	0.000256	-0.000719 to -0.000516	0.060970
soilm2	-0.000227	-0.000869 to -0.000296	-0.019176
smliq2	-0.000211	-0.000861 to -0.000290	-0.017890

Table 3.4: The Pearson correlation coefficient between Averaged Soil Moisture Change and the 32 variables in descending order

Variable	Pearson Corr.
rain	0.7202
uzfwc	0.6340
subflow	0.5784
rmlt	0.5205
smliq1	0.3809
soilm1	0.3806
lzfsc	0.3232
uztwc	0.3196
runoff	0.1199
pevap	-0.1190
tem	-0.1035
smliq2	0.0903
soilm2	0.0902
soilt1	-0.0844
adimpc	0.0784
soilt4	-0.0778
soilt3	-0.0775
soilt2	-0.0753
accmax	0.0377
liqw	0.0325
snow	0.0324
snowfrac	0.0308
lzfpc	-0.0259
lztwc	-0.0247
twe	0.0239
swe	0.0232
sndpt	0.0199
soilm3	-0.0170
smliq3	-0.0170
Smliq4	-0.0168
soilm4	-0.0168
evap	-0.0046

Table 3.5: The sectional correlation between Soil Moisture Anomaly (SMA) Change and the 32 variables in descending order

Variable	Sectional Corr.	Section	Pearson Corr. (Section)
rain	0.191556	-0.007169 to 0.320463	0.572737
rmlt	0.123601	-0.006195 to 0.273381	0.433080
uzfwc	0.077683	-0.005509 to 0.189988	0.389252
subflow	0.028043	-0.006572 to 0.090444	0.283157
runoff	0.026271	-0.005188 to 0.204315	0.122837
evap	-0.004374	-0.013031 to -0.000597	-0.344606
pevap	-0.002191	-0.008989 to -0.000752	-0.260532
soilt1	-0.001673	-0.007643 to -0.000855	-0.241447
liqw	0.001430	-0.005206 to 0.037460	0.032820
soilt2	-0.001377	-0.007718 to -0.001325	-0.211081
tem	-0.001376	-0.007873 to 0.000311	-0.164659
soilm1	0.001119	-0.010015 to 0.003034	0.084015
smliq1	0.001112	-0.009928 to 0.003181	0.083113
soilt3	-0.000994	-0.007342 to -0.001537	-0.167769
snow	0.000959	-0.004599 to 0.025468	0.031257
soilt4	-0.000898	-0.007498 to -0.001612	-0.149449
accmax	0.000660	-0.004679 to 0.009958	0.044139
lzfsc	0.000373	-0.005596 to 0.008065	0.026729
sndpt	0.000332	-0.004686 to 0.009947	0.022233
smliq4	-0.000216	-0.006832 to -0.003934	-0.072845
soilm4	-0.000216	-0.006832 to -0.003934	-0.072845
lzfpc	-0.000206	-0.007885 to -0.003579	-0.046885
soilm3	-0.000199	-0.006966 to -0.003954	-0.064731
lztwc	-0.000198	-0.006609 to -0.003987	-0.074149
twe	0.000191	-0.005039 to 0.016850	0.008568
smliq3	-0.000189	-0.006925 to -0.003929	-0.061691
swe	0.000163	-0.004618 to 0.014894	0.008172
snowfrac	0.000148	-0.004860 to -0.000432	0.032757
uztwc	0.000116	-0.009689 to -0.002265	0.015309
smliq2	-0.000081	-0.006056 to -0.004684	-0.058112
adimpc	-0.000061	-0.005907 to -0.004533	-0.043200
soilm2	-0.000054	-0.005934 to -0.004727	-0.043443

Table 3.6: The sectional correlation between Standard Precipitation Index-3 Months (SPI3) Change and the 32 variables in descending order

Variable	Sectional Corr.	Section	Pearson Corr. (Section)
smliq1	0.001184	-0.044666 to 0.025251	0.206464
soilm1	0.001167	-0.043389 to 0.026321	0.204064
lzfsc	-0.000503	-0.113098 to 0.010251	-0.049713
evap	-0.000445	-0.024685 to 0.018285	-0.126130
uztwc	0.000390	-0.012467 to 0.020024	0.146311
soilm2	0.000318	-0.019462 to 0.016714	0.107177
rain	0.000242	-0.084454 to -0.000189	0.035064
adimpc	0.000230	-0.011558 to 0.017245	0.097476
smliq2	0.000216	-0.017765 to 0.014513	0.081430
rmlt	0.000142	-0.079751 to 0.006073	0.020140
subflow	-0.000106	-0.133745 to 0.004728	-0.009305
uzfwc	-0.000091	-0.075682 to 0.007675	-0.013274
accmax	0.000082	0.005131 to 0.049675	0.022318
tem	-0.000074	-0.017383 to 0.010009	-0.032791
snow	0.000059	-0.007034 to 0.007596	0.049052
twe	0.000047	-0.000279 to 0.005690	0.096248
swe	0.000045	-0.004708 to 0.007787	0.043823
sndpt	-0.000026	-0.007037 to 0.007341	-0.022338
runoff	-0.000022	0.000275 to 0.001858	-0.169761
pevap	-0.000022	-0.002480 to 0.005464	-0.033393
snowfrac	0.000020	-0.001599 to 0.015525	0.014075
liqw	0.000017	-0.004915 to 0.007261	0.017280
soilt2	-0.000012	-0.009930 to 0.007750	-0.008373
lzfpc	-0.000008	-0.008252 to 0.004877	-0.007452
soilt4	-0.000008	-0.010186 to 0.007980	-0.005101
lztwc	0.000007	-0.006417 to 0.008814	0.005597
smliq3	0.000006	-0.007864 to 0.009230	0.003967
soilt3	-0.000005	-0.009641 to 0.006205	-0.003666
soilm3	0.000005	-0.009017 to 0.008750	0.003176
smliq4	0.000004	-0.009125 to 0.008655	0.002699
soilm4	0.000004	-0.009125 to 0.008655	0.002699
soilt1	-0.000001	-0.014970 to 0.009738	-0.000445

Table 3.7: The sectional correlation between Standard Precipitation Index-6 Months (SPI6) Change and the 32 variables in descending order

Variable	Sectional Corr.	Section	Pearson Corr. (Section)
evap	-0.000513	-0.014620 to 0.009245	-0.218319
soilm1	0.000403	-0.022067 to 0.011630	0.121586
smliq1	0.000391	-0.021283 to 0.010848	0.123603
soilm2	0.000386	-0.014371 to 0.008279	0.173056
smliq2	0.000385	-0.013256 to 0.008938	0.176402
lzfsc	-0.000377	-0.064102 to 0.005038	-0.055385
adimpc	0.000267	-0.009533 to 0.007962	0.155166
rain	0.000201	-0.044527 to -0.000089	0.045963
rmlt	0.000191	-0.041310 to 0.001251	0.045504
smliq3	0.000186	-0.012779 to 0.005653	0.102512
lzfpc	-0.000183	-0.013599 to 0.006556	-0.092379
soilm3	0.000183	-0.012494 to 0.005938	0.100863
smliq4	0.000174	-0.011195 to 0.007045	0.096839
soilm4	0.000174	-0.011195 to 0.007045	0.096839
uzfwc	-0.000160	-0.042310 to 0.002821	-0.036140
subflow	-0.000150	-0.059057 to 0.001465	-0.025130
lztwc	0.000087	-0.006833 to 0.005254	0.072801
runoff	-0.000057	-0.004286 to 0.016172	-0.028286
pevap	-0.000051	-0.003040 to 0.002864	-0.086958
liqw	0.000041	0.003416 to 0.010921	0.055726
snow	-0.000040	-0.008057 to 0.005448	-0.030010
swe	0.000029	0.003404 to 0.013956	0.028013
uztwc	-0.000026	-0.006002 to 0.006191	-0.021887
twe	0.000022	-0.000644 to 0.012327	0.016917
sndpt	0.000020	0.003140 to 0.016190	0.015255
soilt3	-0.000014	-0.001544 to 0.000268	-0.080020
snowfrac	-0.000008	0.002215 to 0.003589	-0.060510
accmax	0.000008	0.004149 to 0.013910	0.007983
tem	-0.000005	-0.003516 to 0.000995	-0.010139
soilt2	-0.000002	-0.003642 to 0.000688	-0.005522
soilt1	-0.000002	-0.002241 to 0.001147	-0.006029
soilt4	-0.000002	-0.001152 to -0.000867	-0.063009

Table 3.8: The sectional correlation between Standard Precipitation Index-9 Months (SPI9) Change and the 32 variables in descending order

Variable	Sectional Corr.	Section	Pearson Corr. (Section)
evap	-0.000540	-0.013133 to 0.008252	-0.218330
rain	0.000329	-0.048778 to 0.004502	0.053463
lzfsc	-0.000262	-0.028119 to 0.000736	-0.078404
smliq1	0.000231	-0.014764 to 0.007941	0.087838
soilm1	0.000216	-0.014132 to 0.008573	0.082107
smliq2	0.000211	-0.011272 to 0.007639	0.096639
uzfwc	-0.000162	-0.025846 to 0.000092	-0.053927
subflow	-0.000139	-0.031726 to -0.001210	-0.039273
soilm2	0.000124	-0.008813 to 0.005616	0.074577
rmlt	0.000094	-0.019743 to -0.001050	0.043719
accmax	0.000089	0.004569 to 0.050182	0.016853
liqw	0.000063	0.000775 to 0.013172	0.044202
swe	0.000062	0.004234 to 0.013456	0.057991
twe	0.000059	0.004625 to 0.013530	0.057539
soilm3	0.000055	-0.004962 to 0.001415	0.075150
pevap	-0.000055	-0.002755 to 0.003528	-0.075796
smliq4	0.000052	-0.005704 to 0.000660	0.071201
soilm4	0.000052	-0.005704 to 0.000660	0.071201
soilt2	-0.000046	-0.002671 to 0.004510	-0.055272
lztwc	0.000045	-0.003291 to 0.001550	0.080996
sndpt	0.000042	0.002444 to 0.017639	0.024147
runoff	-0.000032	-0.002798 to 0.008540	-0.024356
soilt4	-0.000022	-0.001309 to 0.002225	-0.054160
uztwc	0.000019	-0.004118 to 0.003999	0.020577
soilt3	-0.000018	-0.001908 to 0.002776	-0.032909
snow	0.000017	0.000529 to 0.005610	0.029044
soilt1	-0.000017	-0.002203 to 0.004252	-0.022626
tem	-0.000015	-0.000630 to 0.000347	-0.130390
adimpc	0.000014	-0.004885 to 0.004383	0.013296
lzfpc	-0.000010	-0.004895 to 0.004591	-0.009511
snowfrac	0.000003	0.001413 to 0.006940	0.005426
smliq3	0.000001	-0.003678 to 0.002611	0.001637

Table 3.9: The sectional correlation between Standard Precipitation Index-12 Months (SPI12) Change and the 32 variables in descending order

Variable	Sectional Corr.	Section	Pearson Corr. (Section)
smliq1	0.000431	-0.012397 to 0.006616	0.185879
soilm1	0.000420	-0.012450 to 0.006562	0.180795
lzfsc	-0.000278	-0.024245 to 0.001519	-0.088522
smliq2	0.000236	-0.009408 to 0.003897	0.145220
soilm2	0.000233	-0.009031 to 0.004275	0.143491
subflow	-0.000219	-0.048983 to 0.002044	-0.035137
lzfpc	-0.000176	-0.008378 to 0.003599	-0.120168
smliq4	0.000165	-0.007611 to 0.002744	0.130825
soilm4	0.000165	-0.007611 to 0.002744	0.130825
rain	0.000138	-0.035433 to 0.002190	0.030025
smliq3	0.000133	-0.007251 to 0.003153	0.104578
adimpc	0.000133	-0.005523 to 0.003877	0.115462
soilm3	0.000111	-0.006750 to 0.004316	0.082361
evap	-0.000106	-0.007073 to 0.004669	-0.073702
rmlt	0.000090	-0.019916 to -0.001438	0.039713
accmax	0.000062	0.002541 to 0.027747	0.020036
soilt1	-0.000061	-0.005380 to 0.001889	-0.068365
tem	-0.000057	-0.006178 to 0.001823	-0.058012
soilt2	-0.000050	-0.003191 to 0.001731	-0.083785
uzfwc	0.000045	-0.024914 to -0.001673	0.015909
uztwc	0.000037	-0.003100 to 0.000942	0.074033
soilt3	-0.000029	-0.002057 to 0.000947	-0.079545
soilt4	-0.000023	-0.002243 to 0.001956	-0.045390
runoff	-0.000023	-0.000592 to 0.005469	-0.031013
lztwc	0.000020	-0.004023 to 0.001402	0.030734
snow	0.000019	-0.014456 to 0.003948	0.008340
sndpt	0.000017	0.004718 to 0.005980	0.109514
twe	0.000017	0.004442 to 0.005956	0.090287
swe	0.000003	0.004206 to 0.009591	0.005180
pevap	-0.000003	-0.002209 to -0.000764	-0.015618
snowfrac	0.000000	0.002802 to 0.003448	0.003679
liqw	0.000000	0.004095 to 0.004129	0.000000

3.5.2 Feature Selection

In this work, we focus on finding which variables related to change of the drought indices which are used to determine the drought condition. Thus, the target variables used in the experiment are the change of drought indices.

The result shows that rain, smliq1 and soilm1 obviously play the key role in mitigating the drought condition, i.e. it is related to the increase of the drought index. Evap is the key factor in decrease of most the drought index. Other important decreasing factors are smliq-n, soilm-n, and soilt-n where n is the layer number from 1 to 4.

Table 3.10: The selected variables and their weights that related to the Averaged Soil Moisture Change

Selected Variable	Weight
rain	0.9879
evap	-0.0438
rmlt	-0.0291
swe	0.0162

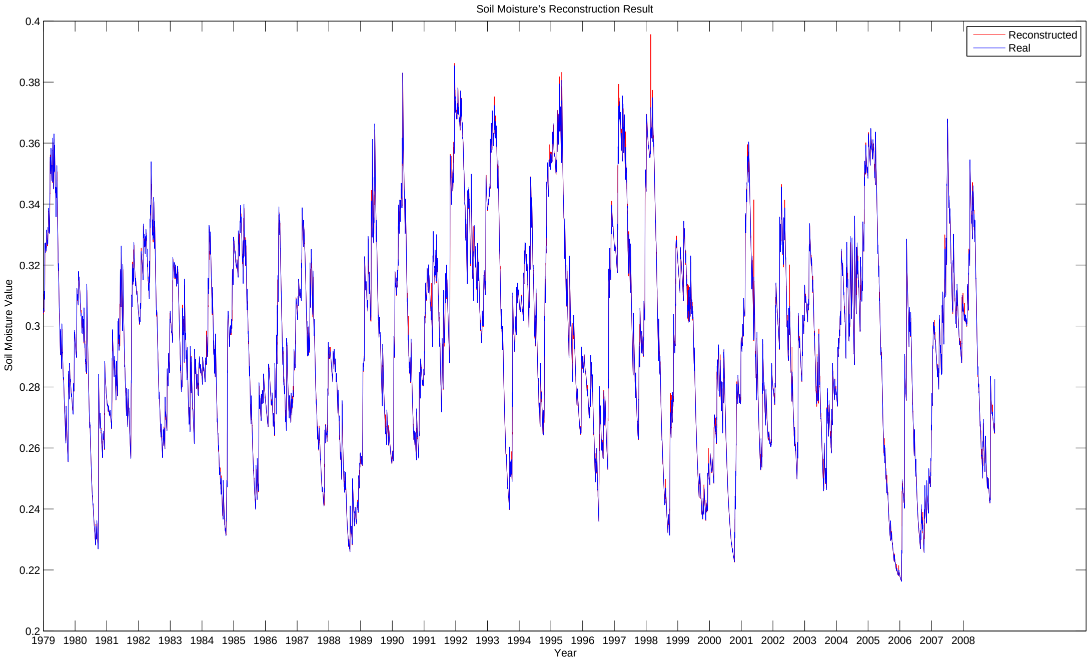


Figure 3.8: Soil moisture's reconstruction result

Table 3.11: The selected variables and their weights that related to the Soil Moisture Anomaly (SMA) Change

Selected Variable	Weight
rain	0.3364
evap	-0.0144
subflow	0.0539
pevap	0.0022
smliq3	-0.2078
smliq4	0.2033
liqw	0.0233
twe	0.0021

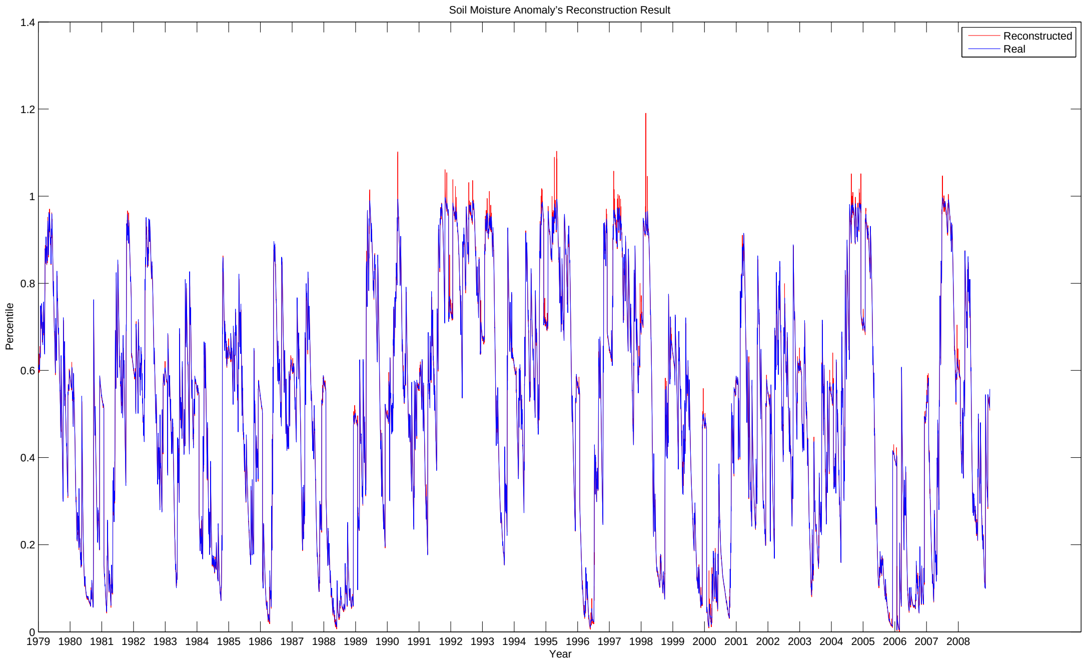


Figure 3.9: Soil moisture anomaly's reconstruction result

Table 3.12: The selected variables and their weights that related to the Standard Precipitation Index-3 Months (SPI3) Change

Selected Variable	Weight
smliq1	0.0057
twe	0.0739
lzfsc	-0.0034
liqw	0.0212
snow	0.0025
adimpc	0.0013
sndpt	-0.0473

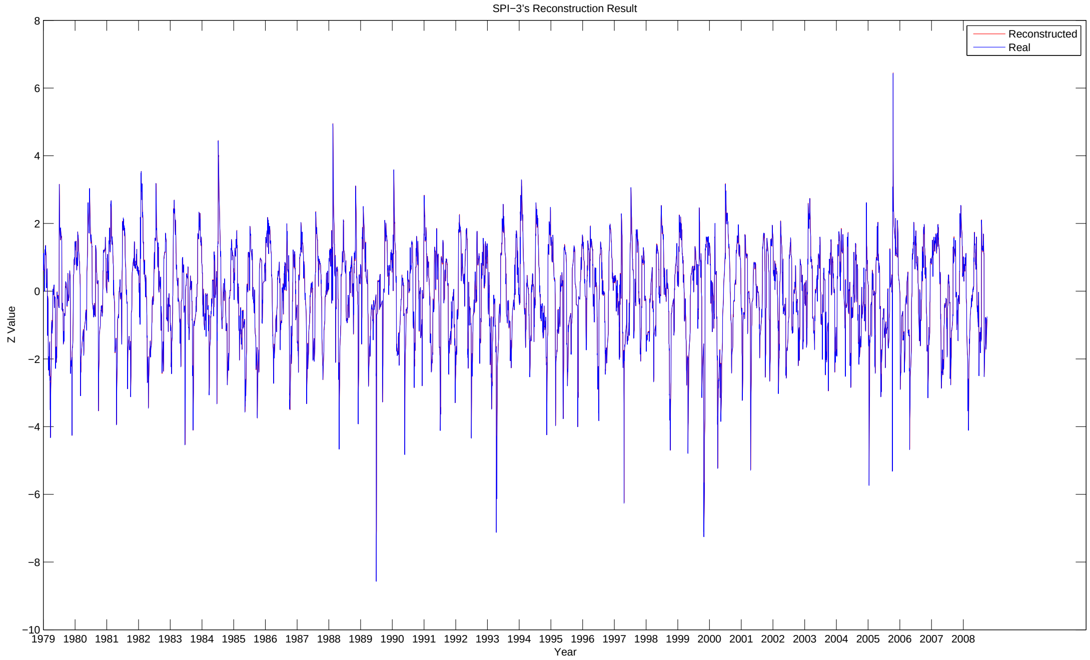


Figure 3.10: SPI-3's reconstruction result

Table 3.13: The selected variables and their weights that related to the Standard Precipitation Index-6 Months (SPI6) Change

Selected Variable	Weight
Lzfsc	-0.0013
Liqw	0.0095
Smliq1	0.0020
Swe	0.0338
Rmlt	0.0004
Evap	-0.0019
Sndpt	-0.0232

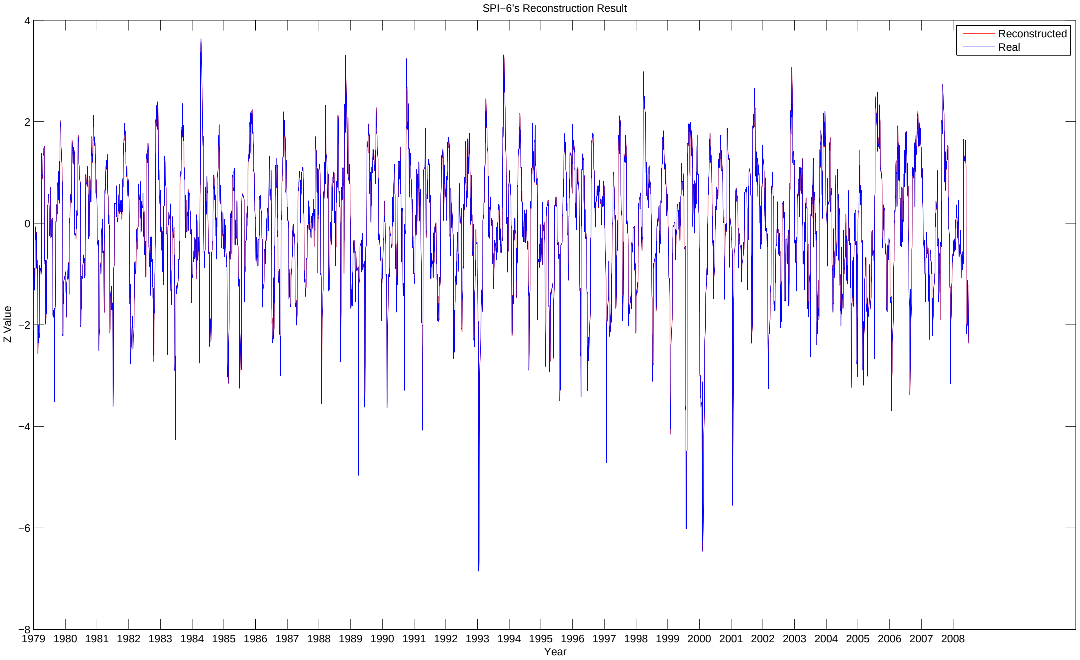


Figure 3.11: SPI-6's reconstruction result

Table 3.14: The selected variables and their weights that related to the Standard Precipitation Index-9 Months (SPI9) Change

Selected Variable	Weight
Evap	-0.0023
Sndpt	-0.0028
Subflow	0.0030
Accmax	0.0074
Soilt1	-0.0005

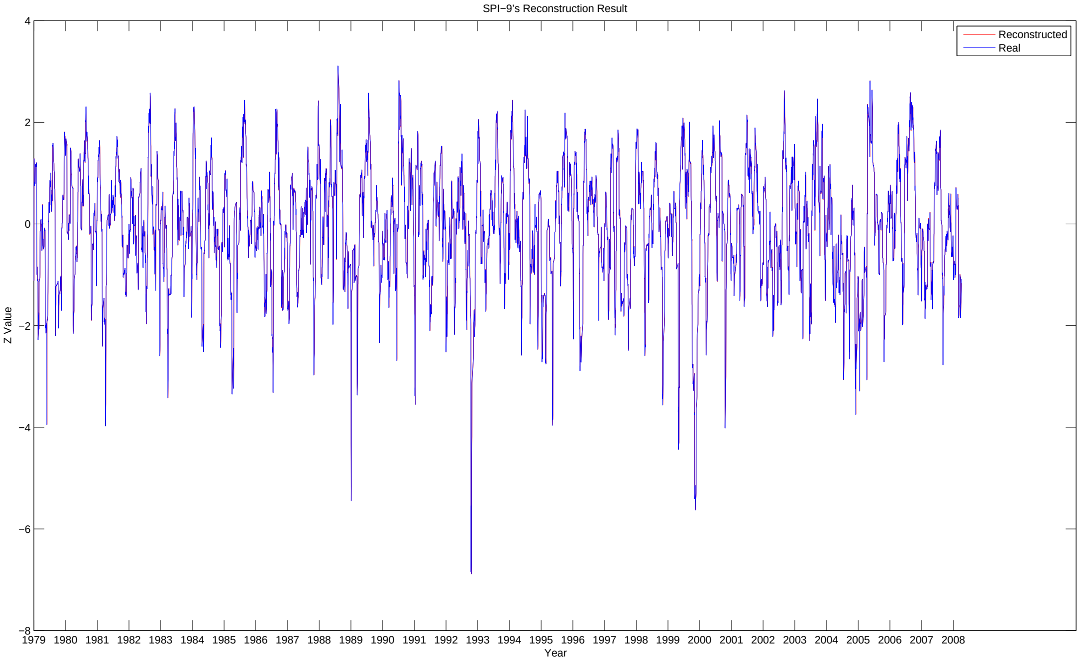


Figure 3.12: SPI-9's reconstruction result

Table 3.15: The selected variables and their weights that related to the Standard Precipitation Index-12 Months (SPI12) Change

Selected Variable	Weight
Soilm1	0.0025
Twe	0.0053
Soilt4	-0.0009
Lzfsc	-0.0013
Accmax	0.0073

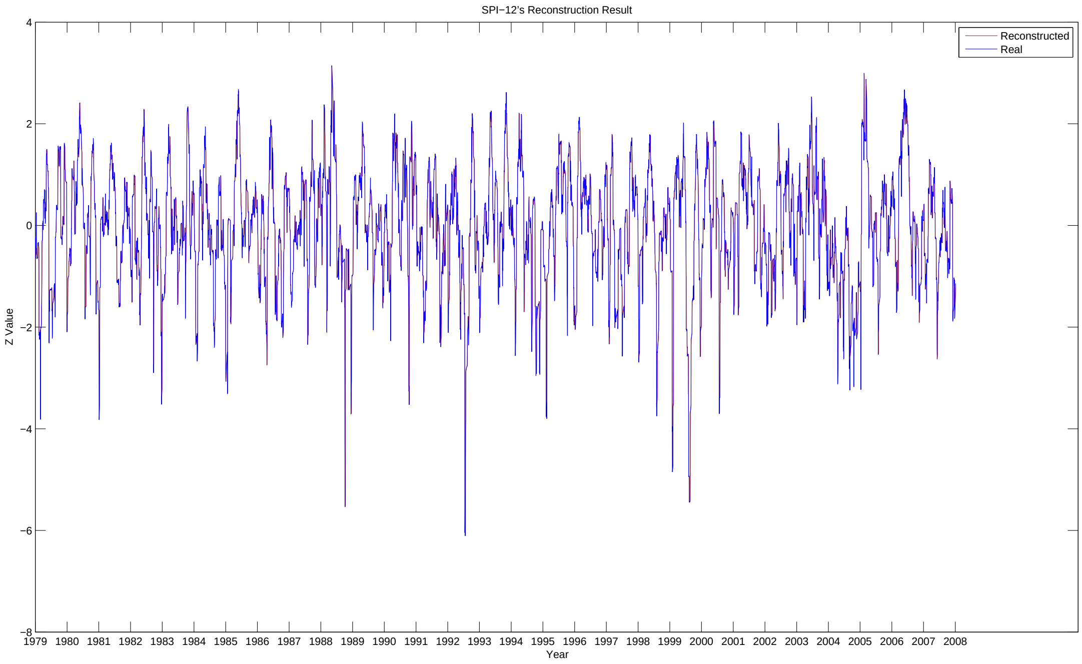


Figure 3.13: SPI-12's reconstruction result

Then, the drought indices are re-calculated using its previous value and the regression of the selected variables to the drought indices change. We use the Support Vector Regression in the regression process.

$$Reconstructed\,TS_i = Original\,TS_{i-1} + Regression\,result_i, \quad \forall i \in I \tag{3.14}$$

where TS is a time series and $I = \{1 \,...\, \text{length of the time series}\}$.

Table 3.16: The target time series reconstruction result

Target Variable	RMSE	Range of Value	Error Compared to Range(%)
Soil moisture	0.0019	0.2 to 0.4	0.95
SMA	0.0270	0 to 1	2.70
SPI-3	0.3620	-8.6 to 6.4	2.41
SPI-6	0.2812	-6.9 to 3.6	2.68
SPI-9	0.2550	-6.9 to 3.1	2.55
SPI-12	0.2471	-6.1 to 3.1	2.69

The result in Table 3.16 shows that the regression of the variables from the feature selection to the drought indices change, in combination with the previous drought index value, can be used to calculate the current drought index value with small error. Thus, these chosen variables are key variables to determine how and how much the drought index change from the previous value.

3.6 Conclusion

We developed a new correlation measurement algorithm that can find a sectional correlation and also helps identifying highly distant noises in data. This algorithm finds the similar section of the target data and the compared data by regression which we choose the Support Vector Regression to do the job. This sectional correlation algorithm can be applied to time series motif discovery to create a robust any-length motif discovery algorithm.

In addition, this sectional correlation is used as an objective function in the forward feature selection to find the variables in OHD-NOAA hydrological data set that are related to how the drought condition change from one day to the next. This feature will be further analyzed and might be useful in drought prediction later.

APPENDIX A

DROUGHT INDICES CALCULATION

In this appendix, the calculation of two drought indices used in this work are presented.

A.1 Soil Moisture Anomaly's Calculation

In this method, the factor that used to determine drought is the soil moisture. The idea is to compare the given soil moisture with the historical data and use its position in the distribution to determine how severe the drought is. The calculation procedure is the following steps.

1. From 30 years historical soil moisture data, reshape it into a matrix as in Figure A.1.
2. Create 365 distributions for each Julian day by collecting data of the same Julian day in 30 years.
3. To have more data for creating distribution, time window is used. In this work, time window size is 49 days which means the data that falls in between ± 24 days of that Julian day are used to create the distribution. For example, the data for creating a distribution of the day-50 is as in Figure A.2.
4. If the drought condition of the day-x is interested, put the soil moisture of that day into the according distribution. Then use its percentile position to

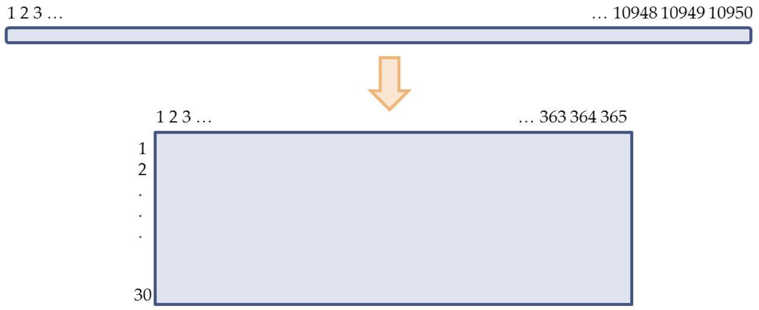


Figure A.1: Reshape the historical soil moisture data

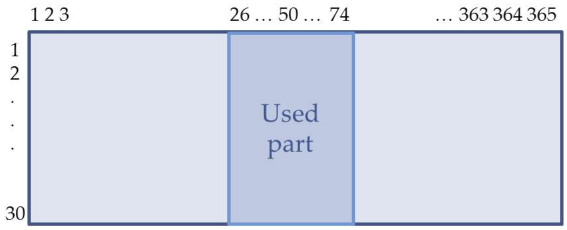


Figure A.2: Data used for creating the distribution of day-50

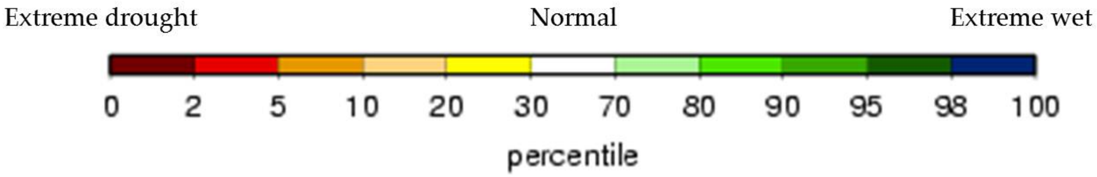


Figure A.3: How to determine the drought severity from the percentile

determine the drought severity. See Figure A.3 for ranges of percentile value and their drought severity.

A.2 Standard Precipitation Index's Calculation

In this method, the factor that used to determine drought is the rainfall value. The idea is to compare the given rainfall value with the previous n-months data where n in this work are 3, 6, 9, and 12. Then use its position in the distribution to determine how severe the drought is. The calculation procedure is the following steps.

1. Create distributions for each day in the data set from the previous $30\times n$ days rainfall data. For example, the data for creating a distribution of the day-301 is as in Figure A.4.

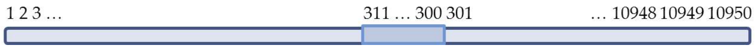


Figure A.4: Data used for creating the distribution of day-301

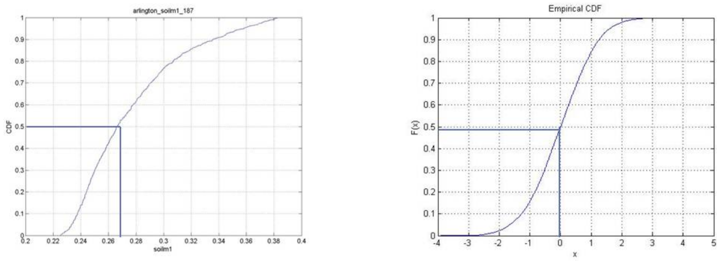


Figure A.5: Example of CDF Matching method

exceptionally dry	extremely dry	severely dry	moderately dry	abnormally dry	near normal	abnormally moist	moderately moist	very moist	extremely moist	exceptionally moist
⬛	🟫	🟥	🟧	🟨	⬜	🟩	🟩	🟦	🟦	🟪
-2.00 and below	-1.99 to -1.60	-1.59 to -1.30	-1.29 to -0.80	-0.79 to -0.51	-0.50 to +0.50	+0.51 to +0.79	+0.80 to +1.29	+1.30 to +1.59	+1.60 to +1.99	+2.00 and above

Figure A.6: How to determine the drought severity from the z value

2. The z value in the standard normal distribution is needed to determine the drought condition but the result rainfall distribution is usually not a normal distribution. The CDF matching method must be used to map from a rainfall value to z value. For example, the rainfall value 0.267 can be mapped to CDF value of 0.5. The the CDF value is used to map to z value by inverse normal distribution. The result z value is 0. (See Figure A.5)
3. Then the z value is available to determine the drought severity as in the Figure A.6.

REFERENCES

[1] (2012) Hydrodesktop - cuahsi hydrologic information system desktop application. The U.S. National Science Foundation supported Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI). [Online]. Available: http://hydrodesktop.codeplex.com/

[2] (2012) Arcgis - mapping and spatial analysis for understanding our world. Esri. [Online]. Available: http://www.esri.com/software/arcgis/arcgis10/index.html

[3] (2011) Gishydro: A gis-based hydrologic modeling tool. The Department of Civil and Environmental Engineering and the Maryland State Highway Administration, Office of Structures (OOS). [Online]. Available: http://www.gishydro.umd.edu/

[4] Hrap.c. The ARM External Data Center (XDC) at Brookhaven National Laboratory. [Online]. Available: http://www.xdc.arm.gov/xds/abrfc/hrap.c

[5] (2012) Php: Hypertext preprocessor. [Online]. Available: http://www.php.net/

[6] (2012) Html5 tutorial. [Online]. Available: http://www.w3schools.com/html5/

[7] (2012) Html5 canvas tutorial. [Online]. Available: http://www.w3schools.com/html5/html5_canvas.asp

[8] (2012) Webgl - opengl es 2.0 for the web. Khronos Group. [Online]. Available: http://www.khronos.org/webgl/

[9] (2012) Three.js - documentation. [Online]. Available: http://mrdoob.github.com/three.js/docs/49/

[10] C. Spearman, “Correlation calculated from faulty data,” *British Journal of Psychology, 1904-1920*, vol. 3, pp. 271–295, Oct. 1910.

[11] M. Kendall, “A new measure of rank correlation,” *Biometrika*, vol. 30(1-2), pp. 81–89, 1938.

[12] F. Mosteller and J. W. Tukey, *Data analysis, including statistics.* New Jersey: Murrey Hill, 1986, ch. 15.

[13] E. Pedhazur, *Multiple regression in behavioral research: explanation and prediction.* Harcourt Brace College Publishers, 1997. [Online]. Available: http://books.google.com/books?id=h-wkAQAAIAAJ

[14] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968

[15] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, “Ranking a random feature for variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1399–1414, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944980

[16] M. A. HALL, “Feature subset selection : a correlation based filter approach,” *Proc Fourth International Conference on Neural Information Processing and Intelligent Information Systems, Dunedin, New Zealand, 1997*, 1997. [Online]. Available: http://ci.nii.ac.jp/naid/10020952250/en/

[17] L. Yu and H. Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution,” *In Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03).*

[18] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226 –1238, aug. 2005.

[19] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, Dec. 1997. [Online]. Available: http://dx.doi.org/10.1016/S0004-3702(97)00043-X

[20] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Information Sciences*, vol. 179, no. 13, pp. 2208 – 2217, 2009, ¡ce:title¿Special Section on High Order Fuzzy Sets¡/ce:title¿. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025509000917

[21] T. N. Lal, O. Chapelle, J. Weston, and A. Elissee, *Feature Extraction, Foundations and Applications.* Springer, 2006, vol. 207, no. 11, ch. Embedded methods, pp. 137–165. [Online]. Available: http://eprints.ecs.soton.ac.uk/11922/

[22] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded methods."

[23] K. Kancherla and S. Mukkamala, "Feature selection for lung cancer detection using svm based recursive feature elimination method," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, ser. Lecture Notes in Computer Science, M. Giacobini, L. Vanneschi, and W. Bush, Eds. Springer Berlin / Heidelberg, 2012, vol. 7246, pp. 168–176.

[24] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. 26, no. 9, pp. 917–922, Sept. 1977. [Online]. Available: http://dx.doi.org/10.1109/TC.1977.1674939

[25] J. Sheffield, G. Goteti, F. Wen, and E. F. Wood, "A simulated soil moisture based drought analysis for the united states," *Journal of Gephysical Research*, vol. 109, 2004.

[26] T. B. McKee, N. J. Doesken, and J. Kleist, "The relationship of drought frequency and duration to time scales," in *Proc. Conference on Applied Climatology*, Anaheim, CA, Jan. 1993, pp. 179–184.

BIOGRAPHICAL STATEMENT

Piraporn Jangyodsuk was born in Bangkok, Thailand, in 1983. She received her B.S. degree in Computer Engineering from Kasetsart University, Thailand, in 2006, her M.S. degrees from The University of Texas at Arlington in 2012 in Computer Science. Her current research interest is in the area of machine learning and data mining especially in time series mining.