Analysis of North Atlantic Tropical Cyclone Intensify Change Using Data Mining

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Jiang TANG
Master of Science
GEORGE MASON UNIVERSITY, 2005

Director: Ruixin Yang, Associate Professor
College of Science

Summer Semester 2010
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my husband Yujie Zhao and my son Eric Zhao, from whom I have had the most happiness.

ACKNOWLEDGEMENTS

I am heartily thankful to my committee members, Dr. Ruixin Yang, Dr. Daniel Babara, Dr. Kirk Borne, Dr. Donglian Sun, Dr. Menas Kafatos (former committee member), and Dr. Long Chiu (former committee member). Dr. Ruixin Yang's guidance and support from the initial to the final level enabled me to develop an understanding of the subject and complete this dissertation. It is Dr. Daniel Barbara who brought me to the wonderful world of association rule mining and Dr. Long Chiu who showed me the direction of studying the latent heat release in tropical cyclone. If it was not because of the generous support from Dr. Menas Kafatos, I would not have been possible to come to the United States and study at George Mason University.

This dissertation would not have been possible without Dr. Mark DeMaria and Mr. John Kaplan who kindly provided the SHIPS dataset, Dr. Christian Borgelt who provided an implementation of the Apriori algorithm, Dr. Owen Kelley who explained in details of how to define and search for a hot tower in the TRMM 2A25 data product, and TSDIS team who developed the TRMM Orbit Viewer.

In addition, I offer my regards and blessings to all of those who supported me in any respect during the completion of my dissertation – Dr. David Wong in Department of Earth Systems and Geo-information Sciences at George Mason University, Dr. Wanting Wang, Mr. Brain Fleeger, Mr. Yang Liu, Ms. Chungguan Yu, Dr. Yixiang Nie, Dr. Yukun Xing, Dr. Guido Cervone, and Dr. Hesham El-Askary in the Center for Earth Observing and Space Research at George Mason University, Dr. Xintao Wu and Mr. Yong Ye in University of North Carolina at Charlotte, and Mr. Otmane Benali in SWIFT Inc..

Last but not least, I would like to thank my family: my husband Yujie Zhao and my son Eric Zhao. They made my world more bountiful. To them I dedicate this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AODT	–	Advanced Objective Dvorak Technique
CDO	–	Central Dense Overcast defined in Dvorak Technique
DT	–	Dvorak Technique
ESMR	–	Electrically Scanning Microwave Radiometer aboard Nimbus satellites
GMS	–	Geosynchronous Meteorological Satellite
GOES	–	Geostationary Operational Environmental Satellites
HDF	–	Hierarchical Data Format
IR	–	Infrared
KD03	–	Reference of Kaplan and DeMaria, 2003
LHR	–	Latent Heat Release
MPI	–	Maximum Potential Intensity
MW	–	Microwave
NCEP	–	National Centers for Environmental Prediction
NHC	–	National Hurricane Center
NOAA	–	National Oceanic and Atmospheric Administration
PCT	–	Polarization Corrected Temperature
PIP	–	Precipitation Intensity Parameter
PR	–	Precipitation Radar aboard TRMM satellite
RI	–	Rapid Intensification
RIP	–	RI probability
RR	–	Rain Rate
RSS	–	Remote Sensing Systems Inc.

SAL	–	Sahara Air Layer
SCAMS	–	Scanning Microwave Spectrometer aboard Nimbus satellites
SHIPS	–	Statistical Hurricane Intensity Prediction Scheme
SDS	–	Scientific Datasets
SSM/I	–	Special Sensor Microwave Imager aboard Defense Meteorological Satellite Program (DMSP) satellites
SST	–	Sea Surface Temperature
TB	–	Brightness Temperature
TC	–	Tropical cyclone
TMI	–	TRMM Microwave Imager aboard TRMM satellite
TPC	–	Tropical Prediction Center
TRMM	–	Tropical Rainfall Measuring Mission satellite
VIS	–	Visible
VIRS	–	Visible Infrared Scanner aboard TRMM satellite

ABSTRACT

ANALYSIS OF NORTH ATLANTIC TROPICAL CYCLONE INTENSITY CHANGE USING DATA MINING

Jiang Tang, Ph. D.

George Mason University, 2010

Dissertation Director: Dr. Ruixin Yang

Tropical cyclones (TC), especially when their intensity reaches hurricane scale, can become a costly natural hazard. Accurate prediction of tropical cyclone intensity is very difficult because of inadequate observations on TC structures, poor understanding of physical processes, coarse model resolution and inaccurate initial conditions, etc. This study aims to tackle two factors that account for the underperformance of current TC intensity forecasts: (1) inadequate observations of TC structures, and (2) deficient understanding of the underlying physical processes governing TC intensification.

To tackle the problem of inadequate observations of TC structures, efforts have been made to extract vertical and horizontal structural parameters of latent heat release from Tropical Rainfall Measuring Mission (TRMM) Precipitation Radar (PR) data products. A case study of Hurricane Isabel (2003) was conducted first to explore the feasibility of using the 3D TC structure information in predicting TC intensification.

Afterwards, several structural parameters were extracted from 53 TRMM PR 2A25 observations on 25 North Atlantic TCs during the period of 1998 to 2003. A new generation of multi-correlation data mining algorithm (Apriori and its variations) was applied to find roles of the latent heat release structure in TC intensification. The results showed that the buildup of TC energy is indicated by the height of the convective tower, and the relative low latent heat release at the core area and around the outer band. Adverse conditions which prevent TC intensification include the following: (1) TC entering a higher latitude area where the underlying sea is relative cold, (2) TC moving too fast to absorb the thermal energy from the underlying sea, or (3) strong energy loss at the outer band. When adverse conditions and amicable conditions reached equilibrium status, tropical cyclone intensity would remain stable.

The dataset from Statistical Hurricane Intensity Prediction Scheme (SHIPS) covering the period of 1982-2003 and the Apriori-based association rule mining algorithm were used to study the associations of underlying geophysical characteristics with the intensity change of tropical cyclones. The data have been stratified into 6 TC categories from tropical depression to category 4 hurricanes based on their strength. The result showed that the persistence of intensity change in the past and the strength of vertical shear in the environment are the most prevalent factors for all of the 6 TC categories. Hyper-edge searching had found 3 sets of parameters which showed strong intramural binds. Most of the parameters used in SHIPS model have a consistent “I-W” relation over different TC categories, indicating a consistent function of those parameters in TC development. However, the “I-W” relations of the relative momentum flux and the

meridional motion change from tropical storm stage to hurricane stage, indicating a change in the role of those two parameters in TC development.

Because rapid intensification (RI) is a major source of errors when predicting hurricane intensity, the association rule mining algorithm was performed on RI versus non-RI tropical cyclone cases using the same SHIPS dataset. The results had been compared with those from the traditional statistical analysis conducted by Kaplan and DeMaria (2003). The rapid intensification rule with 5 RI conditions proposed by the traditional statistical analysis was found by the association rule mining in this study as well. However, further analysis showed that the 5 RI conditions can be replaced by another association rule using fewer conditions but with a higher RI probability (RIP). This means that the rule with all 5 constraints found by Kaplan and DeMaria is not optimal, and the association rule mining technique can find a rule with fewer constraints yet fits more RI cases. The further analysis with the highest RIPs over different numbers of conditions has demonstrated that the interactions among multiple factors are responsible for the RI process of TCs. However, the influence of factors saturates at certain numbers.

This study has shown successful data mining examples in studying tropical cyclone intensification using association rules. The higher RI probability with fewer conditions found by association rule technique is significant. This work demonstrated that data mining techniques can be used as an efficient exploration method to generate hypotheses, and that statistical analysis should be performed to confirm the hypotheses, as is generally expected for data mining applications.

CHAPTER 1 Introduction

Tropical Cyclone (TC), especially with high intensity, is among the most deadly weather systems on the Earth. It acts as a giant energy vessel carrying massive water and blowing strong wind from ocean to land as well as pumping heat from the surface to the upper troposphere. Excessive rain induces flooding and land sliding. Strong wind gust blows off trees and buildings and breaks electricity power line. On average, two hurricanes strike the mainland of the United States every year and result in almost 29 deaths (FEMA 1995; Herbert et al. 1996). Average losses caused by hurricane are around of $4.8 billion per year in 1995 dollars (Pielke and Landsea 1998). Therefore, accurately forecasting tropical cyclone track and intensity is of utmost concerns, especially for marine warnings, landfall evacuation planning and decision-making. Knowing the track of a hurricane ahead of time can help people find out the impacted region and when the hurricane will reach that region. Knowing the intensity of a hurricane can help people estimate the regional damage and make massive evacuation decision when it is needed.

Scientists have started the study of hurricane track and intensity forecast since 1970's. The skill of track forecasting has made much bigger progress than that of intensity forecasting (Aberson 2001; Wang and Wu 2004) without mention that the operational track models outnumbers the intensity models in National Hurricane Center

(NHC) (DeMaria 1997). After assembling the predicted results from the 18 worldwide models available to Tropical Prediction Center/National Hurricane Center (TPC/NHC) hurricane specialists, Aberson (2001) found that more than 90% of the current track forecasts span the true path of tropical cyclones at all forecast times. And the long-term 72hr track forecast has a 2.4% improvement rate from 1976 to 2000. In contrast among the primary models currently used as guidance for TC intensity prediction at the National Hurricane Center (NHC), the Statistical Hurricane Intensity Prediction Scheme – SHIPS (DeMaria and Kaplan 1994b, 1999, 2005) or its variants can only explain 30% to 50% variance in tropical cyclone intensity that can be accounted for by the climatological, persistence and synoptic predictors (DeMaria and Kaplan 1994b), even though they produced the best forecasting skills on average (Franklin 2008).

The difficulty of the TC intensity forecasting reflects the fact that many factors control TC intensity changes (DeMaria et al. 2007). Broad studies have been carried out on the favorable factors for predicting TC intensification. These favorable factors include warm ocean eddies (Shay et al. 2000; Hong et al. 2000; Wu et al. 2007), the contraction of an outer eyewall (Willoughby et al. 1982; Willoughby and Black 1996; Lee and Bell 2007), an environment with low vertical shear (Gray 1968; Merrill 1988; DeMaria and Kaplan 1994b; DeMaria 1996; Frank and Ritchie 1999, 2001; Zeng et al. 2007, 2008), interactions between the upper-level trough and a TC (Molinari and Vollaro 1989, 1990; DeMaria et al. 1993; Bosart et al. 2000), dissipative heat (Jin et al. 2007), cloud microphysics (Wang 2002a) and isotopic concentrations (Gedzelman et al. 2003).

Usually TC intensity change studies start with individual TC cases, and case study is still a major approach for examining mechanisms supporting TC genesis, formation, and intensification (e.g., Bosart et al. 2000; Zehr 2003; McTaggart-Cowan et al. 2006; Montgomery et al. 2006; Lowag et al. 2008). Case studies were also performed by numerical simulations with specific initial conditions and environmental conditions such as the numerical simulations of Hurricane Bonnie (1998) by Rogers et al. (2003), Zhu et al. (2004), Zhu and Zhang (2006), and Cram et al. (2007). Model simulation can also be used for studying isolated factors such as the impacts on TC intensity and structure by sea spray evaporation (Wang et al. 2001) and Rossby waves (Wang 2002b).

Composite analysis is another major approach for identifying common factors affecting TC intensity changes and for understanding the underlying physics (e.g., Hanley et al. 2001; Knaff et al. 2004, 2008; Zehr and Knaff 2007; Camargo et al. 2007a, b; Ventham and Wang 2007). More advanced composite analysis uses statistical forecasting models for TC intensity forecasting. One of the successful examples is the Statistical Hurricane Intensity Prediction Scheme (SHIPS) model based on linear regression analysis from a set of selected parameters (DeMaria and Kaplan 1994b, 1999; DeMaria et al. 2005). The SHIPS model uses a group of predictors (a.k.a., parameters, or factors) which represent most of the environmental factors known to affect TC intensity changes. The derived parameters for SHIPS model construct a SHIPS database, which was also used for studies of probabilities of rapid intensification (RI) of TCs (Kaplan and DeMaria 2003; Yang et al. 2007, 2008).

Composite analysis of TC intensity change usually is based on traditional statistical analysis methods and is used extensively to find associations between TC intensity changes such as intensifications or rapid intensifications and their environmental properties. It can be considered as a type of "one-to-one" relation analysis technique. In recent years, data mining techniques are becoming a widely used approach to have extensive and sometimes exhaustive search of hidden relationships in vast amounts of data. For example, comparing with the "one-to-one" relation analysis, the technique of association rules in data mining can explore associations among multiple conditions because it automatically examines all possible combinations of frequent condition sets in a large complex dataset (Agrawal et al. 1993). This data mining technique provides an as complete as possible picture of the dataset to scientists so that the connections among multiple conditions are not overlooked by a theory-driven analysis approach.

The goal of this study is to apply the association rule technique as an "unsupervised," "automatic" data exploration method to discover "multiple-to-one" associations of a large number of geophysical characteristics with the weakening, intensifying, or rapidly intensifying tropical cyclones. The data mining results can then be used to shed light on the hypothesis generation regarding the underlying physical mechanisms, which can be used as guidance in the traditional statistical analysis as well as TC intensity forecasting.

The rest of this thesis is organized as follows. A literature review of tropical cyclone intensification theory, model and measurements are given in chapter 2. In chapter 3, the data sets for this study including TRMM PR products, NHC best track data, parameters used in the SHIPS model and the association rule data mining concept will be introduced.

The analysis of using the 3D structural parameters of latent heat release from TRMM PR 2A25 data products will be discussed in chapter 4. The TC stratification and data discretization method will be discussed in chapter 5, and data mining results for stratified TCs will be deliberated in chapter 6. Chapter 7 will be designated to the studies of rapid intensification, one major error source for TC intensity forecasting. Chapter 8 will summarize the data mining results for TC intensity changes, discuss the process of applying association rule mining technique in tropical cyclone intensity forecasting, and point out several directions for future study.

CHAPTER 2 Historic Research on Tropical Cyclone Intensification

The term “hurricane” or “typhoon” is a regionally specific name for a strong “tropical cyclone.” A tropical cyclone is “a non-frontal synoptic scale low-pressure system over tropical or sub-tropical waters with organized convection and definite cyclone surface wind circulation” (Landsea 1993, Holland 1997). That is, a tropical cyclone is a natural phenomenon at synoptic scale. It can be easily recognized by its convection and cyclonic wind circulation at the surface. It is not caused by the front system (i.e., the interaction between warm air body and cold air body), and its occurrence indeed restricts to over tropical or sub-tropical waters only.

2.1 Measuring the current intensity of a tropical cyclone

A precise measure of the current wind strength is a must to accurately predict TC intensity. Usually a wind (and pressure) measurement is obtained by aircrafts flying through the cyclones. However it is too difficult and too expensive to monitor the wind speed in a tropical cyclone when it takes place in the middle of the ocean, which is a typical situation in the tropical cyclone study. Therefore, tropical meteorologists began to use satellite images to infer the wind strength since the early 1960’s.

The most prodigious work had been done by Dvorak in the 1970's. In his work Dvorak (1972, 1975, and 1984) recognized that the cloud patterns of tropical cyclones evolve not continuously but in surges from one recognizable stage to another as the intensity of the cyclone changes. Dvorak then defined several features (called patterns in his work) from geostationary visible/infrared (VIS/IR) image from GOES satellites and related them to the current intensity of a cyclone. The first feature, curved band measurement, is defined by the arc length of the curved band fitted to a 10 °logarithm spiral overlay. The second feature, shear pattern, measures the distance between the low cloud center and the dense, cold overcast. Dvorak argued that the appearance of shear inhibits the cold clouds from bending around the cloud system center as they do in the curved band patterns. The third feature, the eye, is a relatively warm spot surrounded by the coldest, curved overcast. The eye pattern measures the minimum width and the gray scale of the temperature of the surrounded cold clouds. Sometimes the eye is covered by a dense, solid-looking mass of clouds, which is called the central dense overcast (CDO), identified as the fourth feature. This feature can be found in the VIS image or as the embedded center in the counterpart IR image. The CDO pattern evaluates the abruptness of the edge of the cloud mass and measures the diameter of the CDO (called the CDO size). The embedded center pattern, similar to the eye pattern, measures the embedded distance from the center of eye for small eyes or from the inner wall of the eye for large eyes to the nearest outside edge of the dense overcast.

In spite of its long history in operational state, Dvorak's technique requires manual interpretation of the features from image which is labor-intensive and tends to cause

errors. To overcome the subjectivity of this technique, Zehr (1989) developed a computer-based objective technique using enhanced IR satellite data, which can objectively identify pattern types, calculate the eye/convection temperatures, apply selected rules and derive intensity estimates. The crucial step in Dvorak Technique (DT) is the determination of the eye or cloud center. This objective Dvorak Technique (ODT) needed manual input of cloud center locations. And it worked only for cases when the tropical cyclone was well organized and possessing intensity at or greater than minimum hurricane intensity (Velden, et al, 1998). So the journey of the automated Advanced Objective Dvorak Technique (AODT) had begun in 1998 (Velden, et al 1998) to expand the range of applicability to weak tropical cyclone and to automatically select cloud center and continued to present (Wimmers and Velden, 2004, Olander, Velden and Kossin, 2004).

Due to the very short wavelength of the VIS and IR channels, the cloud particles interfere so heavily with VIS and IR signals that even a thin shield of cirrus on top of the tropical cyclone eye can inhibit the interpretation of the location of the eye thus undermine the credibility of DT forecast. Therefore, scientists begin to use the microwave (MW) data from TRMM and SSM/I as a supplement to Dvorak Technique since the non-rain clouds have little effect on MW signal (Edson 2004). Nevertheless, it cannot be used as a replacement for the VIS/IR technique due to its low resolution and lack of continuous coverage.

2.2 Theory of TC intensification

2.2.1 Tropical cyclone genesis

The formation of a tropical cyclone is the premise of its intensification. The consensus of the favorable environmental conditions in tropical cyclone genesis includes at least six conditions (Landsea 1997). The first condition is the presence of warm ocean waters (>= 26.5^{o}C) throughout a sufficient depth (on the order of 50m) because warm water is the energy source of tropical cyclones. The second condition is an atmosphere which cools fast with height so as to be unstable to moist convection. The third condition is the relatively moist layers near the mid-troposphere (around 5 km) for continuous development of thunderstorm activities. The fourth condition, a minimum distance of at least 500km from the equator, is a requirement for non-negligible amounts of the Coriolis force to produce gradient wind and maintain a low pressure center. The fifth condition is a pre-existing near-surface disturbance with sufficient vorticity and convergence acting as a seed. The last but not least condition is the weak vertical wind shear (<10m/s) between the surface and the upper troposphere.

Although they are only the necessary but not the sufficient conditions to form a tropical cyclone, most of the favorable conditions in tropical cyclone genesis continue to help a cyclone to maintain or keep it growing. A cyclone cannot live without its energy source – the warm water. It cannot maintain its strenght without the sufficient Coriolis force, a weak vertical wind shear condition, and the convective activities. Only the pre-existing vortical disturbance is not that important after a tropical cyclone is formed.

2.2.2 Maximum potential intensity theory

Recognizing the warm ocean water is the solo energy source of a cyclone, attempts have been made to estimate an upper bound of tropical cyclone intensity, referred as maximum potential intensity (MPI), of given atmospheric and oceanic conditions. There are two schools of MPI theory (Wang and Wu 2004). One was proposed by Emanuel (1988). He assumed that the tropical cyclone behaves like a classic Carnot heat engine with energy added at the underlying warm ocean surface and lost in the cool outflow area, and argued that the MPI can be determined by the maximum entropy difference between the cyclone center and the environment. The other theory was proposed by Holland (1997). He argued that the cyclone intensification is due to the warming from moist-adiabatic ascent in the eyewall convection thus the MPI is largely determined by the temperature anomaly in the upper-troposphere. More thorough discussions on MPI can be found in Holland (1997) and Camp and Montgomery (2001).

Statistics has shown that most storms could only reach 55% of their MPI, which leads to an accompanied issue with MPI theory, i.e., what factors prevent the cyclone from reaching its MPI? The commonly recognized limiting factors, indicated by Wang and Wu (2004), include the asymmetry in the inner core region induced by the vertical shear and the cooling of sea surface due to the oceanic upwelling under the eyewall region. Wang and Wu (2004) also pointed out that the mesoscale processes such as the inner and outer spiral rainbands, convectively coupled vortex Rossby waves, eyewall cycles, and embedded mesovortices play key roles in tropical cyclone intensity change.

2.2.3 Hot-tower theory in TC intensification

Postulating that most heating of the upper troposphere in the Tropics is due to towering cumulonimbus clouds rather than broad uplifting, Riehl and Simpson (now Joanne Malkus) started the "hot tower hypothesis" in the late 1950s (Kelley and Stout 2004). These upper-troposphere cumulonimbus clouds are also called as "convective towers" or "hot towers." They can exist inside tropical cyclones and keep warming the upper troposphere inside a cyclone. Therefore, they have been seen as one of the mechanisms that maintain the intensity of a tropical cyclone.

Before TRMM data was available, very few dataset could show the presence of a hot tower in a tropical cyclone and most discussions on the hot tower were descriptive. Using the exact measurement of the convection height from TRMM data, Kelley and Stout (2004) found that intensifying cyclones are more than twice likely to have at least one convective tower in their eye walls than non-intensifying cyclones, which testified the hot tower theory.

2.3 Models of TC intensification forecast

2.3.1 Dvorak's decision rules based on remote sensing image recognition

Dvorak (1984) proposed a model of tropical cyclone development in terms of the day by day cloud pattern changes observed by satellite VIS/IR images. He argued that the weakening in a storm's cloud pattern precedes the weakening of its pressure pattern. In his model, a typical growth rate is as one "T-number" per day for 70% Atlantic cyclones.

A northwestward-moving storm is expected to peak 5 days after the T1 classification; westward movers in 6 days; and northward mover in 4 days. The unfavorable signs around the cloud center identified by Dvorak include persistent warming of cloud patterns for more than 12 hours, a central cold cover persisting for more than 3 hours, a field of stratocumulus clouds in the path of the storm, the cirrus cloud lines, and the cloud pattern elongating. The unfavorable signs in the environment include stratocumulus clouds, land, southward-moving cirrus to the north or west of the storm, unidirectional flow, and a southward surge of the westerly. A 24-hour forecast can then be made based upon Dvorak's model.

Dvorak's work has been in the operational use for almost 40 years. The longevity of Dvorak Technique, argued by Guard (2004), roots from the fact that the proposed patterns relate well to the underlying kinematics and thermo-dynamics processes of tropical cyclone intensification. The rotation pattern of clouds is the result of the internal vorticity, i.e., the strength and distribution of circular winds in a tropical cyclone. The elongated patterns at a cyclone's periphery are the evidence of the environmental destructive force, vertical wind shear, which distorts the internal vorticity of a tropical cyclone and decouples the deep convection and the low-level circulation. The eye temperature, eye distinctness, the depth and the thickness of the eyewall convection are a reflection of the convective vigor, which is the dominant influence in the intense tropical cyclones.

2.3.2 Statistical models based on persistence, climatological and synoptic parameters

In addition to Dvorak's technique (Dvorak 1972, 1975, 1984) which is an image classification approach, there are four other types of forecast models running in National Hurricane Center (NHC) based on DeMaria's description (DeMaria 1997). One type is the statistical model, for example, Statistical Hurricane Intensity Forecast (SHIFOR) model. The second type is the statistical-synoptic model such as Statistical Hurricane Intensity Prediction Scheme (SHIPS) in 1994 version. Another type is the statistical-dynamical model such as SHIPS in 1997 version. The last type is the dynamic model such as Geographical Fluid Dynamics Laboratory (GFDL) model and Interpolated GFDL (GFDI) model. The dynamic models utilize three-dimensional primitive equations so they are not considered in this study.

SHIFOR (Statistical Hurricane Intensity FORecast) (DeMaria 1997) is a benchmark for tropical cyclone intensity forecast. It uses climatological and persistence predictors to forecast intensity change (Jarvinen and Neumann 1979). SHIFOR is only valid for storms over the ocean.

SHIPS (Statistical Hurricane Intensity Prediction Scheme), originally was a statistical-synoptic model that utilizes climatological, persistence and synoptic predictors (DeMaria 1997; DeMaria and Kaplan 1994b). Similar to SHIFOR, SHIPS was developed from cases where the storm track did not cross land. The primary predictors include the difference between the maximum possible intensity and the current intensity, the vertical shear determined by the difference between the 850- and 200-mb wind vectors, the eddy

flux convergences related to the radial wind and its distance from the storm center, the previous 12 hr intensity change, and the zonal wind and temperature within 1000 km of the storm center at high altitude. At present the SHIPS model has been continuously updated, and become more dynamical (DeMaria and Kaplan 1999, DeMaria et al 2005). Although a dynamical model simulates our present understanding about the physical processes in cyclone development, the limited understanding affects the accuracy of model outputs.

2.4 Predictive parameters of TC intensification

2.4.1 Climatology and persistence parameters in TC intensification

Climatology parameters describe the mean or the most frequent behaviors of tropical cyclone intensification in hurricane seasons from history data. For example, observing that September 10 is the peak of hurricane season, that is, on September 10 (Julian date 253) the probability of the existence of a tropical cyclone in the Atlantic basin is maximized, SHIPS (DeMaria and Kaplan 1994b and 1999) uses the difference of the current Julian date of a tropical cyclone to the Julian date 253 as one of the climatology predictors (JDAT). Persistence parameters describe the inertial or the past behavior of the tropical cyclone, for example, the initial storm intensity (V0), intensity change during previous 12 hours (IV12), initial storm location (LAT, LON), and storm motion (ZONX, MERY). Together the climatology and persistence variables can explain 30%-40% of the variance of the intensity changes over different forecast intervals, e.g.

12-, 24-, until 72-hour (DeMaria and Kaplan 1994b). Table 2-1 shows a list of climatology and persistence parameters used in SHIPS model.

Table 2-1 Climatology and persistence predictors in SHIPS

(DeMaria and Kaplan 1994b).

No.	Description of predictors	Name in DK 1994b	Name in this study
1)	Absolute value of Julian date – 253	JDATE	JDAT
2)	Initial storm intensity	VMX	V0
3)	Intensity change during previous 12 h	DVMX	IV12
4)	Initial storm latitude (^{o}N)	LAT	LAT
5)	Initial storm longitude (^{o}S)	LONG	LON
6)	Eastward component of storm motion	USM	ZONX
7)	Northward component of storm motion	VSM	MERY
8)	Magnitude of storm motion vector	CSM	SPD

2.4.2 Synoptic parameters of TC intensification

Most of the synoptic parameters are used to describe the environment around a cyclone. They include information about the temperature of ocean surface, the water vapor content (humidity), and the wind activity in the atmosphere. Table 2-2 shows a list of synoptic parameters used in SHIPS model. DeMaria and Kaplan (1994b) had explained in great detail of the definition of each parameter and its relation to TC intensification. Below listed is a brief description for each parameter from DeMaria and Kaplan's study.

Table 2-2 Synoptic predictors in SHIPS

(DeMaria and Kaplan 1994b, 1999, DeMaria, et al. 2005)

No.	Name	Description	Unit
1)	POT	Maximum potential intensity – current intensity,	knots
2)	SHRD	850-200 hPa vertical wind shear	knots
3)	REFC	200 hPa eddy momentum flux convergence,	m/sec/day
4)	Z850	850 hPa relative vorticity	s^{-1}
5)	DTL	Distance to the nearest landmass	km
6)	T200	200 hPa temperature	^oC
7)	U200	200 hPa zonal wind	knots
8)	D26C	upper ocean heat content through the depth of the 26^oC isotherm	m
9)	D200	200 hPa divergence, unit:	s^{-1}
10)	PSLV	Pressure level of storm steering, unit:	mb
11)	RHLO	850-700 hPa relative humidity	%
12)	RHHI	500-300 hPa relative humidity	%
13)	EPOS	Surface–200 hPa θe deviation of lifted parcel	^oC

POT refers to intensification potential, i.e., the difference between the maximum potential intensity and the current intensity. Based on the MPI theory in which the SST appears to provide an upper bound on the intensity of the storm (Merrill 1987, Emanuel 1988), POT has a positive correlation with future intensity change. DeMaria and Kaplan (1994a) used an empirical relationship to compute MPI for given SST,

$MPI = A + B\exp[C(SST - SST_0)]$, where A = 66.5 knots, B = 108.5 knots, C = 0.1813$^oC^{-1}$, SST_0 = 30.0^oC, and SST is the sea surface temperature (^oC).

SHRD, the magnitude of the vertical shear between 850- and 200-mb, has a negative influence on TC intensification (e.g., Gray 1968, Merrill 1988).

REFC, the relative momentum flux at 200 mb, is included to account for positive interactions between the tropical cyclone and synoptic-scale systems. It provides a measure of whether the large-scale flow increases or decreases the azimuthally averaged tangential wind of the storm. It plays a role that is analogous to relative vorticity advection in quasi-geostropic theory (e.g., Holton 1992). Holland and Merrill (1984) and Molinari and Vollaro (1989) suggested that when a storm interacts with the large-scale flow in a way that makes the upper-level flow more cyclonic, the intensification rate of the storm may increase. The interaction of the storm with the environment is more likely to occur in the upper level. At lower levels, the large inertial stability due to the rapid rotation of the storm circulation limits the interaction of the storm and environment.

REFC can be computed from $REFC = -r^{-2} \frac{\partial(r^2 \overline{U_L' V_L'})}{\partial r}$, where r is the radius from the storm center, U is the radial wind, V is the tangential wind, the overbar denotes an azimuthal average, the prime represents a deviation from the azimuthal average, and the subscript L indicates that these quantities (U or V) are evaluated in a coordinate system moving with the storm. Qualitatively, REFC tends to be large when a storm is moving towards an upper-level trough.

Z850, the 850-mb environmental vorticity parameter, is positively related to future intensity change, indicating that intensification is favored when the synoptic environment at 850 mb is more cyclonic than average. It is proposed to substitute for another parameter SIZE (DeMaria and Kaplan 1994b) as a measure of the extent of the

outer circulation of the tropical cyclone. It is averaged over a circular area within a radius of 1000 km of the storm position.

DTL, the distance to the nearest landmass, is positively related to the future intensity change. The landmass cuts off the energy source of a TC, thus devastates its development.

T200, the 200-mb temperature averaged over a circular area with a radius of 1000 km centered on the initial storm position, has a negative correlation coefficient with future intensity change, which indicates that intensification is favored when the 200-mb temperature is colder than normal. This parameter is proposed to account for several TCs such as Hurricane Florence (1994), Isidore (1990), Bonnie (1992), and Charley (1992), which intensified over relatively cool water and had anomalously cold environment at 200 mb. In most of the 1995 hurricane season, T200 in the tropical regions was 1~2 ^{o}C warmer than normal, which resulted in an underestimate of the storm intensity.

U200, the 200-mb zonal wind averaged over the same area as T200, is negatively related to future intensity change. Since a positive U200 is heading to east, called a westerly wind, this negative relationship indicates that intensification is favored when the 200-mb zonal wind is more easterly than normal. This relationship is consistent with operational forecast guidelines since 200-mb winds are more easterly than normal equatorward of an upper-level ridge, which is considered a favorable region for intensification.

D26C is the upper ocean heat content through the depth of the 26^{o}C isotherm. And D200, the 200-hPa divergence, is added as a measure of synoptic-scale forcing.

PSLV, the steering layer pressure, is used to find the weights on the horizontal winds from 1000 to 150 hPa surrounding the storm (200 to 800 km annular average) that provide the best match to the observed storm motion. It was found that this variable is negatively correlated with intensity change because for highly sheared storms that became decoupled in the vertical, the PSLV occurs with a relatively high value.

RHLO, the relative humidity at 850-700 hPa from the NCEP global model, is used to estimate the Sahara Air Layer (SAL). SAL is characterized by very dry air in the 850-500 hPa layer. Dunion and Velden (2004) showed that SHIPS model overestimated intensification for storms that interacted with the Sahara Air Layer (SAL), such as for Hurricane Joyce (2000) and Debby (2000). However, Dunion and Velden (2004)'s result also suggested that the NCEP global analysis does not always represent the magnitude of the low-level drying associated with the SAL.

EPOS, the averaged positive differences between the surface θ_e of a lifted parcel and that of the original NCEP profile, is a crude representative of the atmospheric stability. SHIPS performed poorly for tropical storm Dolly in 2002 (DeMaria et al 2005), which did not intensify despite the favorable low-shear environment with fairly warm water. This motivated the testing of a new time-dependent thermodynamic predictor EPOS.

2.4.3 Physical structure parameters based on VIS/IR images

When using VIS/IR images to forecast tropical cyclone intensity, Dvorak technique (Dvorak 1972, 1975, and 1984) is the classic approach. The structure parameters are extracted from cloud patterns and shown in Table 2-3.

Table 2-3 Cloud pattern parameters in Dvorak Technique

(Dvorak 1972, 1975, 1984)

No.	Pattern name	Measurement of the pattern
1)	Curved band pattern	Arc length of the band, curved band type
2)	Shear pattern	Distance from the eye center to the dense overcast
3)	Eye pattern	Band width around the eye, distance from eye to the outer edge of the band, surrounding TB gray scale
4)	Central dense overcast	Edge of the overcast, diameter of the CDO
5)	Embedded center pattern	distance from eye to the outer edge of the band, surrounding TB gray scale
6)	Brightness temperature	TB gray scale at eye, TB gray scale around the eye

Table 2-4 explained how to use the change of these patterns to predict the TC intensity change.

Table 2-4 Changes of cloud pattern in Dvorak Technique for predicting TC intensity change

(Dvorak 1972, 1975, 1984)

Pattern	Tropical cyclone will intensify	Tropical cyclone will weaken
Curved band	Curved band coils	Curved band untwists
CDO	CDO becomes larger or banded	CDO declines Or cold CDO persists (>= 3hr)
Shear	CSC fits curved cloud lines or closes to the dense overcast	CSC far away from the curved line
TB	The TB gray scale has little changes	TB gray scale changes significantly

In addition to those parameters, researches on Nimbus IR sensors, GOES and GMS have shown that the minimum cloud-top temperatures occurred approximately 24 hours prior to intensification (Rodgers, Chang and Pierce 1994).

2.4.4 Physical structure parameters from RS MW images

Although VIS/IR observations of tropical cyclones have a long history, they can only indirectly estimate the TC intensity. Due to the comparable short wavelength of VIS/IR radiation to cloud ice parcels, the cirrus shield on top of a tropical cyclone may block the VIS/IR emissions from the underlying convective cloud bands. Fortunately clusters of deep convection in convective cloud bands can be seen clearly through the shield when using MW images. Therefore, the rain rate and the latent heat release in both cyclone's inner core and outer bands can be derived directly from MW measurements.

The rain rate (RR), the latent heat release (LHR) and the brightness temperatures (TB) correlate to the future (24-72 h) tropical cyclone intensity have been derived in the studies using the MW sensors, such as Nimbus ESMR (Rodgers and Adler 1981) and SCAMS (Kidder et al. 1978; Velden and Smith 1983), DMSP SSM/I (Rao and McCoy 1997; Rao and MacArthur 1994a, 1994b; Cecil and Zipser 1999), TRMM TMI (Hoshino and Nakazawa 2004; Nakazawa and Hoshino 2004) and PR (Kelley and Stout 2004).

Rao and McCoy (1997) centered a set of rectangular grid boxes along the direction of motion of 13 typhoons in 1987 season in totally 27 map times. Those boxes have various annuli ranging from 27.5 km to 888 km. Using these settings, they found that the mean 85 GHz V TB in the left half box of 111 km, the TB variance in the left 111km half box, and the outer-inner TB anomaly (mean right 888 km half box annular TB – mean right 444 km half box TB) correlated well to the 24-h intensity (in terms of the wind speed), especially for the tropical cyclones moving northwestwards or those moving slowly (translation speed is less than 6 knots).

Rodgers, Chang and Pierce (1994) used a stereographic horizontal grid for the rain rate and latent heat release quantities derived from SSM/I. The grid is circular and the annuli are 55.5 km in width. They tested 103 observations of 18 named western North Atlantic tropical cyclones that occurred during 1987 to 1989. They found that intensifying tropical cyclones had greater PIP (the Precipitation Intensity Parameter) within the inner rings than that of weakening or steady-state systems. PIP, a measurement of the heavy convective precipitation, was defined in their research as the fraction of rainfall contributed by rain rates greater than 5 mm/hr within a given annular area.

Similar relationship has been found by Chang et al. (1995) for five typhoons passing Taiwan in 1994 season observed by SSM/I.

Since the late 1970s, scientists have successfully estimated tropical cyclone central pressure (a measure of cyclone intensity) using passive microwave sounding instruments aboard polar-orbiting satellites (Kidder, Gray, and Vonder Haar 1978; Velden and Smith 1983). In these techniques, the minimum central pressure through the hydrostatic balance of a tropical cyclone was related to the microwave TB anomaly between the tropical cyclone's upper-level (250 mb) warm core and the environment (2^{o} latitude from the track center) at the same level. The MW TB anomaly was measured by the 55-GHz sounding channel in the Nimbus-6 SCAMS. UW-CIMSS (Herndon et al. 2004) extended this relationship to operational applications using the observations from the AMSU onboard NOAA polar-orbiting satellites.

Cecil and Zipser (1999) defined the SSM/I 85 GHz PCT (Polarization Corrected Temperature) as a proxy for inner core convective updraft strength in a tropical cyclone. They revealed that the mean PCT and the area with PCT below 250K have higher correlation with future tropical cyclone intensity (in terms of maximum wind speed) than the minimum PCT. Hoshino and Nakazawa (2004) and Nakazawa and Hoshino (2004) verified this relationship using TRMM TMI 1B11 data from 12 tropical cyclones in 2000.

Kelley and Stout (2004) used the maximum height in the inner core of a tropical cyclone observed by TRMM PR as a proxy of the strength of the convective towers in eyewalls. They found that this parameter can explain 15% of variations of the change rate of hurricane intensity.

2.5 Data mining in tropical cyclone intensification studies

2.5.1 Image mining for estimating current intensity

Accurate forecast requires precise intensity measurements. However, there are still limitations on estimating current intensity. For most of the tropical cyclones, intensity estimation is done from satellite observations. Many data mining techniques are used to estimate current intensity from satellite images.

The classic Dvorak approach (Dvorak 1972, 1975) is a pilot example on applying image pattern recognition and decision tree scheme on estimating current tropical cyclone intensity. It depends on manual interpretation of the cloud shape and color from geostationary VIS/IR imagery.

May et al. (1997) calculated the empirical orthogonal functions (EOFs, also called Principle Component Analysis (PCA)) using SSM/I images. A neural network was then trained and used to automate tropical cyclone intensity estimation utilizing those EOFs as input.

Kitamoto (2001, 2002a) also applied PCA on a GMS-5 typhoon image collection with approximately 34,000 pictures. He found that the maximum variance in typhoon cloud patterns is in the north-south slope of the cloud amount.

Bankert and Tag (2002) extracted over 100 features from the 85 GHz H TB data and the derived rain rate data centered on 142 tropical cyclones from 1988 to 1998 using SSM/I observations. Using a backward sequential selection algorithm for the feature

selection, they found 15 optimal features from 100 candidates. Later on the k-nearest-neighbor algorithm was served as the pattern recognition algorithm to estimate the current intensity associated with the testing images using the selected 15 features.

2.5.2 Trend forecasting for intensity change

A general solution for TC intensity forecast is to extrapolate the trend of the past and current intensity – a typical question in time series analysis. Statistical regression has been adopted most widely for extrapolation, as in the cases of SHIFOR (DeMaria 1997) and SHIPS (DeMaria and Kaplan 1994b and 1999).

Although none of them were applied in tropical cyclone intensity change, the recent researches on mining streaming data proposed some promising algorithms. One is MUSCLE (Yi et al. 2000; Faloutsos 2004) which is a multi-variant linear prediction method with the co-evolving sequences. Another algorithm F4 (Chakrabarti and Faloutsos 2002; Faloutsos 2004) is a nonlinear regression method based on fractal dimension. The "StatStream" (Zhu and Shasha 2002) can find correlation on multiple real-time streams.

2.5.3 Other mining algorithms in intensification study

In addition to using time series analysis algorithms for forecasting, other approaches were also used in tropical cyclone forecast. For example, decision rules were adopted in Dvorak's later work (Dvorak 1984), in which the change of environmental

parameters was combined with intensity estimates over time to make a 24-hr forecast. In another example, Kitamoto (2002a, 2002b) proposed an analog-based forecast model. This model searches past similar patterns in his typhoon collection and forecast the TC intensification based on the evolution of those past similar cases identified by the searching process.

Principal component analysis (PCA), Discrete Fourier transform (DFT) and Discrete Wavelet transform (DWT) are widely used for feature extraction and dimension reduction in tropical cyclone analysis (Kitamoto 2001, 2002a and 2002 b, Bankert and Tag 2002). K-means clustering and self-organizing maps (SOM) are used to unveil the internal similarity among data as shown in Kitamoto's work (Kitamoto 2002a and 2002b). Furthermore, when a cluster is regarded as a discrete state, the state transition rules can be generated from sequences of tropical cyclone images (Kitamoto 2002a and 2002b).

2.6 Application of association rule mining in geo-science

A new data mining algorithm – Apriori and its variation – has been tentatively applied in geo-sciences data for finding frequent association patterns. For example, Tan et al. (2001) applied it for mining anomalous associated patterns between the net primary production (NPP) and other environmental variables such as soil moisture, temperature and precipitation. They found that the standard measures of interestingness (e.g., support, confidence, and lift) could not help them very much. The deficiency of interestingness measure has been well documented in other research as well. The log-linear model and

the chi-square test have been proposed (DuMouchel and Pregibon 2001, Wu, Barbara, and Ye 2003) to find infrequent but surprising multi-item association rules from market-basket type data.

In order to apply association rule algorithm, most of the geo-science data, including tropical cyclone images, have to be transformed into market-basket type of data. The major concern is how to extract and transform the implicit spatial and temporary relationship from geo-science data. Koperski and his colleageus (Koperski and Han 1995; Koperski, Han and Stefanovic 1998) proposed a spatial predicate for the vector-format spatial data as digital maps, which is widely used in a geographical information system. Their approach is rooted on the Apriori algorithm, but with more considerations and modifications on how to generate a spatial predicate and how to generalize/specialize it. Thus the attribute selection and generalization are the two most important steps in their approach. A RELIEF algorithm was proposed to select relevant attributes. They also demonstrated how to generalize a spatial predicate such as "close_to()." However the image type of spatial data such as the satellite imagery is out of their consideration.

The collocation miner, developed at the University of Minnesota (Shekhar and Huang 2001; Huang et al. 2003; Xiong et al. 2004; Huang, Shekhar and Xiong 2004), was designed for mining spatial co-locations based on vector-type of spatial data. It enumerates the neighborhood of a point-type event and forms transaction baskets based on the neighborhood definition. Then a subset of Boolean spatial features can be identified whether they have occurred together in the neighborhood or not. The

concurrent features are defined as a co-location. A co-location rule can be generated if several statistical requirements are satisfied.

To author's knowledge, no attempts have been made yet on applying association rules in the tropical cyclone study. In this study, the Apriori association rule mining is used to identify combined conditions that affect TC intensity changes for the first time.

CHAPTER 3 Data and Methodology

The datasets for this study are the NHC HURDAT file (Jarvinen et al. 1984), the TRMM PR 2A25 data in 1998-2003, and the SHIPS 1982-2003 database (DeMaria et al. 2005). The association rule data mining technique (Agrawal et al. 1993) and the software implemented by Borgelt (2009) are used in this study.

3.1 NHC best track data

The best track (BT) data used in this study is from NOAA National Hurricane Center (NHC) for the Atlantic basin (http://www.nhc.noaa.gov/pastall.shtml). It is an ASCII (text) file containing the 6-hourly (0000, 0600, 1200, 1800 UTC) center locations (latitude and longitude in tenths of degrees) and intensities (maximum 1-minute surface wind speeds in knots and minimum central pressures in millibars) for all tropical storms and hurricanes since 1851 (Jarvinen et al. 1984). It is so called the "best" because it has assembled multiple post-analysis estimates primarily based on multiple satellite observations. The nominal accuracy of the maximum wind estimates is 5 knots, since the wind values are rounded to this interval in the best track data (Jarvinen et al. 1984). Comparisons of satellite and aircraft reconnaissance intensity estimates suggested that the average error in the minimum central pressure is about 10 mb, which corresponds to a

wind error of 10-15 knots (DeMaria and Kaplan 1999). The best track data from 1982 to 2003 is used to obtain TC locations and intensity change in this work.

3.2 TRMM observations

Tropical Rainfall Measurement Mission (TRMM) satellite carries several sensors that are useful for tropical cyclone monitoring, including Precipitation Radar (PR), TRMM Microwave Imager (TMI), and Visible Infrared Scanner (VIRS). As an example, Figure 3-1 illustrates the horizontal coverage of VIRS versus PR of the Hurricane Frances (2004).

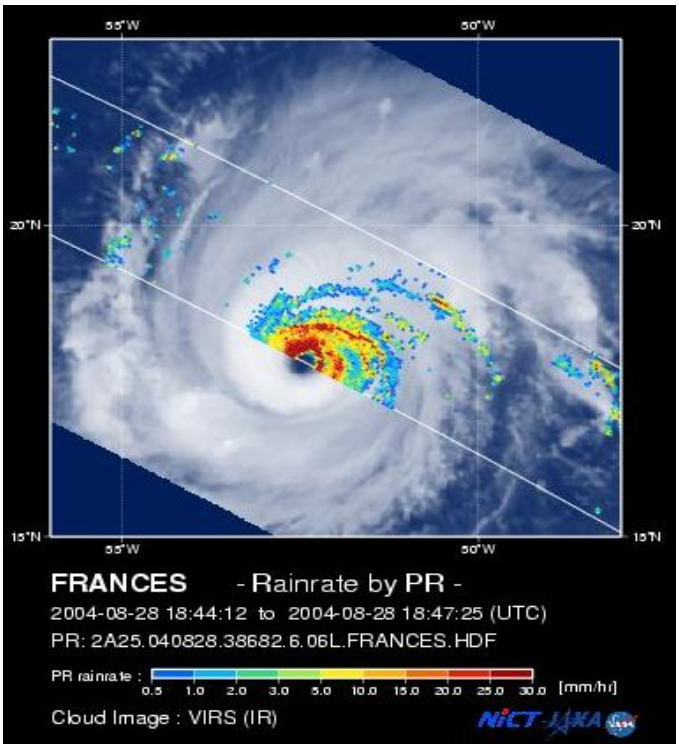


Figure 3-1 Coverage of TRMM VIRS and PR of Hurricane Frances (2004)

Although TRMM VIRS has the widest observation range (the swath width) of about 720 km pre-boost and 833 km post-boost and the finest footprint (pixel resolution) of 2.2 km pre-boost and 2.4 km post-boost, it inherits the incapability of VIS/IR sensors of penetrating the cirrus top to the convective eye wall of a tropical. However, we do not have to worry about the data quality of the PR and TMI microwave-based sensors, when a cloud shields over the cyclone. Meanwhile, since PR can tell the difference among objects at different altitudes, it produced the first real three-dimensional observation of the rainfall structure in a tropical cyclone.

3.2.1 TRMM PR 2A25

TRMM PR is a kind of unique space-borne rain radar that can directly measure the 3-D rainfall distributions over both land and ocean. Before TRMM was boosted on Aug 7, 2001, PR swept 215 km from left to right across track in every 0.6 second, and measured 49 surface pixels in a footprint of 4.3 km x 4.3 km in each scan. After that date, the swath width of PR increased to 247 km and the ground resolution increased to 5 km. Measurements of the columnar rainfall rate of each cubic cell were obtained at 0.25 km vertical resolution from the surface to 20km altitude (TRMM PR Team, 1999; Goddard DAAC 2004).

The level-2 TRMM PR 2A25 dataset in HDF format stores the instantaneous rainfall retrieved from radiation attenuation. It contains scientific datasets (SDS) such as rain rate profile, profile of attenuation corrected reflectivity factor Z, near surface rain, near surface Z, and parameters for the rainfall retrieval algorithm (Goddard DAAC

2004). Each SDS contains at least a two-dimensional matrix with one dimension as the number of sweeps and the other as the number of radar beams in each sweep. Both rain rate profile and Z profile have an additional dimension describing 80 vertical layers evenly distributed from surface to 20 km altitude. This study focuses on the rain rate profile since the rain rate profile is highly correlated to the Z profile. Actually the rain rate profile is derived from the Z profile. The TRMM 2A25 data associated with a particular tropical cyclone over the globe can be found at http://www.eorc.jaxa.jp/TRMM/typhoon/index_e.htm.

3.2.2 TRMM TMI SST product produced by RSS Inc.

The optimally interpolated (OI) SST product (1998-2003) used in this study is derived from TRMM TMI by the Remote Sensing Systems (RSS), Inc., a world leader in processing and analyzing microwave data collected by SSM/I, TMI, AMSR, QSCAT and MSU. The binary data and accessing codes can be downloaded at http://www.remss.com/sst/microwave_oi_sst_data_description.html .

The OI TMI SST dataset covers a region from 40^{o}S to 40^{o}N at the resolution of 0.25 deg (about 25 km). The daily measurements are saved in a separate data file for each day. There are a total of 7 parameters in a daily data file, including time (T), SST (S), 10-meter surface wind using 11 GHz channel (Z), 10-meter surface wind using 37 GHz channel (W), atmospheric water vapor (V), cloud liquid water (L) and rain rate (R). Since TRMM can visit a given location at various local times, the data file contains both ascending and descending observations for every pixel. Therefore the data file contains a

14-layer data cube with the first 7 layers recording the ascending data maps for the 7 geophysical parameters respectively and the second 7 layers recording the descending maps of the corresponding 7 parameters. Some locations at high latitudes can be observed multiple times in a day when TRMM orbit segments overlap within local regions and within a short measurement time. In these cases, data are averaged. However one exception is that when the first and last orbit segments of a day overlap, data at this overlapped region are not averaged. Instead, the later swath data overwrite previous data measured at the beginning of the day.

3.2.3 TRMM data used in this study

Totally 45 hurricanes were recorded in NHC's best track data from 1998 to 2003 in the North Atlantic basin. Most hurricanes were observed multiple times by TRMM except for Hurricane Noel (2001), which was not observed by TRMM at all. Due to the narrow swath of TRMM PR (see Figure 3-1) , sometimes TRMM observes the eye or/and eye wall near a tropical cyclone center, while in other times it only sees part of the outer rain band far away from the eye center. Therefore in this study only 53 TRMM PR observations were selected from a total of 25 hurricanes. Each selected observation captured at least half of the tropical cyclone around its center.

3.3 SHIPS 2003 data

The SHIPS database is used here because of its completeness and its forecast performance. The SHIPS dataset is the most complete dataset so far for hurricane intensification study, not only because of the number of parameters but also the type of parameters it contains.

The SHIPS 1982-2003 database includes information for all 219 tropical cyclones in the Atlantic basin from 1989 to 2003. Additional 59 tropical cyclones from 1982 to 1988 are also included. For each cyclone, the SHIPS 2003 database contains 12-hour synoptic information. More recent SHIPS data use a 6-hour resolution (DeMaria et al. 2005).

The SHIPS 2003 data file records a total of 2 persistence, 3 climatology and 28 atmosphere parameters at every 12-hour intervals (-24hr, -12hr, 0, 12hr, 24hr, …, 120hr) relative to the time and date of a particular TC case. As summarized in Table 3-1, other linear parameters such as the relative Julian date (JDAT), the zonal and meridian storm motions (ZONX and MERY) have to be calculated from the best track data. The potential to change (POT) has to be calculated from the recorded RSST values. Additional 5 quadratic parameters can be easily calculated from those linear parameters. All of the parameters have covered the persistence, climatology, synoptic and nonlinear measurements of a tropical cyclone during its entire lift span, except when the storm track crosses the land (DeMaria and Kaplan 1994b, 1999; DeMaria et al. 2005). All of the linear atmospheric parameters are extracted from NCEP global model analyses, and are measured relatively to the storm center. The storm center was determined from the NHC

best track (DeMaria and Kaplan 1994b, 1999; DeMaria et al. 2005). The persistence parameters include the initial intensity of a tropical cyclone and the past intensity change over different time intervals. The climatology parameters include the cyclone's location, the distance to the nearest land mass, and the relative number of days to the peak of the hurricane season in the Atlantic region. The condition of the presence of warm ocean water was measured by sea surface temperature. The condition of the vertical shear was measured by a set of parameters related to vertical wind distributions. The condition of an unstable atmosphere was measured by the difference of the potential temperature of a lifted air parcel to its environment. The condition of moist atmosphere was measured by the relative humidity at various atmospheric levels. The condition of sufficient vorticity and convergence was measured by the relative momentum flux.

A subset of 20 parameters out of the pool of 37 linear parameters and 5 quadratic parameters were selected by DeMaria and Kaplan (1994b, 1999) and DeMaria et al. (2005) as the most significantly correlated predictors for the future intensity change by a series of F tests. Furthermore after introducing the concept of static versus time dependent parameters, DeMaria et al. (2005) claimed that the SHIPS model after 1999 was a statistical, synoptic and dynamic model and the SHIPS models described in DeMaria and Kaplan (1994b, 1999) were statistical and synoptic only. The value of a static parameter at the initial time (T=0) was used directly to forecast the intensity change at the future time. However the value of a time dependent parameter was averaged over the value at the initial time (T=0) and those along the future storm track. For example, in order to predict the intensity change in the next 24 hours, the time dependent parameter

SHRD was obtained by averaging the recorded value at T=0, T=12, and T=24 from the SHIPS data file. The averaged SHRD value is then used to predict the intensity change in the next 24 hours. On the other hand the SHIPS model used the value of the static parameter PSLV at the initial time T=0 to predict the intensity change in the next 24 hours. Table 1 in DeMaria et al. (2005) provided a complete list of the most significantly correlated static and time dependent predictors. It is worth to point out that REFC had been listed twice, one as a static value, the other as a time dependent value.

In this study, in order to simplify the comparison of our analysis with DeMaria's work (i.e., DeMaria and Kaplan 1994b, 1999; DeMaria et al. 2005), the same 20 parameters and the meridian storm motion (MERY) were selected initially for the weakening and intensifying rules in stratified TCs discussed in Chapter 5 and 6. All of these parameters were treated as static parameters. That is, we only seek for the relationships between the values at the initial time and the future intensity change in the next 12 hours. Names and corresponding physical definitions of selected predictors are listed in Table 3-1. Additional parameters are used for studying rapid intensifications, which will be discussed in Chapter 7.

Table 3-1 Selected SHIPS parameters

No.	Name	Description	Unit
1	**JDAT**	**Absolute value of (Julian Day – peak season value)**	**day**
2	*JDTE*	*Gaussian function of (Julian Day-peak value)*	*none*
3	V0	Initial maximum winds (current intensity)	knot
4	IV12*	Maximum wind change during the past 12 h	knot
5	*VV*	*Initial max winds times previous 12 h change*	*knot*knot*
6	PSLV*	Pressure level of storm steering	mb
7	D200	200 hPa divergence	10^5 s^{-1}
8	REFC*	200 hPa relative eddy momentum flux convergence	m/sec/day
9	**POT***	**Maximum possible intensity – current intensity**	**knot**
10	SHRD*	850-200 hPa vertical shear	knot
11	U200*	200 hPa zonal wind	knot
12	T200	200 hPa temperature	^oC
13	RHLO*	850-700 hPa relative humidity	%
14	RHHI	500-300 hPa relative humidity	%
15	Z850	850 hPa relative vorticity	$10^5 s^{-1}$
16	EPOS	Surface–200 hPa θ_e deviation of lifted parcel	^oC
17	*SRLA*	*Vertical shear times sine of storm latitude*	*knot*
18	*POT2*	*Square of potential-current intensity*	*knot*knot*
19	*SRV0*	*Initial intensity times shear*	*knot*knot*
20	**ZONX***	**Zonal component of storm motion**	**km/(12hours)**
21	**MERY**	**Meridian component of storm motion**	**km/(12hours)**
22	LAT*	Latitude of TC centers	degree
23	LON*	Longitude of TC centers	degree
24	SST*	Sea Surface Temperature	^oC

Note:
Parameters from No.1 to No. 21 are used in the stratification study.
Parameters with asterisk (*) are the 11 parameters used in the rapid intensification study.
The parameters in the regular font are those with values given in SHIPS database.
The **boldfaced** ones represent derived parameters for which values should be estimated from Best track data.
The *italicized* ones are the 5 quadratic parameters which can be computed based on other parameters.

Among those parameters, most of values are given in SHIPS database directly. Five of them, JDAT, JDTE, POT, ZONX, and MERY, are derived from these direct values based on the same procedures for SHIPS model. JDAT is obtained by subtracting the peak Julian data for Atlantic TC activities, 253, from the Julian date of the current TC (DeMaria et al. 2005). The quadratic parameter JDTE is simply the Gaussian function of JDAT with standard deviation of 25 days based on DeMaria et al. (2005). For POT values, the differences between the maximum possible intensity (MPI) and the current intensity, the empirical relationship between MPI and SST established by DeMaria and Kaplan (1994a, b) was used. The zonal component (ZONX) and meridional component (MERY) of the TC motion are estimated based on the respective longitude and latitude changes in the past 12 hours using NHC best track data. The values of other quadratic parameters in Table 3-1, VV, SRLA, POT2, and SRV0, are obtained by simple mathematical functions of the known parameters, either given in SHIPS database or derived as described above.

The SHIPS model is based on multiple linear regression technique. Significant predictors (i.e., parameters) are chosen based on standard F test at 1% significant level for specific forecasting periods. Essentially, the predictor selection procedure used by SHIPS is a one-to-one procedure based on the correlation between TC intensity changes and a specific parameter. One objective of this study is to explore a deeper understanding of the interactions among various physical processes that may affect hurricane intensification, i.e., to discover “multiple-to-one” associations among a large number of factors controlling TC intensities. Data mining techniques or more specifically, an

association rule algorithm is applied in later chapters to explore the potential combinations of the conditions.

3.4 Association rule algorithm

Association rule induction (Agrawal et al. 1993) was developed originally for market basket analysis, which aims at finding regularities in the shopping behavior of customers. An association rule is a rule like "Z←X, Y." The items X and Y are called antecedents in the rule and Z is the consequent. This rule expresses an association between items X, Y, and Z. It states that if a randomly selected customer picked items X and Y, it is likely that the customer also picked item Z. The number of antecedents can range from one to the total number of items in a database.

Usually, three parameters, i.e., support, confidence, and lift, are reported for mined association rules. The support estimates the probability P({X,Y,Z}), and the confidence estimates the probability P(Z|{X, Y}). An association rule "Z←X, Y" is strong if it has a large support and a high confidence. The third parameter, the lift (Silverstein et al. 1998) was introduced as the ratio between the actual probability of the item set containing both antecedent and consequent divided by the product of the individual probabilities of the antecedent set and the consequent. That is, lift = P({X,Y,Z})/[P({X,Y})*P(Z)]. The lift measures the dependency among the antecedents and the consequent of a rule. Lift values greater than 1 indicate positive dependence, while those less than 1 indicate negative dependence.

The association rule algorithm used in this study is implemented by Borgelt (2009). The support value in this implementation is defined as P({X,Y}) instead of P({X,Y,Z}). The practical description of the rule measurements (support, confidence, and lift) for this application is postponed to a later section with a real rule for TC intensity change.

CHAPTER 4 Structures of Latent Heat Release in Tropical Cyclone Intensity Change

Accurate tropical cyclone (TC) intensity prediction is a challenging research topic. Many reasons can be accounted for the poor skill of the TC intensity forecasting (Wang and Wu 2004), such as inadequate observations on TC structures, deficient understanding of physical processes governing the intensity change, coarse model resolution and poor initial conditions, etc. Efforts have been conducted to find additional parameters using satellite images. Studies using MW sensors such as Nimbus ESMR (Rodgers and Adler 1981) and SCAMS (Kidder et al. 1978; Velden and Smith 1983), DMSP SSM/I (Rao and McCoy 1997; Rao and MacArthur 1994a, 1994b; Cecil and Zipser 1999; Rodgers et al. 1994; Chang et al. 1995), TRMM TMI (Hoshino and Nakazawa 2004; Nakazawa and Hoshino 2004) and TRMM PR (Kelley and Stout 2004) have obtained concordant results that the rain rate, latent heat release (LHR) and the brightness temperatures (TB) correlate to the future (24-72 h) tropical cyclone intensity.

Since TRMM is the first satellite that captures the 3-dimentional rain rate structure within a tropical cyclone, we believe that its 3D rainfall product will be informative to TC intensity forecast. For example, when using the rain rate as the indicator of the eye of a tropical cyclone, the 3D structure of eye wall, including the height of the eye wall, the size of the eye, the slope of the eye wall as well as the eye wall asymmetry, can be

extracted. This chapter discusses these features extracted from TRMM 2A25 data and their association with the tropical cyclone intensity change. TSDIS orbit viewer is used here to provide most of the basic functionalities (Tang et al. 2005).

4.1 A case study of the 3D structure of TC eye wall

Due to the limited observations from TRMM PR a case study of Hurricane Isabel (2003) is performed first to demonstrate the eye wall structure change over TC lifetime. The recorded life of Hurricane Isabel (2003) from the Best Track data covers a total of 15 days with one observation per 6 hours. However there are only 6 TRMM PR observations that captured at least half of the tropical cyclone around its center. Figure 4-1 grouped the 6 TRMM PR observations of Hurricane Isabel (2003), along with its intensity change from the Best Track data.

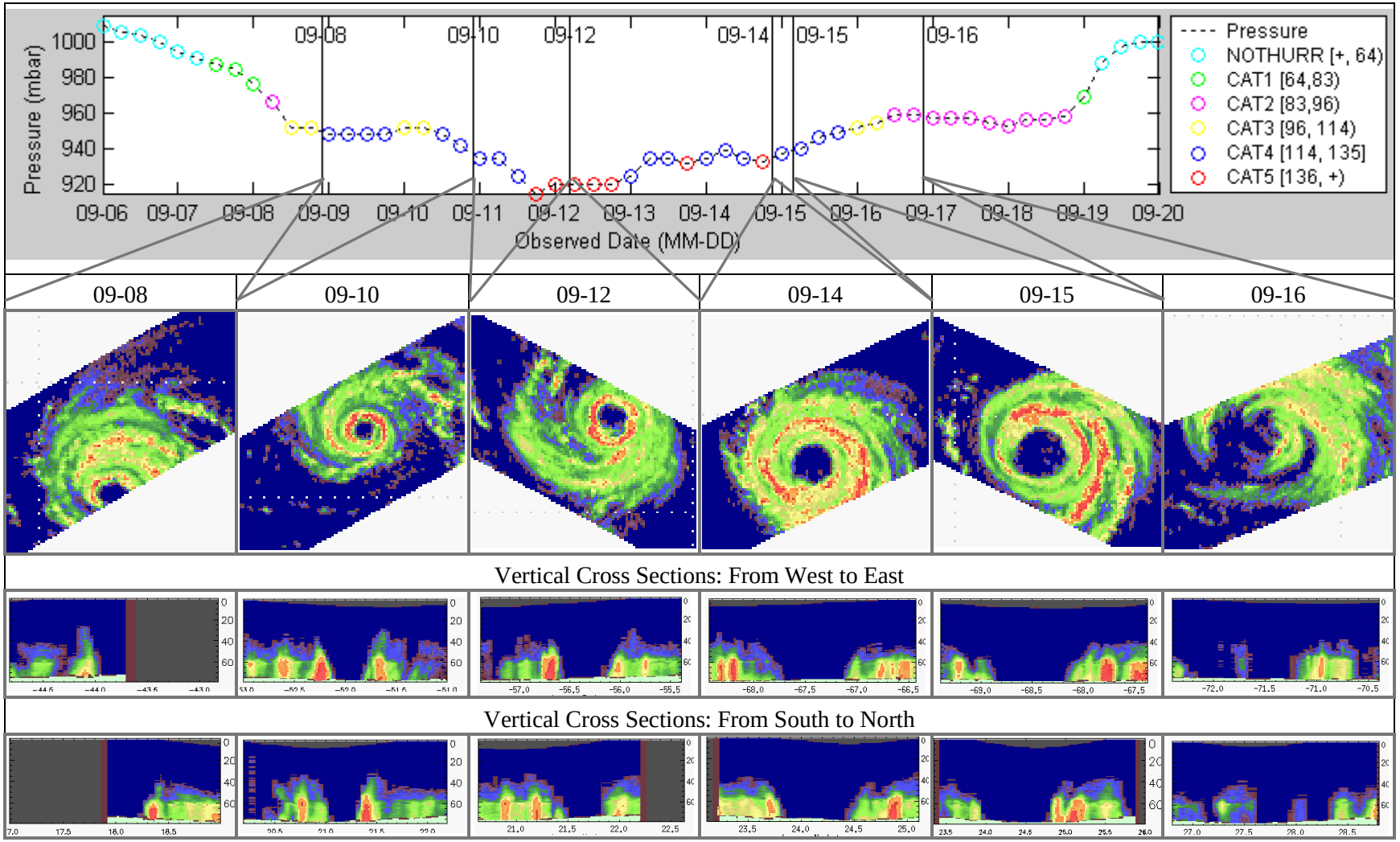


Figure 4-1 Hot towers during the lifetime of Hurricane Isabel (2003) captured by TRMM PR

The top panel in Figure 4-1 shows the central pressure change of Hurricane Isabel with time. The 6 vertical lines represent the time when the 6 TRMM observations were made. The right box gives the legend of the central pressure and 6 stages of tropical cyclone strength of Hurricane Isabel (2003) based on the Saffir-Simpson Scheme. The minimum central pressure data are extracted from the Best Track data. Different color schemes are used to distinguish the stages of hurricane development, including tropical storm, category 1 hurricane, category 2 hurricane, and so on. On September 12, 2003 Hurricane Isabel reached its maximum strength as a category 5 hurricane. The maximum wind is used to classify hurricane stages in the top panel. That also explains why the colors do not exactly match the value of minimum central pressure. The mismatch also shows that the relationship of max wind and the central pressure is not monotonic.

The second panel in Figure 4-1 shows the 2D near surface rain rate from the six TRMM PR observations using TSDIS orbit viewer. The warmer the color, the higher the rain rate is. The convective ring can then be easily identified from the snapshots. From this panel we can find that the inner convective ring expanded then dissipated when Hurricane Isabel weakened. This finding verified the conclusion in the previous work by Rodgers, Chang and Pierce (1994). Using a dynamic model in their study, they found out that tropical cyclone intensification was observed when the convective rings propagated into the inner core of a tropical cyclone, and tropical cyclone weakening was observed when these inner-core convective rings dissipated.

The last two panels in Figure 4-1 contain the cross sections of the vertical rain rate (a.k.a., columnar rain rate) around the eye in the North-South direction and West-East

direction, for each TRMM observation. The x axis is the longitude for the 3rd panel, and is the latitude for the 4th panel in degree. The y axis is the altitude in km. One can see that while Hurricane Isabel (2003) was intensifying on 10 September 2003 and on 12 September 2003, the inner eye wall had the largest rain rate, but when it was weakening from 14 September to 16 September, the largest rain column moved out of the inner eye wall. Looking at the slope of the inner eye wall, one can see that when Hurricane Isabel was growing, that is, from 10 September 2003 to 14 September 2003, the eye wall was steep, and when it was decaying, as shown in pictures from 14 September to 16 September 2003, the eye wall slope reduced. On the observation of 16 September 2003, the inner eye wall had deconstructed. In these two cross section panels one can also see that the slopes of eye walls are asymmetric. The tilting of eye walls may indicate the existence of a vertical shear of horizontal wind, which, argued by both earlier observational and modeling studies (Wang and Wu 2004), inhibits the development of the cyclone.

Using the similar layout as in Figure 4-1, Figure 4-2 gives us a direct view of the rain rate change over the life time of Hurricane Isabel (2003). The top panel is the same as in Figure 4-1, displaying the intensity variation of Hurricane Isabel (2003) over time. The middle panel shows the 3D snapshot of columnar rain rate of Hurricane Isabel (2003) when an observation cube has a rain rate no less than 20 mm/hr. This panel clearly shows the formation and dissipation of the convective ring inside Hurricane Isabel (2003).

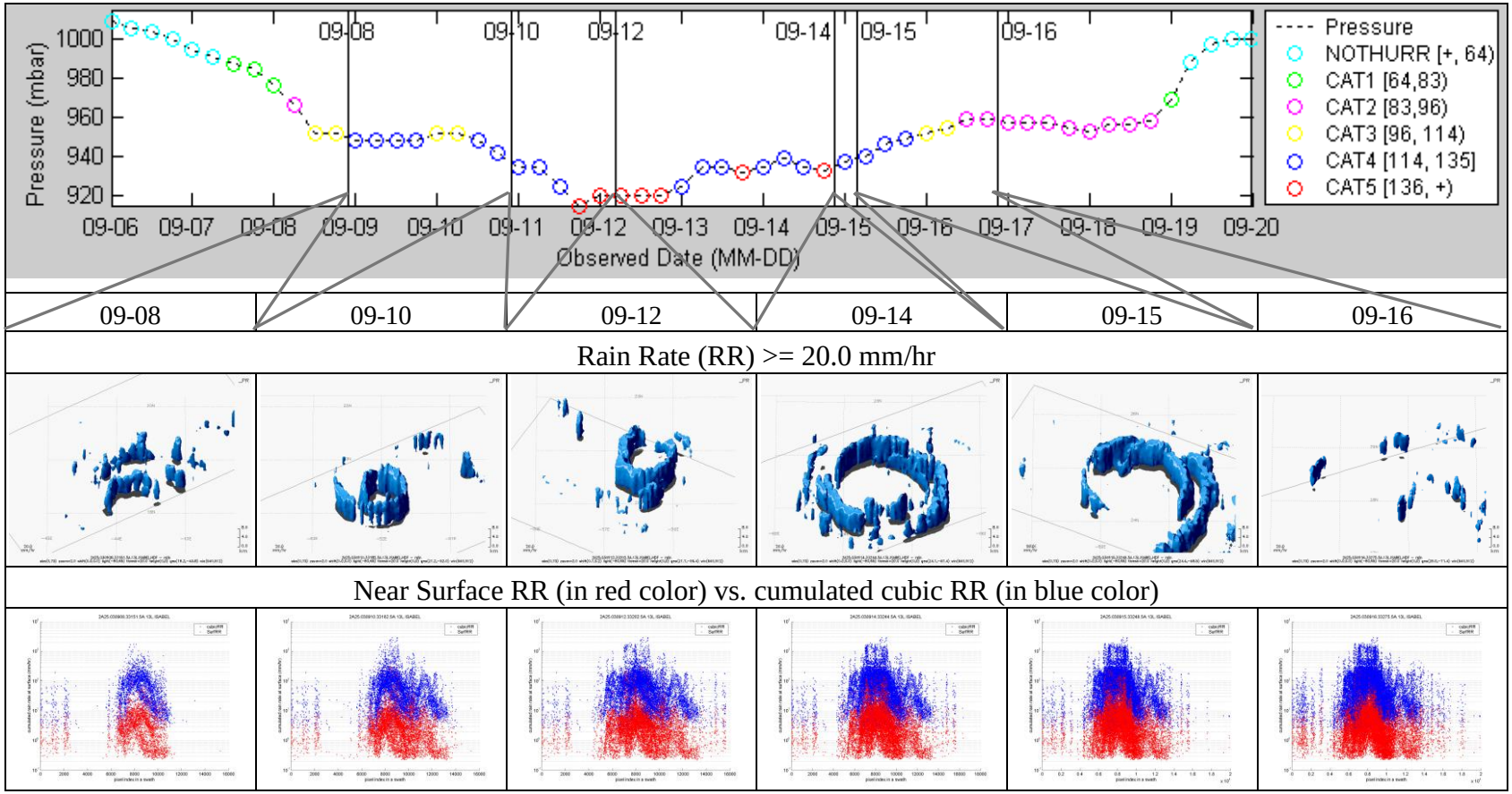


Figure 4-2 Cumulated Rain Rate over the lifetime of Hurricane Isabel (2003)

The bottom panel in Figure 4-2 shows the relation of surface rain rate and the cumulated cubic rain rate in every surface cell. The x axis is the index of each observation cell on the surface. The y axis is the log of the rain rate value. The surface rain rate is in red, and the cumulated cubic rain rate is in blue. Our assumption here is that when a tropical cyclone is at the intensification stage, it is accumulating energy to grow therefore the entire column of rain will not be released as the surface rain. However when a tropical cyclone is at the weakening stage, the energy will be released with rain and wind, thus the amount of the near surface rain will be close to the sum of the entire column of rain cubes. In other words, the difference of the sum of columnar rain rate and the surface rain rate may indicate the potential of intensifying. In the first two pictures when Hurricane Isabel (2003) was observed on September 8, 2003 and on September 10, 2003, we can clearly see the white gap between the cumulated cubic rain rate and the surface rain rate.

A convective tower was defined in Kelley and Stout (2004) as a convective cell with at least 2 mm/hr precipitation rate at an altitude higher than 14 km. Following the same definition of the convective cell, the maximum height of convective towers in the 6 TRMM PR observations of Hurricane Isabel (2003) was extracted and shown as the blue bars in Figure 4-3. The distance from the hot tower to the center of the eye was extracted and shown as yellow diamonds. In order to correlate the hot tower data with hurricane intensity change, the intensity measured by the maximum wind was displayed as red dotted line in Figure 4-3.

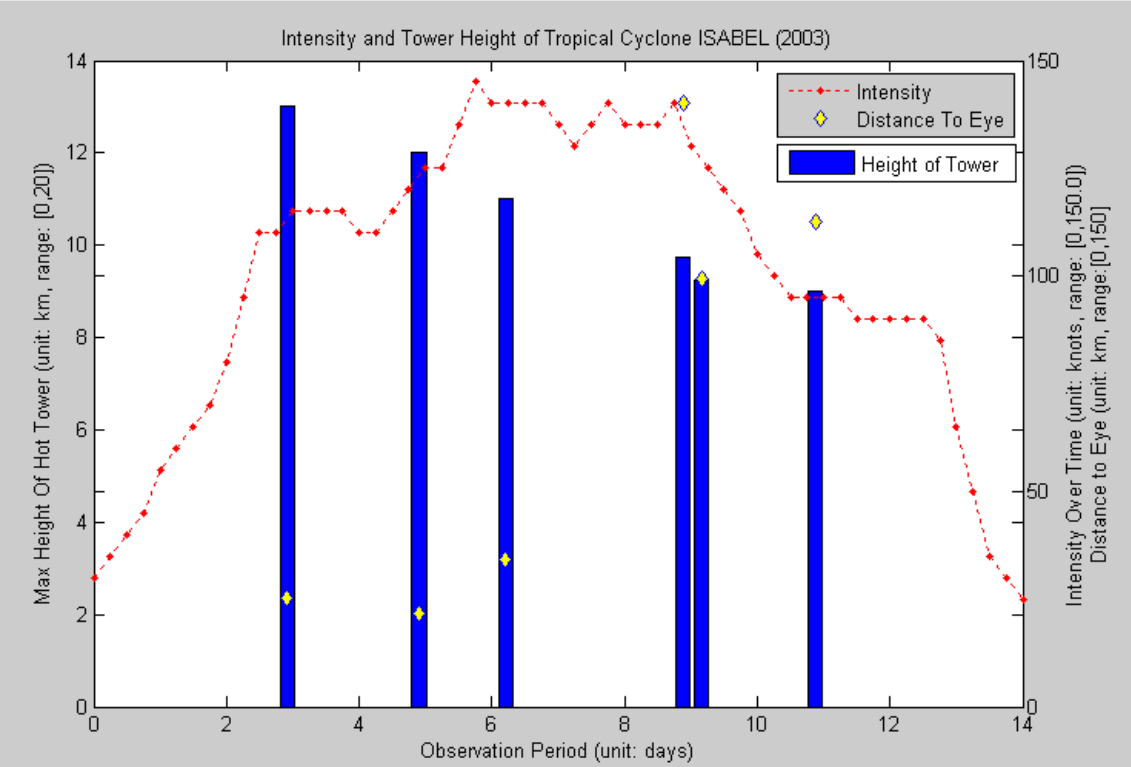


Figure 4-3 Structure of inner convective ring of Hurricane Isabel (2003)

One can easily find out from Figure 4-3 that the height of the hot towers keeps decreasing over the lifetime of Hurricane Isabel (2003). However the hot tower had moved closer to the center when Hurricane Isabel (2003) was growing. When hurricane was decaying the hot tower was rather far away from the eye center, generally speaking.

Overall the result of this case study is very promising. We verified the expansion of convective ring when Hurricane Isabel (2003) decayed. This means that the energy (mostly the released latent heat from precipitation) distribution in horizontal dimensions may be a good feature for intensification research. When using the columnar rain as the measurement of eye wall, the size of eye, the slope of eye wall and the eye wall asymmetry can also be

identified from TRMM PR 3D observations. The difference of the cumulated columnar rain rate and the surface rain rate may indicate the potential ability of TC intensification. The height of the hot tower and its distance to the TC center also were extracted from TRMM PR data. They showed interesting correlation with TC intensity change. However to extract those measurements from TRMM PR observations for supplying the real-time TC forecast, a major barrier is the limited coverage of TRMM observation over the lifetime of a TC due to the small swap width of TRMM PR.

4.2 Feature extraction of the latent heat release near the TC center

Encouraged by the above case study of Hurricane Isabel (2003), several features related to the structure of the latent heat release near the TC center were extracted from TRMM dataset.

As explained in Chapter 3, although there were 45 TCs in North Atlantic basin from 1998 to 2003, only 53 TRMM PR observations for 25 TCs covered at least half of TC center due to the narrow swap width of the TRMM PR instrument. Using these 53 TRMM PR observations and the Best Track data, a total of 13 attributes were extracted (see Table 4-1). The first 12 attributes in Table 4-1 are used as independent variables and the last attribute (DV6) as the dependent variable. All of the attributes are interpolated, if necessary, to the TRMM observation time. They are organized as a relational table with the 13 attributes as columns and the 53 observations as rows. The data sources of each attribute are also listed in Table 4-1.

Table 4-1 Attributes used in the latent heat release study

Attr.	Description	Source
JDT	Julian date of the observation – peak value (253), unit: day, range: [-58, 67]	TRMM metadata
JHR	UTC hour of the observation, unit: hour, range: [0.067, 22.57]	TRMM metadata
LAT	Latitude of the TC center at the observation, unit: degree, range: [14.11, 35.47]	TRMM PR 2A25
LND	Distance from TC center to nearest land mass, unit: km, range: [0.00, 1739.49]	TRMM TMI RSS
SST	Averaged sea surface temperature within 1^o of the TC center at the observation time, unit: ^{o}C, range: [24.53, 30.28]	TRMM TMI RSS
HGH	Max height of rain cells with rain rate >= 2mm/hr and inside TC at the observation time, unit: km, range: [8.50, 17.75]	TRMM PR 2A25
CRR	Averaged rain rate within 111km radius of the TC center at the observation time, unit: mm/hr, range: [0.804, 15.183]	TRMM PR 2A25
CPP	PIP index (the fraction of rain rate > 5 mm/hr over the total rain rate) within 111km radius of the TC center at the observation time, unitless, range: [0.286, 0.966]	TRMM PR 2A25
ORR	Averaged rain rate within the 111km to 222km annuli around TC center at the observation time, unit: mm/hr, range: [0.121, 6.74]	TRMM PR 2A25
OPP	PIP index within the 111km to 222km annuli around TC center at the observation time, unitless, range: [0.184, 0.888]	TRMM PR 2A25
SPD	Averaged translation speed in the past 6hr to the observation time, unit: km/hr, range: [2.41, 33.18]	NHC Best Track
V0	The TC intensity at the observation time, unit: knots, range: [45.28, 140.0]	NHC Best Track
DV6	The intensity change from the observation time to 6 hours later, unit: knots, range: [-23.93, 11.83]	NHC Best Track

Among these 13 attributes, the observation date (JDT) and the observation time (JHR) are coded in TRMM PR 2A25 metadata. Since the measurements from TRMM PR are

fluctuated from daylight to night time, the observation time (JHR) is included in this dataset to accommodate the impact.

The center location is critical information if we want to extract any measurements of the eye wall structure. In this study, the center location of a TC is manually identified from TRMM PR 2A25 images via TSDIS orbit viewer.

Using the center information, the latitude of the TC location (LAT) is extracted in our study because we believe that the Coriolis force varies at different latitude and its strength may be one factor to affect the TC movement. The averaged SST within the 1^{o}-radius area around the center location is computed from the optimally interpolated TMI SST products. The distance from the center location to the nearest land mass (LND) is calculated using the land masks in the OI TMI SST product.

The vertical rain rate profile SDS in the TRMM PR 2A25 product is used to extract the maximum height of rain cells with rain rate > 2 mm/hr (HGH). Following the method in Chang et al. (1995), the near surface rain SDS dataset in TRMM PR 2A25 data product is used to calculate four rain rate attributes to describe the latent heat release in the inner core region – the rain rate at the TC core area (CRR), the PIP index (Precipitation Intensity Parameter) at the TC core area (CPP), the rain rate at TC outer rings (ORR), and the PIP at TC outer rings (OPP). The center location information is used to form the 111km-radius circle as the inner band and 111km-to-222km annuli around the center as the outer band of a TC during the calculation.

Other attributes such as the current intensity in terms of wind speed (V0), the moving speed of the TC (SPD), and the intensity change in the next 6 hours (DV6) are calculated from NHC's best track data.

4.3 Association Rule mining of the structure of latent heat release

4.3.1 Data Pre-processing

The attribute V0 was broken down into 6 ranges according to Saffir-Simpson scale: (1) tropical storm $V0 \in [26,63] knots$, (2) hurricane category 1 when $V0 \in [64,82] knots$, (3) hurricane category 2 when $V0 \in [83,95] knots$, (4) hurricane category 3 when $V0 \in [96,113] knots$, (5) hurricane category 4 when $V0 \in [114,135] knots$, and (6) hurricane category 5 when $V0 > 135 knots$. There are totally 6 fields used to describe the attribute V0. Here is only an illustration. Suppose the attribute V0 has a value of 65 knots, it can be represented as "01 0000" where the second field has a "1" and other 5 fields have "0"s indicating V0 falls into the [64, 82] knots range.

The remaining 11 predictive attributes were discretized and transformed into 3 fields each, using the following method. Each of them (X) was broken down into three ranges by its own mean (m) and half of its standard deviation (s) over the entire observation sets. That is, one range for high values ("HI") when $X \geq m + \frac{s}{2}$, one range for low values ("LO") when $X \leq m - \frac{s}{2}$, and the third for normal values ("N") that $m - \frac{s}{2} < X < m + \frac{s}{2}$.

The dependent variable DV6 was categorized into 3 groups, including the weakening group when $DV6 \le -5knots$ which contains 12 records, the intensifying group when $DV6 \ge 5knots$ which contains 15 records, and the stable group when $-5knots < DV6 < 5knots$ which contains 26 records. Table 4-2 lists the mean and half of the standard deviation for each predictive attribute except V0.

Finally the 13 numeric attributes were transformed into a total of 42 Boolean fields in a relational table with each of the 53 observations as a row. The transformed data can then be used in any of the association rule discovery implementations.

Here we have chosen the implementation by Dr. Christian Borgelt. His program (http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html) takes the transformed data as input and outputs association rules or frequent item sets. The program was written in C. We used its Windows version.

Table 4-2 Statistics of the 11 predictors in latent heat release study

Attr.	*mean*	*std/2*
JDT	0.9623	11.9968
JHR	11.5999	3.2724
LAT	27.5170	3.1544
LND	557.3287	221.3429
SST	28.2219	0.6263
HGH	12.5566	1.2021
CRR	5.1356	1.4938
CPP	0.7773	0.0574
ORR	2.1999	0.6708
OPP	0.6278	0.0713
SPD	19.0881	3.5514

4.3.2 Mined Association Rules

Our target is to find out the associations among multiple attributes and the intensity change in next 6 hours, so the rules are restricted to those with only attribute DV6 appearing in the consequent.

Using the minimum support as 7%, the minimum confidence as 80% and the minimum lift as 100%, 2 rules were found which indicating at what kind of a situation a TC will intensify in the next 6 hours. Similarly, 24 rules were found for weakening situations and 582 rules for stable situations. Table 4-3 listed rules for intensifying and weakening cases when the minimum support is 7%, the minimum confidence is 80% and the minimum lift is 100%. However in order to control the number of the stable rules, the thresholds of support and confidence were changed to 10% and 100% respectively, which resulted in 2 of such rules (Table 4-3).

Table 4-3 Rules for TC Intensifying/Weakening/Stable in next 6 hours

Intensifying Situation (s=7%, c=80%, l=100%)	
DV6=I <- HGH=H SPD=N (7.5, 80.0, 353.3)	IN.1
DV6=I <- CRR=L ORR=L SPD=N (7.5, 80.0, 353.3)	IN.2
Weakening Situation (s=7%, c=80%, l=100%)	
DV6=W <- V0=H4 CRR=H (7.5, 80.0, 282.7)	WK.1
DV6=W <- V0=H4 LAT=L (7.5, 80.0, 282.7)	WK.2
DV6=W <- V0=H4 JDT=N (7.5, 80.0, 282.7)	WK.3
DV6=W <- HGH=H CRR=N (7.5, 80.0, 282.7)	WK.4
DV6=W <- SPD=L LAT=H (7.5, 80.0, 282.7)	WK.5
DV6=W <- LND=N CPP=H (7.5,80.0,282.7)	WK.6
DV6=W <- V0=H1 HGH=L (7.5, 80.0, 282.7)	WK.7
DV6=W <- HGH=L OPP=H (7.5, 80.0, 282.7)	WK.8
DV6=W <- SPD=H JHR=N (7.5, 100.0, 353.3)	WK.9
DV6=W <- LAT=H OPP=H (9.4, 83.3, 294.4)	WK.10
DV6=W <- SST=L OPP=H (9.4, 83.3, 294.4)	WK.11
~~DV6=W <- V0=H4 CRR=H LAT=L (7.5, 80.0, 282.7)~~	WK.12
~~DV6=W <- V0=H4 CRR=H JDT=N (7.5, 80.0, 282.7)~~	WK.13
~~DV6=W <- V0=H4 LAT=L JDT=N (7.5, 80.0, 282.7)~~	WK.14
~~DV6=W <- HGH=H JDT=H CRR=N (7.5, 80.0, 282.7)~~	WK.15
~~DV6=W <- HGH=H CRR=N LND=L (7.5,80.0,282.7)~~	WK.16
~~DV6=W <- HGH=L V0=H1 LAT=H (7.5, 80.0, 282.7)~~	WK.17
DV6=W <- HGH=L JHR=N LAT=H (7.5, 80.0, 282.7)	WK.18
DV6=W <- V0=H1 LAT=H ORR=N (7.5, 80.0, 282.7)	WK.19
DV6=W <- LAT=H SST=L CPP=H (7.5, 100.0, 353.3)	WK.20
DV6=W <- LAT=H SST=L OPP=H (9.4, 100.0, 353.3)	WK.21
DV6=W <- JDT=H CRR=N LND=L (7.5,80.0,282.7)	WK.22
~~DV6=W <- V0=H4 CRR=H LAT=L JDT=N (7.5, 80.0, 282.7)~~	WK.23
~~DV6=W <- HGH=H JDT=H CRR=N LND=L (7.5,80.0,282.7)~~	WK.24
Stable/Neutral Situation (s=10%, c=100%, l=100%)	
DV6=S <- JDT=L SST=H JHR=N (11.3, 100.0, 203.8)	NT.1
DV6=S <- JDT=N CRR=N CPP=N (11.3, 100.0, 203.8)	NT.2

As an illustration, the first rule in Table 4-3 ("DV6=I <– HGH=H SPD=N (7.5, 80.0, 353.3)") can be interpreted as "when the translation speed of a TC is in normal range (between 15.53 km/hr to 22.64 km/hr) and the maximum height of rain cells reaches the high range (higher than 13.77 km), a TC will intensify in the next 6 hours with 80% probability."

In Table 4-3 the rules with 100% confidence are highlighted in gray shade, are of the highest interest in this study. On the other hand the redundant rules are marked as strikethrough since they are not the most concise rules, i.e., the rules are implied by other rules, and will not be discussed in this study.

From Table 4-3 we can see that the favorable situations that help a TC intensify in next 6 hours are:

- IN.1 – when the maximum height of rain cells of a TC reaches the high range (HGH >= 13.75 km), and the translation speed of a TC is in normal range (between 15.53 km/hr to 22.64 km/hr), the TC will intensify in the next 6 hours with about 80% probability.
- IN.2 – when a TC moves in normal speed (SPD between 15.54 km/hr and 22.64 km/hr), the latent heat release at the core area is low (CRR <= 3.64 mm/hr), and the latent heat release at the outer band is low (ORR <= 1.53 mm/hr), the TC will intensify in the next 6 hours with 80% probability.

The most adverse situations that cause a TC to weaken are:

- WK.9 – When a TC moves very fast (SPD >= 22.64 km/hr) and the TRMM observation time is the UTC (JHR) time between 8.32 and 14.86 UTC, the TC will decay in the next 6 hours with 100% probability. Note that UTC is not a local time.

However in this study since we focus on the North Atlantic TCs, and the local time is about 3 to 5 hours later than the UTC, and the UTC 8-13 leads to local time period 3-12am. As a result, we can say that the JHR condition in this rule means roughly the local morning time.

- WK.20 – When a TC moves to a high latitude area (LAT >=30.66 ^{o}), the SST is low there (SST <= 27.66 ^{o}C), and there are very intensive latent heat release in the inner band (CPP >=0.83), the TC will decay in the next 6 hours with 100% probability.
- WK.21 – When a TC moves to a high latitude area (LAT >=30.66 ^{o}), the SST is low (SST <= 27.66 ^{o}C) there, and there are very intensive latent heat release in the outer band (OPP >= 0.70), the TC will decay in the next 6 hours with 100% probability. This weakening rule has proved Bister's argument that latent heat release in the outer rainbands may have adverse effect on tropical cyclone intensity (Bister 2001).

The conditions which do not significantly affect TC intensity change in the next 6 hours are:

- NT.1 – When a TC is developed in the early hurricane season (JDT <= -11.03 day), even though it enters an area with high SST (SST >= 28.84 ^{o}C) around local morning time (JHR between 8.32 and 14.86 UTC), this TC will likely keep the same intensity.
- NT.2 – When a TC is developed around the peak of hurricane season (JDT between -11.03 and 12.96 day), the latent heat release at the inner band is moderate (CRR between 3.64 and 6.62 mm/hr) and the percentage of very intensive rain among the total rain in the inner band is moderate (CPP between 0.72 and 0.83), it will likely maintain the same intensity.

Overall, the discovered rules have shown that TC intensification requires to avoid the adverse conditions such as (1) entering a higher latitude area where the underlying sea is relatively cold, (2) moving too fast to absorb the warming energy from the underlying sea, or (3) strong energy loss at the outer band. The energy buildup of a TC can be indicated by the height of the convective tower and by the relatively low latent heat release in the core area and in the outer band. When the adverse conditions and the amicable conditions reached some kind of balance, the tropical cyclone intensity would keep stable.

4.4 Discussion

One may notice that we chose a low value for the minimum support, i.e., 7%. This is resulted from the process of transforming a numeric attribute to several Boolean fields. If the number of intervals for a numeric attribute is large (here is 3 for most of the attributes), the support for any individual interval can be low. Hence we have to decrease the minimum support in order to select some rules involving this attribute. This reduced-support problem is inherited in every discretization situations. Some scientists have already been working on this issue (Srikant and Agrawal 1996) and we will consider their results in our future work.

Discretization is a crucial step in our process. A different set of discretization criteria has been tested using the same data set. Instead of using the half of the standard deviations of the 11 predictors, we use the full length of the standard deviations to separate them around their means. Nine rules were found for favoring TC intensification. Twenty three rules were found related to TC weakening, and 497 rules for keeping TCs stable when the minimum support is 7%, minimum confidence is 80% and minimum lift is 100%.

In Table 4-3 some rules are associated with each other such as the three rules for weakening TCs in Table 4-3 "(WK.1) DV6=W <- V0=H4 CRR=H (7.5, 80.0, 282.7)," "(WK.2) DV6=W <- V0=H4 LAT=L (7.5, 80.0, 282.7)," and "(WK.12) DV6=W <- V0=H4 CRR=H LAT=L (7.5, 80.0, 282.7)". Both antecedents in the first two rules are different subsets of the antecedent in the third rule. However the support and confidence are the same across all of these 3 rules. Although the 3rd rule is redundant based on the principle of concise rule pruning, we really do not know if DV6=W is the consequence of a strong interaction of the triplet < V0=H4 CRR=H LAT=L > or because of a combination of two attributes (eg. < V0=H4 CRR=H > or <V0=H4 LAT=L >). Wu et al. (2003) proposed a method to select the multi-item associations that cannot be explained by the subset associations in the item set. Their method fits the frequent item sets to a log-linear model by classic statistical theory and uses the fitted coefficients to determine the importance or interestingness of a frequent item set. This method should be considered in our future study to identify interesting rules.

Considering the fact that the Apriori-base algorithm found much more rules for situations to keep TCs stable than those to intensify or weaken TCs, it is natural to ask why this happened. The major reason is that there are more training cases for stable TCs than intensifying or weakening TCs under the current discretization scheme on attribute DV6. There are 26 cases among the total of 53 observations for stable TCs, 12 cases for intensifying TCs, and 15 cases for weakening TCs. Although we can change the separation criteria to increase cases for weakening TCs and intensifying TCs, we did not do so because the nominal error in the NHC best track intensity records is 5 knots (Jarvinen et al. 1984). In the future we plan to increase the size of the entire data set by adding more observations.

CHAPTER 5 TC stratification and data preprocessing

Although the structural parameters of latent heat release extracted from TRMM PR data are very promising, we are limited to a small amount of available observations. Therefore, SHIPS 2003 dataset was explored to understand the internal coupling of the environmental factors and the tropical cyclone intensity change.

5.1 TC Stratification

The first step of data preprocessing is to remove records with missing data for any of selected parameters. After removing entries with missing values, there are a total of 2,003 observation records. Since TC intensity changes depend on initial intensity, the 2003 observations were stratified to seven groups, including Tropical Depression (TD), Tropical Storm (TS), and the five categories of hurricanes (H1-H5) based on the Saffir-Simpson Hurricane Scale (NOAA/NWS 2007). Data mining techniques are then exercised on stratified TC records. The numbers of records for each group in the total of 2003 records are listed in Table 5-1.

To explore the factor combinations affecting TC intensity changes with association rule data mining technique, TC cases are further separated into intensifying cases and weakening cases. Since there is a nominal error of 5 knots in the intensity records, TCs with intensity

decrease of 5 knots or greater in the next 12 hours are considered as weakening cases. In the opposite, TCs with intensity increase of 5 knots or greater are classified as intensifying cases. The records with intensity change between -5 knots to +5 knots in the next 12 hours are considered as stable cases. The number of intensifying and weakening cases and the sample percentage values for each individual group are listed in Table 5-1. For example, there are a total of 478 records of category 1 hurricanes, among which 168 (35.1%) records were weakening, and 181 (37.9%) records were intensifying.

Table 5-1 Number of total, weakening, and intensifying cases of stratified TCs

Category	Total	Weakening(%)	Intensifying(%)
Tropical Depression (TD)	312	41(13.1)	80(25.6)
Tropical Storm (TS)	841	253(30.1)	348(41.4)
Category 1 Hurricane (H1)	478	168(35.1)	181(37.9)
Category 2 Hurricane (H2)	174	85(48.9)	50(28.7)
Category 3 Hurricane (H3)	97	42(43.3)	38(39.2)
Category 4 Hurricane (H4)	93	43(46.2)	22(23.7)
Category 5 Hurricane (H5)	8	6(75.0)	1(12.5)

Note:

The numbers in parentheses are percentages of weakening and intensifying cases in each category.

Looking at the case numbers in Table 5-1, one can see that the percentage of intensifying cases in weak TCs (TD, TS and H1) is always higher than that of weakening cases.

Nevertheless more weakening cases were recorded than the intensifying cases for hurricanes from category 2 to category 5. Therefore, it is easier to identify conditions for TC weakening in stronger hurricanes

Staying with Table 5-1, one may expect that the percentage of intensifying cases should decrease with increasing initial intensity and that of weakening cases should increase. However this expectation does not fit the data shown in Table 5-1 , especially for the intensifying category 3 hurricanes and tropical depressions and the weakening category 2 hurricanes. For instance, the intensifying percentage for category 3 hurricanes (39.2%) is higher than those of category 2 hurricanes (28.7%) and even category 1 hurricanes (37.9%). The same happens for weakening category 2 hurricanes (48.9%), higher than those for category 3 (43.3%) and category 4 (46.2%). Those out-of-trend numbers indicate that forecasting intensity changes for categories 2 and 3 hurricanes may be more difficult because their behavior is more erratic.

5.2 Data preprocessing for the stratified TCs

Conceptually, the association rule algorithm was designed to deal with data sets of Boolean type attributes, such as “market basket” transaction data with items either present or absent. As of a scientific dataset with continuous numeric attributes such as the case here, all of the continuous attributes have to be broken down into disjoint predicates or conditions corresponding to a set of Boolean conditions.

There are various ways to decompose a numeric attribute into several range predicates. For example, Tan et al. (2001) used sample mean and standard deviation as the splitting

criteria to convert an attribute into three range conditions – "normal," "above normal" and "below normal" to apply association rules for finding abnormal ecological patterns. The main interest in this study is to distinguish intensifying TCs from weakening TCs. Therefore, the entire range of each attribute has to be split into a "Low" range and a "High" range. Thus the problem of finding the best split value of an attribute for the intensifying and weakening TC groups is a two-class classification problem using only one attribute. Based on Bayesian theory, the optimal split value for an attribute to distinguish 2 classes is where the histograms of two classes reach the same frequency. If the two classes have different number of observations, the optimal split value shifts from the class with larger population to the smaller class. Therefore, the estimated best split value should be weighted by the population sizes. In this stratified study, populations of the two groups are not significantly biased except for category 5 hurricanes (see Table 5-1). So the sample means of the 2 classes are used as the estimate of the split value for each parameter.

Although the 21 parameters listed in Table 3-1 are significantly correlated to the future intensity change based on the SHIPS model, it is suspected that this result is skewed by the dominate number of tropical storm cases (see Table 5-1). Therefore, we reexamined the parameters on their ability to distinguish TC intensifying and weakening. For each attribute, the two-side two-sample t test is used to test the null hypothesis that the means of the two TC intensity change stages are the same. If the p-value is less than 0.05, the null hypothesis is rejected and the attribute is said to have some ability to distinguish the two different stages. This test is chosen because most of the parameters are nearly normally distributed with equal

variances. In other circumstances other nonparametric tests such as the Wilcoxon-Mann-Whitney rank sum test may be more appropriate (Wilks 2005).

5.3 Attribute selection for various TC categories

The t test results demonstrate that a total of 18 parameters out of the 21 parameters can distinguish intensifying and weakening TCs for one or more of the TC categories as listed in Table 5-2. Among the selected 18 parameters, 17 of them are associated with tropical storms (TS), and fewer parameters are related with other categories. This confirms the suspicion that SHIPS's attribute selection is skewed by dominant tropical storm records. Please note that category 5 cases are excluded in the study of stratified TCs due to the small sample number.

Table 5-2 Selected SHIPS parameters for TC categories using t-test

TD (15)	TS (17)	H1 (13)	H2 (11)	H3 (12)	H4 (6)
V0					
IV12	IV12	IV12	IV12	IV12	IV12
U200	U200	U200	U200	U200	U200
	EPOS	EPOS	EPOS	EPOS	
RHLO	RHLO	RHLO			
RHHI	RHHI				
SHRD	SHRD	SHRD	SHRD	SHRD	
PSLV	PSLV				
Z850	Z850				
D200	D200				
	REFC	REFC		REFC	REFC
POT	POT	POT	POT	POT	
POT2	POT2	POT2	POT2	POT2	
VV	VV	VV	VV	VV	VV
SRLA	SRLA	SRLA	SRLA	SRLA	SRLA
SRV0	SRV0	SRV0	SRV0	SRV0	
ZONX	ZONX	ZONX	ZONX	ZONX	
	MERY	MERY	MERY	MERY	MERY

Note:
The numbers in the parentheses are the total numbers of parameters selected for each category

From Table 5-2 one can see that the parameters demonstrating the distinguishing ability in most of the 6 subsets are those related to the intensity change in the past 12 hours (IV12, VV), the vertical shear (U200, SHRD, SRLA, SRV0), the past storm motion (ZONX, MERY), the difference between the current intensity and the maximum possible intensity (POT, POT2), the instability of the upper atmosphere (EPOS), and the convergence of the

upper atmosphere (REFC). Table 5-2 also shows that the humidity of the atmosphere (RHLO, RHHI), the vorticity of the atmosphere (Z850, D200), and the pressure level of storm steering (PSLV) have the distinguishing ability mainly for tropical depressions and tropical storms. The current intensity (V0) plays a determinate role only for tropical depression, and this is not unexpected because the current intensity was used for the definition of categories and therefore, in each category, the variation of the initial intensity is limited.

5.4 Split values at various TC categories

The split value of each selected parameter is estimated by averaging the means of the intensifying and weakening cases in each subset and those split values are listed in Table 5-3.

Table 5-3 Split values of the t-test selected parameters and the binary directions of the sample means of the intensifying and weakening TC groups

(Units as shown in Table 3-1)

Parameter	TD	TS	H1	H2	H3	H4
V0	28.5 I<W					
PSLV	652 I<W	647 I<W				
Z850	-2.93 **I>W**	16.68 **I>W**				
D200	10.1 **I>W**	16.3 **I>W**				
REFC		1.96 **I>W**	4.23 I<W		3.15 I<W	2.23 I<W
EPOS		11.28 **I>W**	11.05 **I>W**	12.5 **I>W**	14.23 **I>W**	
RHLO	57% **I>W**	58% **I>W**	57% **I>W**			
RHHI	46% **I>W**	47% **I>W**				
IV12	-1.6 **I>W**	0.85 **I>W**	2.1 **I>W**	4.5 **I>W**	4.3 **I>W**	7.2 **I>W**
POT	108.3 **I>W**	86.5 **I>W**	60.0 **I>W**	47.1 **I>W**	42.3 **I>W**	
POT2	12045.2 **I>W**	7943.2 **I>W**	4052.7 **I>W**	2525.9 **I>W**	1969.1 **I>W**	
VV	-45.4 **I>W**	52.8 **I>W**	148 **I>W**	411 **I>W**	442.4 **I>W**	877.3 **I>W**
U200	8.66 I<W	10.5 I<W	10.96 I<W	8.14 I<W	4.23 I<W	1.65 I<W
SHRD	19.3 I<W	20.2 I<W	18.8 I<W	16 I<W	14 I<W	
SRV0	554 I<W	951 I<W	1338 I<W	1425 I<W	1454 I<W	
SRLA	7.66 I<W	8.64 I<W	9.18 I<W	7.39 I<W	6.02 I<W	4.6 I<W
ZONX	93.7 **I>W**	56.5 **I>W**	14.3 **I>W**	36.5 **I>W**	109.9 **I>W**	
MERY		89.3 **I>W**	150.4 I<W	144.3 I<W	143.9 I<W	97.3 I<W

The table also shows the "I-W" relation under each split value. "I" stands for the mean of the intensifying TC cases, and "W" stands for the mean of the weakening cases. The relation of "I>W" (in boldface) in the table indicates that the mean of the intensifying cases is significantly greater than that of the weakening cases for a particular parameter. In other words, the parameter is a favorable condition for TC intensification since a higher value is favored in intensifying TCs when a parameter has an "I>W" relation. Similarly the relation of "I<W" indicates that a lower value of the parameter is preferred in intensifying TCs thus it is an adverse condition for TC intensification. This expression gives us a quick overview of how the parameter will likely behave in the intensifying and weakening TCs. For example, the intensifying TDs and TSs tend to have a low steering pressure (PSLV<652mb for TDs and PSLV<647mb for TSs), or a strong vorticity in the atmosphere (Z850>-2.93 x $10^5 s^{-1}$ and D200>10.1 x $10^5 s^{-1}$ for TDs, and Z850>16.68 x $10^5 s^{-1}$ and D200>16.3 x $10^5 s^{-1}$ for TSs), or a humid atmosphere (RHLO>57% and RHHI>46% for TDs, and RHLO>58% and RHHI>47% for TSs), which is consistent with the general recognition on the necessary conditions for cyclonegenesis (Landsea 1997, http://www.faqs.org/faqs/meteorology/storms-faq/part1/).

As shown in Table 5-3 most of the parameters have a consistent "I-W" relation over various TC categories, indicating a consistent function of that parameter in TC development. However, the "I-W" relations of the convergence of the upper atmosphere (REFC) and the meridional motion (MERY) change their directions from tropical storms to hurricanes.

Specifically, Table 5-3 indicates that for TSs, most of the intensifying storms have a persistence of moving more than 89.3 km northwards along the meridional direction in the next 12 hours (equivalent to a 4.0 knots of the average meridional component of TC motion in

the past 12 hours), while for hurricanes, farther northwards movements involve more weakening conditions. This observation is due to the north-drifting of the Intertropical Convergence Zone (ITCZ) associated with the warm ocean water. Tropical storms usually occur south of the ITCZ and moving northwards means entering a more favorable environment. Northwards hurricanes however usually move away from the ITCZ and enter a more adverse environment. This result is consistent with the findings by Davis et al. (2008) in their studies of TC formation in the eastern Pacific. Certainly, detailed conclusion about the role of MERY should be drawn along with the TC locations to have a better evaluation of the ITCZ (Davis et al. 2008).

The behavior of the attribute REFC shown in Table 5-3 indicates that intensifying tropical storms and weakening hurricanes tend to have higher REFC. Similar negative correlation between REFC and intensity change was found by DeMaria et al. (1993), who investigated the role of upper-level eddy momentum fluxes (EFCs) on TC intensity changes based on the data for all named storms during the 3-year (1989-91) Atlantic hurricane seasons and found that the EFC is negatively correlated with TC intensity in single regression analysis. However, after controlling other environmental factors such as the vertical shear (SHRD), the potential to grow (POT), etc., in a multiple regression model, DeMaria et al. (1993) and DeMaria and Kaplan (1994b, 1999) also showed that the parameter REFC has a positive correlation with the intensity change. The contrary results was actually shown by Merrill (1988) who investigated the 1977-83 Atlantic hurricane cases and concluded that no clear correlation exists between upper-level eddy momentum fluxes and the Atlantic hurricane intensity change. The results shown in Table 5-3 help to explain the contraries. The role of

REFC in TC intensity changes depends on the initial intensity. For weak TCs such as tropical storms, high REFC results in intensification or positive correlations between REFC and intensity changes. For strong TCs at hurricane levels, the role of REFC is just opposite, a negative correlation. As pointed out in Table 5-1, tropical storm is the largest category in the TC stratum, and it is logical to suspect that the positive correlation between the upper-level eddy momentum fluxes and TC intensity changes found in other work is mainly caused by the skewed high number of TS cases.

It is interesting to note that the stronger the intensity scale of a hurricane is, the higher the split values tend to be as for the past intensity change (IV12, VV), the instability of the upper atmosphere (EPOS), and the zonal motion (ZONX), except for H3 with IV12. Considering that they have an "I>W" relation, that is, they are the favorable conditions for TC intensification, this trend indicates the existence of a more restrictive threshold in order for a stronger hurricane to grow persistently. On the contrary, for another favorable condition, the differences between current intensity and the maximum possible intensity (POT and POT2), the thresholds become more relaxed for them because those threshold values decrease with the hurricane strength. Similarly, the threshold values also decrease with the increase of the hurricane intensity for the adverse condition of vertical shear related parameters (SHRD, U200 and SRLA). This means that the thresholds of those unfavorable TC intensifying environmental parameters are lessened with a stronger hurricane.

CHAPTER 6 Roles of associated conditions in stratified tropical cyclones

6.1 General and concise rules

Association rule algorithms implemented by Borgelt (2009) were executed here to find condition combination for both intensifying TCs and weakening TCs with binarized parameter value ranges as the antecedents. Here, the antecedents are the selected parameters for each category associated with the high and low value ranges based on the threshold values given in Table 5-3. The format "parameter=H or L" is used for representing the high or low parameter value ranges in mined rules. For example, "U200=L" for the TS cases means that the 200-hPa zonal wind is less or equal to 10.5 knots. Initially the rule strength control parameters, that is, the minimal support of the antecedent and the minimal confidence of the rules, are set at 1/3, or 33%; and the minimal lift of the rule at 100%. After that the predefined confidence levels are adjusted to obtain manageable numbers of rules for discussion.

A typical rule mined from the tropical storm data subset is of the form, "INTENS ←U200=L, SHRD=L, SRLA=L (37.9, 57.4, 138.6)," which means that when three conditions (U200=L, SHRD=L, SRLA=L) work together, a TC will intensify with a 37.9% of support, a 57.4% of confidence, and a 138.6% of lift. Certainly, those jargons for computer science community should be translated into the language used in the TC research community. Since this rule is mined for the tropical storm category, this rule actually tells us the followings: for

the data records in tropical storm category, there are 37.9% cases satisfying the combination conditions, (U200=L [≤10.5 kt], SHRD=L [≤20.2 kt], SRLA=L [≤8.64 kt]). Among those cases, 57.4% of them underwent intensification, against the sample mean 41.4% for tropical storms. Actually, the definition of lift is equivalent to the ratio of confidence to the sample mean (57.4/41.4=138.6%). Therefore, a larger than 100% lift value guarantees that the percentage of intensifying cases mined by association rules is larger than the sample mean and reinforces the meaning of the rules found since they are true dependencies between the antecedents and the consequents.

One problem with plain association rule data mining technique is the existence of redundant rules which should be removed, and only the concise rules should be considered. A concise rule is a rule that cannot be derived from any subsets of the antecedents of other rules with a higher confidence. In other words, a concise rule contains the least constraints (antecedents). Any rule containing those constraints and any additional constraint without a higher confidence than the confidence of the concise rule is a redundant rule. For example, Table 6-1 lists all the eleven rules mined out by the association rule algorithm based on the predefined conditions for intensifying category 4 hurricanes, and three of them are redundant. Clearly, the rule defined by (VV=H, IV12=H) is of the same confidence with the first two rules defined by VV=H or IV12=H, respectively, and therefore, is a redundant rule. All other rules with two conditions (Rules 4-7) are concise rules because none of those conditions appear individually in rules with one condition only. The eighth rule, (U200=L, SRLA=L, MERY=L), contains a condition combination (SRLA=L, MERY=L) which appears in Rule 5. Since the confidence value in Rule 8 with an additional condition is not higher than the

confidence value of Rule 5, this rule is a redundant rule and should be removed. The same logic works for Rule 9 compared against Rule 6. The conditions for Rule 10, (U200=L, MERY=L, REFC=L), contains two sub-conditions, (U200=L, MERY=L) for Rule 4, and (MERY=L, REFC=L) for Rule 7. However, the confidence value for Rule 10 is higher than the corresponding values in both Rule 4 and Rule 7. As a result, Rule 10 is not a redundant rule. The case for Rule 11 is the same.

Table 6-1 Examples of concise rules and redundant rules from the mining result

Italicized rules are redundant. The columns represented by "s," "c," and "i" give the values of support, confidence, and lift in percentages respectively.

No	Conditions	s	C	i
1	IV12=H	37.6	37.1	157
2	VV=H	37.6	37.1	157
3	*IV12=H, VV=H*	*37.6*	*37.1*	*157*
4	U200=L, MERY=L	38.7	33.3	140.9
5	SRLA=L, MERY=L	37.6	37.1	157
6	SRLA=L, REFC=L	44.1	34.1	144.3
7	MERY=L, REFC=L	48.4	35.6	150.3
8	*U200=L, SRLA=L, MERY=L*	*35.5*	*36.4*	*153.7*
9	*U200=L, SRLA=L, REFC=L*	*38.7*	*33.3*	*140.9*
10	U200=L, MERY=L, REFC=L	35.5	36.4	153.7
11	SRLA=L, MERY=L, REFC=L	34.4	40.6	171.7

6.2 Intensifying/weakening rules

Table 6-2 lists the concise association rules which are the strongest in each TC category for the intensifying cases. To have a manageable number of rules for each category, while the same support level is kept at 33%, the confidence levels are increased from 33% to 35% for

tropical depressions (TD), 55% for tropical storms (TS), 60% for category 1 hurricanes (H1), 50% for category 2 hurricanes (H2), and 70% for category 3 hurricanes. The predefined 33% confidence level is used only for category 4 hurricanes (H4). The corresponding association rules for weakening TC cases are listed in Table 6-3, for which, the confidence levels are adjusted to 20% (TD), 40% (TS), 55% (H1), 70% (H2), and 66% (H3), respectively. Again the predefined 33% confidence level is kept for category 4 hurricanes as well.

Table 6-2 Strong concise intensifying rules mined for various TC categories.

The numbers in the three columns represent support (s), confidence (c), and lift (i) values in percentages, respectively.

Cat	Conditions	s	c	i
TD	D200=H, VV=H	35.9	36.6	142.8
	D200=H, IV12=H	35.9	36.6	142.8
	POT2=H, SRV0=L	33.7	35.2	137.4
	POT2=H, VV=H	36.2	37.2	145
	POT2=H, IV12=H	36.2	37.2	145
	RHHI=H, VV=H	38.1	38.7	150.8
	RHHI=H, IV12=H	38.1	38.7	150.8
	POT=H, VV=H	39.1	35.2	137.5
	POT=H, IV12=H	39.1	35.2	137.5
	POT2=H, SRV0=L, SHRD=L	33	35.9	140.1
TS	VV=H	45.4	68.8	166.4
	IV12=H	45.4	68.8	166.4
	U200=L, SHRD=L	40.7	55.6	134.3
	SHRD=L, SRLA=L	44.6	55.7	134.7
	POT=H, SRV0=L	34.4	57.1	138
	U200=L, SHRD=L, SRV0=L	37.9	55.8	134.8
	U200=L, SHRD=L, SRLA=L	37.9	57.4	138.6
	U200=L, SRV0=L, SRLA=L	37.6	56.3	136.1
	U200=L, SHRD=L, SRV0=L, SRLA=L	35.9	57.6	139.2
H1	VV=H	45.2	61.1	161.4
	IV12=H	45.2	61.1	161.4
	VV=H, SRLA=L	33.9	69.8	184.2
	IV12=H, SRLA=L	33.9	69.8	184.2
	ZONX=H, U200=L, SRV0=L	36.6	60	158.5
	POT=H, SRV0=L, SHRD=L	36.2	60.1	158.8
	ZONX=H, U200=L, SRV0=L, SHRD=L	35.6	60.6	160
	ZONX=H, U200=L, SRV0=L, SRLA=L	36	61	161.2
	ZONX=H, U200=L, SHRD=L, SRLA=L	36.4	60.3	159.4
	ZONX=H, SRV0=L, SHRD=L, SRLA=L	36	60.5	159.7
	POT=H, U200=L, SRV0=L, SRLA=L	33.5	60.6	160.1
	POT=H, U200=L, SHRD=L, SRLA=L	33.5	60	158.5
	POT=H, SRV0=L, SHRD=L, SRLA=L	36	60.5	159.7
	ZONX=H, U200=L, SRV0=L, SHRD=L, SRLA=L	34.9	61.7	162.9

Cat	Conditions	s	c	i
H2	VV=H	46	53.8	187
	IV12=H	46	53.8	187
	ZONX=H, U200=L	33.9	50.8	176.9
	SRV0=L, MERY=L	33.3	51.7	180
	SHRD=L, MERY=L	33.9	50.8	176.9
H3	MERY=L, U200=L	35.1	70.6	180.2
	VV=H, SRLA=L	37.1	72.2	184.4
	VV=H, U200=L	37.1	72.2	184.4
	IV12=H, SRLA=L	37.1	72.2	184.4
	IV12=H, U200=L	37.1	72.2	184.4
	VV=H, SRV0=L, SRLA=L	35.1	73.5	187.7
	VV=H, SRLA=L, SHRD=L	35.1	73.5	187.7
	IV12=H, SRV0=L, SRLA=L	35.1	73.5	187.7
	IV12=H, SRLA=L, SHRD=L	35.1	73.5	187.7
H4	IV12=H	37.6	37.1	157
	VV=H	37.6	37.1	157
	U200=L, MERY=L	38.7	33.3	140.9
	SRLA=L, MERY=L	37.6	37.1	157
	SRLA=L, REFC=L	44.1	34.1	144.3
	MERY=L, REFC=L	48.4	35.6	150.3
	U200=L, MERY=L, REFC=L	35.5	36.4	153.7
	SRLA=L, MERY=L, REFC=L	34.4	40.6	171.7

Table 6-3 Strong concise weakening rules mined for various TC categories

Cat	Conditions	s	c	i
TD	SHRD=H	35.6	22.5	171.4
	SRLA=H	35.6	21.6	164.5
	SRV0=H	37.8	22.9	174.1
	Z850=L	46.5	21.4	162.7
	SHRD=H, SRV0=H	33.7	23.8	181.2
	POT=L, V0=H	33	21.4	162.5
	POT2=L, V0=H	35.6	20.7	157.7
TS	SRV0=H	44.7	41	136.1
	PSLV=H	50.7	40.1	133.4
	IV12=L	54.6	46.4	154.3
	VV=L	54.6	46.4	154.3
	SRLA=H, SHRD=H	33.2	40.9	135.8
	SRV0=H, SHRD=H	40	41.7	138.5
H1	SRLA=H	39.7	55.3	157.2
	SRLA=H, SHRD=H	35.8	55.6	158.1
	SRLA=H, SRV0=H	35.4	56.2	159.9
	SRLA=H, U200=H	34.3	55.5	157.9
	SRLA=H, POT2=L	33.7	59.6	169.7
	SHRD=H, U200=H	34.3	55.5	157.9
	SRV0=H, U200=H	34.5	55.8	158.6
	U200=H, POT2=L	35.4	55.6	158.3
	POT=L, ZONX=L	33.1	55.1	156.7
	IV12=L, POT2=L	35.6	62.4	177.4
	VV=L, POT2=L	35.6	62.4	177.4
	SRLA=H, SHRD=H, SRV0=H	34.9	56.3	160.2
	SHRD=H, SRV0=H, U200=H	33.5	56.3	160

Cat	Conditions	s	c	i
H2	MERY=H, U200=H	33.9	71.2	145.7
	SHRD=H, U200=H	36.2	71.4	146.2
	SHRD=H, VV=L	33.9	72.9	149.2
	SHRD=H, IV12=L	33.9	72.9	149.2
	SRV0=H, VV=L	35.1	72.1	147.7
	SRV0=H, IV12=L	35.1	72.1	147.7
	U200=H, VV=L	35.1	70.5	144.3
	U200=H, IV12=L	35.1	70.5	144.3
	VV=L, POT2=L	34.5	73.3	150.1
	IV12=L, POT2=L	34.5	73.3	150.1
	SHRD=H, SRV0=H, U200=H	35.6	72.6	148.6
H3	SRV0=H, VV=L	34	69.7	161
	SRV0=H, IV12=L	34	69.7	161
	SRLA=H, VV=L	34	69.7	161
	SRLA=H, IV12=L	34	69.7	161
	SHRD=H, VV=L	34	69.7	161
	SHRD=H, IV12=L	34	69.7	161
	U200=H, VV=L	34	72.7	168
	U200=H, IV12=L	34	72.7	168
	U200=H, MERY=H	35.1	70.6	163
H4	REFC=H	33.3	74.2	160.5
	MERY=H	46.2	55.8	120.7
	SRLA=H	47.3	59.1	127.8
	U200=H	49.5	54.3	117.5
	IV12=L	62.4	50	108.1
	VV=L	62.4	50	108.1
	SRLA=H, IV12=L	33.3	61.3	132.6
	SRLA=H, VV=L	33.3	61.3	132.6
	U200=H, IV12=L	33.3	58.1	125.6
	U200=H, VV=L	33.3	58.1	125.6

The most striking observation from Table 6-2 is that the persistence of past intensity increase (IV12=H, or VV=H) appears frequently in the concise intensifying rules for all TC categories. In tropical storms and category 1, 2 and 4 hurricanes, the past intensity increase condition appears alone in the intensifying rules, and those rules are of the highest support values in all intensifying rules in the corresponding categories except for category 4 hurricanes. Moreover, the confidence levels of those rules are relatively high among all intensifying rules in the corresponding categories if not the highest. Those results indicate persistence of intensify increase in the past is the most prevalent and favorable factor for cyclone intensification, at least for TC of strength from tropical storms to category 2 hurricanes. Nevertheless, for tropical depressions the increasing past intensity condition should work together with another condition such as high 200 hPa divergence (D200=H), high 500-300 hPa relative humidity (RHHI=H), or large difference between the maximum possible intensity and the current intensity (POT=H or POT2=H) to establish a condition combination supporting TC intensification.

On the other hand, the strong vertical shear condition (U200=H, SHRD=H, SRV0=H, or SRLA=H) and the past intensity decrease condition (IV12=L, or VV=L) appear frequently in the most concise weakening rules, as shown in Table 6-3. For example, in category 4 hurricanes, the strong vertical shear condition and the past intensity decrease are present alone in the rules, indicating that either of them alone can produce the most prevalent situation to reduce the intensity of strong hurricanes. At the opposite end of the cyclone strength spectrum, the strong vertical shear dominates the weakening factors for tropical depressions and the past intensity decrease plays a significant rule in tropical storms. In other TC

categories including category 1, 2, and 3 hurricanes, both high shear and past intensity decrease appear in the weakening conditions but they must work together with other conditions to have the weakening effects.

The above observations show that these two factors – the persistence of intensity change of a tropical cyclone and the vertical shear in the atmosphere – are the most important factors to explain the future intensity change because they can produce truly correlated intensifying and weakening rules with large supports and large confidences.

One can also observe from Table 6-2 and Table 6-3 that the parameters describing the differences between the current intensity and the maximum possible intensity, or intensity potential, POT and POT2 play a role mainly for weak TCs. For instance, the high intensity potential, that is, POT=H or POT2=H, appears only in intensifying rules in TD, TS, and H1 subset. They do not show up in category 2, 3 and 4 hurricanes as part of intensifying factor. Similarly, the low intensity potential, that is, POT=L or POT2=L, make their appearance in TD, H1 and H2 categories. This tells us that for a major hurricane (category 3 or above), other factors become more important than the intensity potential parameters in terms of controlling the intensity. Another interesting result related to the intensity potential is that the nonlinear parameter, POT2, acts more importantly than the linear counterpart, POT, for hurricanes than for tropical depressions, especially for the weakening H1 cases.

The I-W relation changes with values of parameters REFC and MERY shown in Table 5-3 are not reflected in the mined rules because both of the parameters appear only in rules of strong hurricanes, category 2 or higher and the transition of trend takes place between tropical storms and category 1 hurricanes. As expected according to the threshold value trend, the high

convergence of the upper atmosphere (REFC=H) and the fast northward motion (MERY=H) do appear more often in weakening rules than the opposite conditions REFC=L and MERY=L appear in intensifying rules. Moreover, for category 4 hurricanes, both factors become dominate factors since they appear individually in weakening rules. Although the rule by MERY=H is not significantly stronger compared to the rules with a single low-shear or low-past intensity increase condition, the rule defined by REFC=H shows a significantly higher confidence than all weakening rules in that category. The rule demonstrates that the chance of the high convergence of the upper atmosphere is not high as we only have 1/3 of cases (33.3% of the support) in category 4. Nevertheless, if the REFC=H condition is satisfied, the hurricanes will weaken in the following 12 hours in nearly three out of four cases (74.2% of confidence).

In contrast to the role of northward motion (MERY), the role of the zonal motion is not very significant. This parameter plays roles only for category 1 and category 2 hurricanes in intensifying rules and for category 1 hurricanes in weakening rules. All those mined rules are associated with other conditions. The most noticeable results on lacking of involvements are the roles the 850-700 hPa relative humidity (RHLO) and the atmospheric stability represented by the "surface–200 hPa θe deviation of lifted parcel" (EPOS), which do not appear in any rules at all. One should note that we limit to rules with high support and confidence values here. The t-test results shown in Table 5-2 should lead to rules involving RHLO and EPOS with confidence values larger than sample means although the difference may not be significant.

There is a group of parameters working for tropical depression and tropical storms only including the pressure level of storm steering (PSLV), the 850 hPa relative vorticity (Z850), the 200 hPa divergence (D200), and the 500-300 hPa relative humidity (RHHI). Among them, PSLV and Z850 appear only in weakening rules and D200 and RHHI appear only in intensifying rules for tropical depressions. In the intensifying rules, both D200=H and RHHI=H must work together with high persistence (VV=H or IV12=H) to achieve high probability for TC intensification. However, the other two parameters showed weakening rules with single condition Z850=L or PSLV=H. The Z850=L rule, "*WEAKEN←Z850=L(46.5, 21.4, 162.7)*," has the highest support among all weakening rules for tropical depression. This means that we have more than 46% of chances to encounter TD with low 850 hPa relative vorticity, and those TD has a chance of 21.4% to weaken in the next 12 hours. Although the 21.4% is not a high probability, it is much higher than the sample mean 13.1%. In other words, if we predict that a TD will weaken based on its low Z850 value, our successful rate is 62.7% (from the 162.7% lift value) higher than the successful rate of the prediction purely based on the sample mean. The high pressure level of storm steering (PSLV=H) shows a similar prediction power for weakening tropical storms.

After discussing individual parameters, we can take a large picture to see the overall strengths of the mined rules in different categories of TCs. One can notice that the confidence values are the highest for intensifying category 3 hurricanes and for weakening category 2 hurricanes. However, they are not unexpected because of the high sample mean values for both cases. More meaningful comparisons on the prediction power should be derived based on the lift values which measure the ratio of the rule prediction power (confidence values) to

the prediction by unconditional chance only (sample means). Again, the lift values for intensifying H3 rules are uniformly higher than those in other categories, but in weakening cases, the high values are found in rules for tropical depressions and for category 3 hurricanes. Generally speaking, the lift values are higher for intensifying rules than those for weakening rules in the corresponding categories except for tropical depression. This may result from the data collection at the very beginning, that is, scientists always try to understand the physics of TC intensification thus they tend to include those parameters sensitive to TC intensifications.

6.3 Intramural binding

In the above discussions, two parameters IV12 and VV are considered as predictors of the same physical process, the past intensity changes; and four parameters, U200, SHRD, SRV0, and SRLA, are for the vertical shear. This way of grouping parameters is not only based on the definition of the corresponding parameters but also based on the mining results of the association among the parameters themselves, especially the implicit similarity among the frequent conditions.

To reveal the intramural binds among a set of frequent items, the action of searching for association hyper-edges was carried out using the same association rule algorithm. The concept of association hyper-edges was proposed by Han et al. (1997). A hyper-edge is a set of frequent items (the Boolean condition range in our study) from which all the association rules that can be generated are generated and the averaged confidence of all these rules must satisfy a minimum threshold.

The mining results have showed a perfect bind between IV12 and VV in all subsets. In other words, for all the TC categories from tropical depression to category 4 hurricanes, the IV12=H and VV=H appear together with 100% confidence. A detail check on their scatter plots has shown a narrow straight line for every subset. This result actually means that adding VV, a nonlinear derived variable, cannot result in any benefit to improve the ability of predicting the future TC intensity changes, and that the number of rules in Table 6-2 and Table 6-3 can be reduced further by removing all rules with parameter VV.

Similarly, strong intramural binds among U200, SHRD, SRV0 and SRLA were also identified by the association hyper-edges searching. Any two of them can form a hyperedge with a large support (around 40~50%) and with a large confidence (above 80%) in all 6 subsets. The scatter plots show trends of straight lines but not as perfect as the pair of IV12 and VV. Another strong binding revealed by the hyper-edge relationships is POT and POT2 pairs, as expected. The two strong but not perfect bindings among the vertical shear parameters and between the two intensity potential parameters demonstrate that although the gain is not strong by introducing new derived parameters but they did help us to increase our chances of improved prediction.

CHAPTER 7 Mining associated condition of Rapid Intensification of Tropical Cyclones

In addition to the rules for intensifying and weakening TCs, the association rule data mining technique is also used to investigate the condition combination for rapid intensification (RI) of TCs (Yang et al. 2007, 2008).

7.1 Introduction of Rapid Intensification of Tropical Cyclones

Forecasting of tropical cyclones (TC) intensity changes, in particular rapid intensification (RI) is a challenging problem. As Kaplan and DeMaria (2003) defined, a tropical cyclone undergoes rapid intensification (RI) if its intensity (defined by the maximum wind) has increased at least 15.4 m/s (30 knots) over a 24-hour period. The favorable factors on intensification of tropical cyclones have been broadly studies. Those factors include warm ocean eddies (Shay et al. 2000; Hong et al. 2000), the contraction of an outer eyewall (Willoughby et al. 1982; Willoughby and Black, 1996; Lee and Bell, 2007), an environment with low vertical shear (Gray, 1968; Merrill, 1988; DeMaria and Kaplan, 1994; DeMaria, 1996; Frank and Ritchie, 1999, 2001), interactions between the upper-level trough and a tropical cyclone (Molinari and Vollaro, 1989, 1990; DeMaria et al., 1993), and even cloud microphysics (Wang, 2002) and isotopic concentrations (Gedzelman et al., 2003).

Most of the previous studies were largely focusing on only one of three categories of factors, ocean characters, inner-core processes, and environmental interactions, and it is well known that intensity changes depends on a combination of those factors (Zhu et al., 2004). Holliday and Thompsom (1979) examined the rapidly intensifying northwest Pacific typhoons and observed that a sufficiently deep layer of warm water, the development at night time, and a smaller size of eye were necessary for those RI typhoons. DeMaria and Kaplan (1994) studied Atlantic TCs and observed that the TCs with a smaller size, with a greater potential to reach their maximum potential intensity, with a faster intensification history, and in an environment with low vertical shear and weak upper-level forcing could have the largest 48-hour intensification rates. In a study of the rapid intensification of Hurricane Opal (1985), Bosart et al. (2000) concluded that its RI was resulted from a combination of several factors: enhanced divergence, low vertical shear and the enhanced heat and moisture from a warm Gulf eddy. More recently, Kaplan and DeMaria (2003) examined the large-scale characteristics of rapidly intensifying Atlantic tropical cyclones from 1989-2000 using the NHC HURDAT file and the SHIPS database. Their results confirmed the aforementioned studies. Furthermore a scheme to estimate the probability of RI was developed in their study by combining the thresholds of the five persistence and synoptic predictors: the persistence of intensity change, the vertical shear, the sea surface temperature, the potential to reach the maximum potential intensity, and the moisture content in the lower atmosphere.

The goal of this study is to introduce the technique of association rules from the data mining research area as an “unsupervised,” “automatic” data exploration method to discover “multiple-to-one” associations among a large number of environmental characteristics that are

responsible for tropical cyclones which will be rapidly intensifying. The results from this technique can then be used to shed lights on the hypothesis generation regarding the underlying physical mechanisms, which can be used as the guidance in the traditional statistical analysis.

To satisfy the goal, the SHIPS 2003 dataset which is introduced in Chapter 3 and used in Chapter 5 and 6 are reused here to examine the ability of the association rule algorithm to discover the associated environmental conditions in rapid tropical cyclone development. The dataset construction, the RI definition, RI thresholds and RI probability definitions are followed those in Kaplan and DeMaria (2003, hereafter KD03) as much as possible. The purpose is to create a similar context as KD03 so that the results from the association rules can be easily mapped to and verified by the traditional statistical terms and methods.

7.2 Data set used for RI studies

This RI study follows the work of Kaplan and DeMaria (2003) which used the NHC HURDAT file and the SHIPS database from 1989-2000, examined the large-scale characteristics of rapidly intensifying Atlantic TCs, estimated the RI probability (RIP), and discussed the dependence of RI probability on a combination of factors. The same two data sets, but with the SHIPS 2003 database covering the 1982-2003 time period, are used here. They are merged based on the methodology identical to that described in KD03 except for the consideration of the systems remained both over water and tropical during the period from t – 12hr to t + 24hr, and for the non-developing tropical depressions. In total, there are 34 values (variables) for each case, and each variable was evaluated at the beginning (t=0h) of each 24-

h period. The 0000 and 1200 UTC synoptic predictor values in the SHIPS database were averaged to estimate the magnitude of the corresponding values at 0600 and 1800 UTC. Since the SHIPS data are from 1982 to 2003, longer than the 1989-2000 period studied by KD03, the data is divided into 3 subsets: 1982-88, 1989-2000, and 2001-03 in addition to the whole 1982-2003 period for analysis. It should be pointed out that the 2003 SHIPS database is of a 12-h interval although the more recent SHIPS data are at a 6-h interval (DeMaria et al. 2005). As a result, the SHIPS data are interpolated into 6-h interval for the RI study.

Table 7-1 Total case numbers, RI case numbers, and RI probabilities (RIP) based on the sample means of RI cases for various time periods.

Time Coverage	Total	RI	RIP
1982-1988	1170	60	5.1%
1989-2000	3306	169	5.1%
2001-2003	1029	36	3.5%
1982-2003	5505	265	4.8%

Abided by the RI definition proposed in KD03, a RI case is identified if the intensity of a TC (defined by the maximum wind) has increased at least 30 knots (15.4 m/s) over a 24-hour period. In the merged data set from 1982-2003, there are 7497 records with 1697 RI cases. However, to mine/analyze the data further, record with missing values were removed from the data set, which resulted in total 5505 valid records and only 265 RI cases. Similar process is applied to subsets from 1982-1988, from 1989-2000 and from 2001-2003. The numbers of cases after such division are listed in Table 7-1. For the 1989 to 2000 period, the

reconstructed dataset contains a total of 3306 cases, which were from 135 distinct Atlantic TCs (0 tropical depressions, 54 tropical storms, and 81 hurricanes). Unlike the KD03 dataset, the information of 30 non-developing tropical depressions is not included in this study because we are interested in tropical systems which had developed into hurricanes.

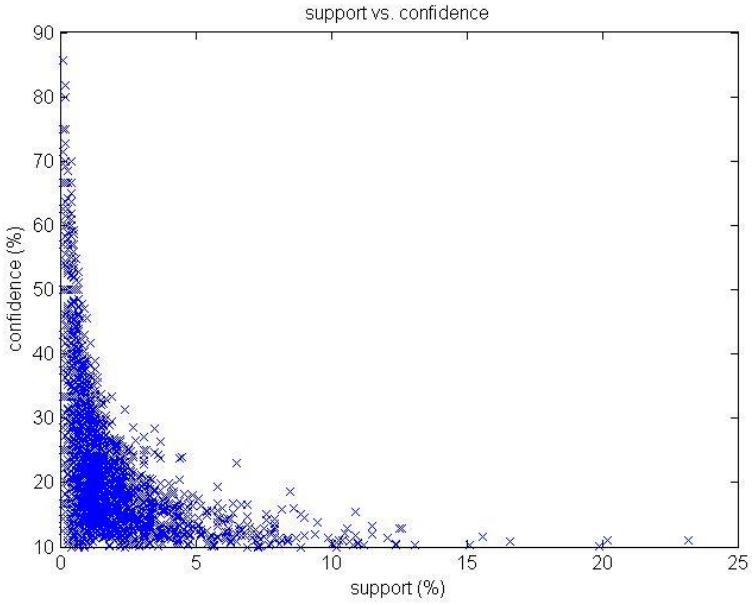
7.3 Data pre-processing for RI studies

As in KD03, eleven independent parameters out of a total of 34 were selected including eight variables used in the stratified TC studies and three new variables. The three variables are the longitude (LON) and latitude (LAT) values of the TC centers and the sea surface temperature (SST) (see the parameters with asterisks in Table 3-1). As in the stratified TC mining procedure, the continuous values of the 11 parameters are converted into disjoint binary conditions, "High" or "Low" value ranges. The threshold values separating "High" and "Low" ranges are initially chosen to be the same as those provided in KD03, the mean values of the RI samples.

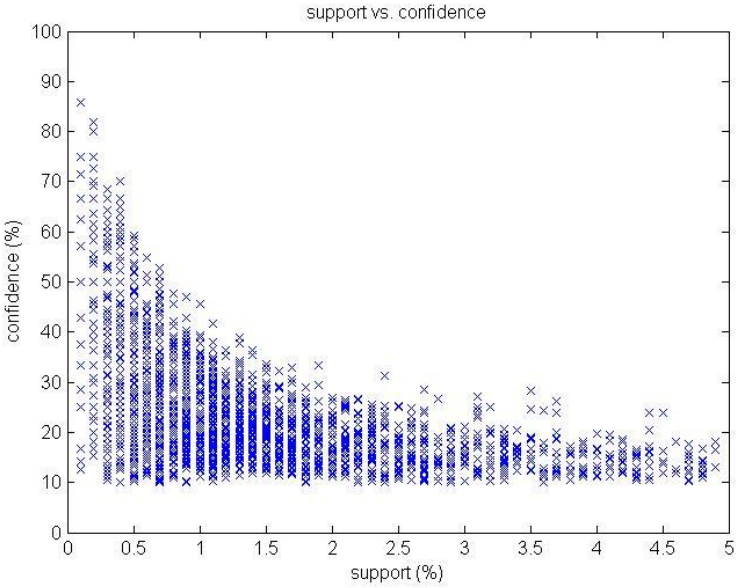
7.4 Mined higher RI probabilities

One challenge of mining RI favorable conditions is the low support values because of the low percentage of RI cases. Figure 7-1 (a) displayed the scatter plot of confidence and support of the mined rules which favor to the RI processes. It is clear that most of those rules are of relatively low supports and low confidences as shown by the dense data cloud in the left-bottom corner. The most interesting part of the figure is at the cases with high confidence,

which are the cases with high RI probabilities. Therefore, an enlarged scatter plot with support up to 5% is shown in Figure 7-1 (b) for the details in the high confidence part. There are quite a few cases with RI probabilities higher than 70% with relatively lower support, either 0.1% or 0.2%. The support level is rounded off to one digit after the decimal point in the software implementation. Due to the large sample, 5505 in this case, each 0.1% means about 5 cases. Considering the sample mean of around 5% of RI probability, one should notice that collectively, the data mining results can catch a much higher RIP with a relatively large case numbers if the high RIP cases with low supports are not heavily overlapped. This result is useful for guiding the TC intensity forecasting in addition to understand the RI processes by studying cases with high confidence values.



(a)



(b)

Figure 7-1 Scatter plots of confidence and support of the mined rules favoring the RI processes.

When observing that the probability of RI for any individual predictor was not particularly high, Kaplan and DeMaria (KD03) tried to find a set of predictors to provide improved probability of RI estimates. They found that employing the set of seven predictors that were significant at the 99.9% level yielded the highest RI probabilities. They then further narrowed down the list to five predictors: PD12, SHRD, SST, POT, and RHLO to construct a composite RI probability and removed LAT and U200 because they are highly correlated with the five remaining factors. They found that "a total of five RI thresholds were satisfied for <2% of the total sample …, a disproportionately large percentage (~10%) of the sample total number of RI cases were observed when all five of the RI thresholds were satisfied …, the probability of RI was 41% when all five of the RI predictors were satisfied."

The above cited result can be easily translated into an association rule as "RI <= SHRD=L, PD12=H, SST=H, POT=H, RHLO=H" with a support of the antecedent 1.5% (J. Kaplan, personal communication, 2005) and a confidence of the rule as 41%. Indeed, the exact associate rule with a support of the antecedent = 0.7%, a confidence = 43.5% and a lift = 850.5 % is found by the association rule technique. The lower support is likely due to the larger portion of non-RI samples in our dataset. However the higher confidence is likely due to the unenclosed non-developing depressions. We suspect that the non-developing depressions (1989-2000) may contain a higher proportion of cases satisfying the five conditions at the same time without undergoing RI, which resulted in the confidence of 43.5% in our study higher than the reported 41% from KD03.

Using the association rule mining technique, the process of finding a set of predictors which have improved RI probabilities can be easily accomplished by means of pruning the association rules. The pruning process would first identify redundant rules, those with confidence values being equal to or smaller than that of a more general rule. A redundant rule cannot provide improved RI probabilities and will be removed, while the ones with improved RI probabilities over other more general rules will survive the pruning process.

During the pruning process, the original rule with the five KD03 constraints satisfied, "RI <= SHRD=L, PD12=H, SST=H, POT=H, RHLO=H (supp=0.7%, conf=43.5%, lift=850.5%)," is removed as a redundant rule from the rule set. A more interesting rule with three constraints satisfied, "RI <- SHRD=L, PD12=H, RHLO=H (supp=1.3%, conf=47.6%, lift=931.5%)," remains. Similar conclusions are found after examining the datasets of 1982-1988, 1982-2003 and 2001-2003. This means that this rule with all 5 constraints satisfied in KD03 is not the best and a rule with fewer constraints fits more RI cases. Although the remaining rule has a shorter list of antecedents, it has a higher confidence than the original rule, which indicates that the 3 constraints, SHRD=L, PD12=H, and RHLO=H, can explain more general RI cases than all the 5 constraints. Therefore, increasing the number of constraints will not improve the accuracy of the determination of RI cases. Meanwhile, without a proper guidance, it is difficult to select RI constraints in the RI probability estimate to achieve the optimal results. The association rule mining with pruning process provides a systematic implementation to reach this goal.

To compare the results with those in KD03, the composite RI probability defined in KD03 is computed with statistical analysis. The results for time periods of 1982-2003, 1982-1988,

1989-2000, and 2001-2003 are displayed in Figure 7-2. The plots for the whole dataset from 1982 to 2003 and from 1989 to 2000 are similar to result of KD03. With the number of satisfied constraints increases, the RI probability increases. However, there are no RI cases which satisfied all 5 constraints at the same time in the datasets for the 1982-1988 and the 2001-2003 periods, thus the probabilities decrease to zeros. The RI probabilities for any 4 satisfied constraints also decrease in these two datasets. This observation indicates that not all 5 conditions are required for a TC to undergo RI. Moreover, the RI probability for any 3 out of 5 satisfied constraints seems more stable than that for more satisfied constraints for the 1982-1988 and 2001-2003 periods.

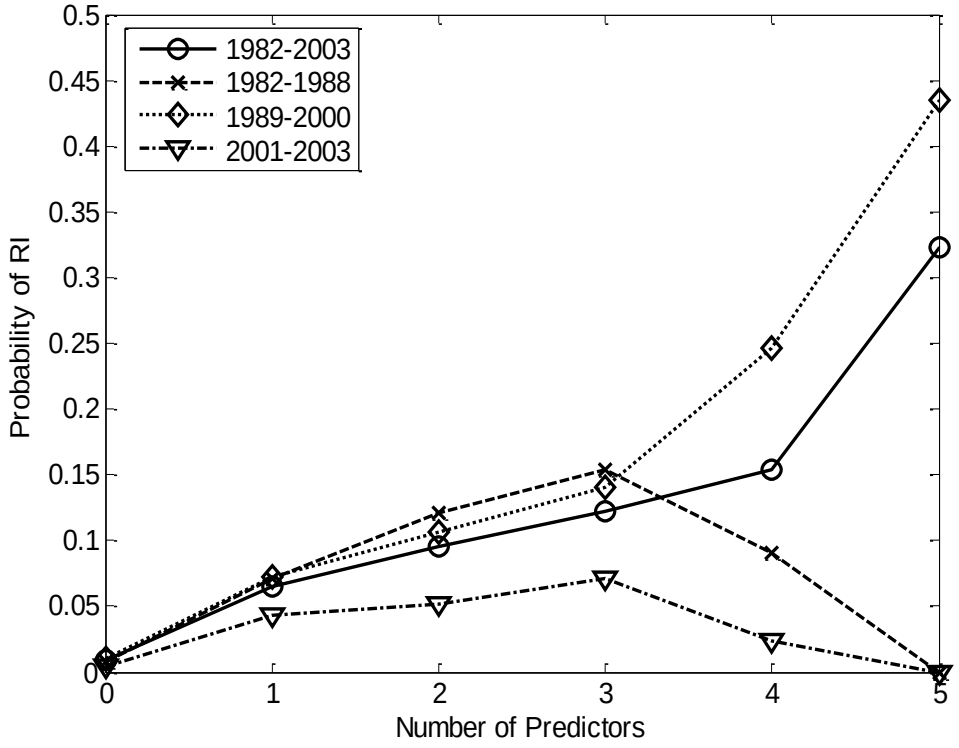


Figure 7-2 Composite RI probabilities with number of predictor constrains satisfied based on KD03 five RI thresholds (SHRD=L, PD12=H, SST=H, POT=H, RHLO=H*)* for different time periods.

The zero probability of RI with all five constraints satisfied is a striking result. More detailed investigation indicates that not only no case satisfying the five constrains at the same time for RI cases, but the number of records with all five identified conditions for both RI and non-RI cases are much lower for the 1982-1988 and 2001-2003 time periods. In the total of 22 years from 1982 to 2003, there are total 31 cases with all five RI favoring conditions. Twenty-three of them fall in the twelve year period, 1989-2000; six cases are in the last three years, 2001-2003. For the remaining sub-period, seven years from 1982 to 1988, only two cases satisfying all five constraints are found. This striking result demonstrates that long-term environmental conditions on interannual and decadal scales may affect the TC intensification as well as RI process.

To emphasize the importance of the narrowed down predictors, a similar figure with three satisfied constraints, "SHRD=L, PD12=H, and RHLO=H," mined via association rule technique is displayed in Figure 7-3. Comparing Figure 7-3 with Figure 7-2, one will find that there is no significant difference between the RI probabilities with at least one constraint satisfied for the cases of the five identified constraints and of the three reduced constraints. However, when at least two constraints are satisfied, the probabilities with three identified constraints are significantly higher than the corresponding probabilities in five constraint cases. The most significant result of this work is that when all three identified constraints are satisfied, the RI probabilities are higher than the probabilities with all five satisfied constraints identified in KD03. This result is significant not only to rapid intensification of hurricanes but also to the data mining procedures in identifying meaningful science results.

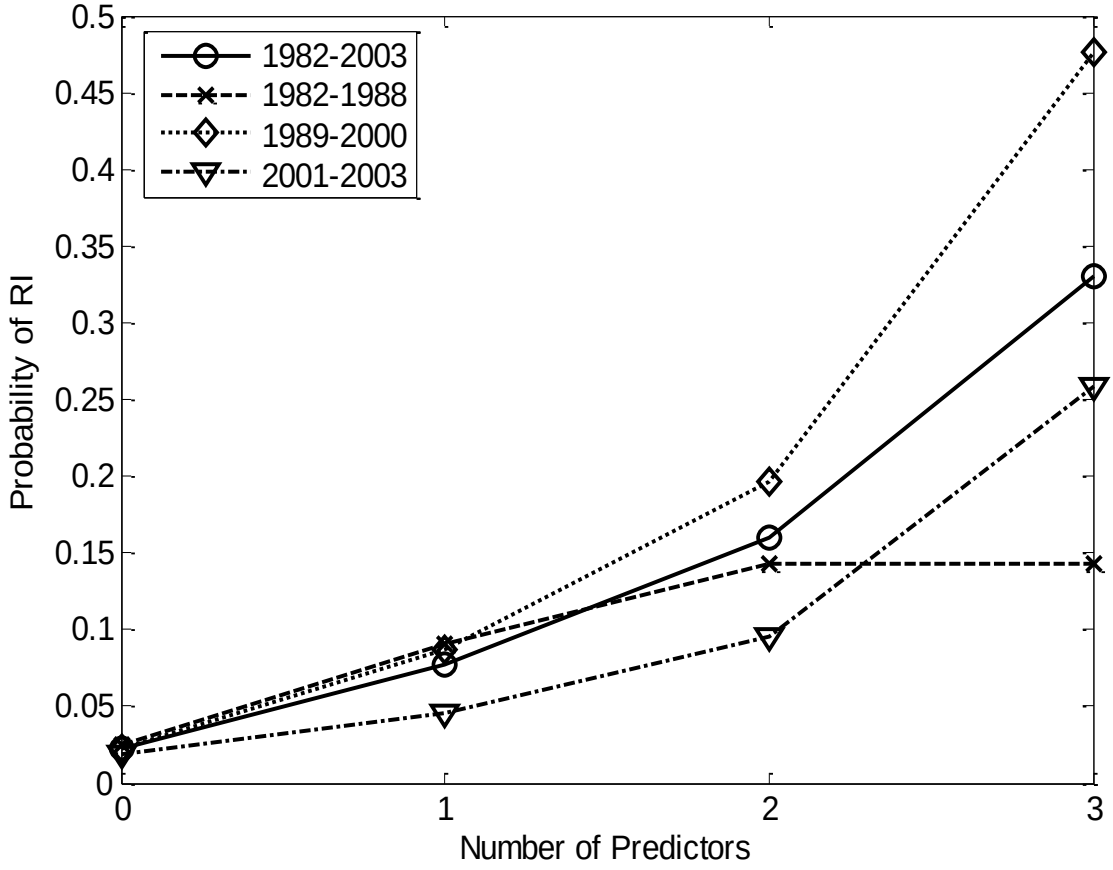


Figure 7-3 Composite RI probabilities with number of predictor constrains satisfied based on the three mined RI thresholds (SHRD=L, PD12=H, RHLO=H).

It is worth noting that the new reduced predictor list does not include "SST=H," high sea surface temperature, which is commonly considered as the main factor influencing the TC intensity and its change. Nevertheless, it was noticed that although SST is an important factor in hurricane intensity, it is not a controlling factor (Evans, 1993; Paterson et al., 2005). Moreover, Emanuel et al. (2004) explored the sensitivities of tropical cyclone intensity changes to initial conditions and environmental factors by using a simple coupled model with

wind shear parameterization. Their results demonstrated that the intensity changes are very sensitive to wind shear and humidity in middle troposphere air, which is well consistent with the short list of RI favoring factors, "SHRD=L, RHLO=H," i.e., low vertical wind shear and high relative humidity in the 850-700hPa layer. Recently, Wu (2007) defined hurricane mean peak intensity and examined the long-term relationship between Atlantic hurricane peak intensity and Saharan air layer (SAL) represented by 850hPa relative humidity and vertical wind shear between 850hPa and 600hPa. Wu et al. (2007) also defined a SAL effect index (SEI) by a linear combination of relative humidity (positive) and wind shear (negative) and obtained a higher correlation between the peak intensity and the new parameter than those with the two parameters separately. The most significant conclusion is that SEI "can much better account for the trend in the mean peak intensity … than SST or AMO index." This result is also highly consistent with our mining results of the improved RI probability by the combined contribution from the conditions "SHRD=L" and "RHLO=H."

7.5 Mining Optimal RIP

One step further, one can use the association rule data mining technique to search for the "optimal" conditions which result in the highest RIP for a given number of conditions among the selected set (Yang et al. 2008). Figure 7-4 shows the changes in highest RI probabilities with the number of the thresholds in the 22 predictor pool for different time periods. All curves demonstrate the same trend. That is, the highest RIPs increase with the numbers of predictors initially, reach the peak values when the number of predictors approaches five to seven (N=5-7), and then decrease with further increases of the numbers of predictors. These

results demonstrate that the multiple factors together are responsible for the RI process of TCs. However, the number of factors will saturate at certain numbers. After that, the impacts of individual factors may cancel each other out, or may be replaced by other factors.

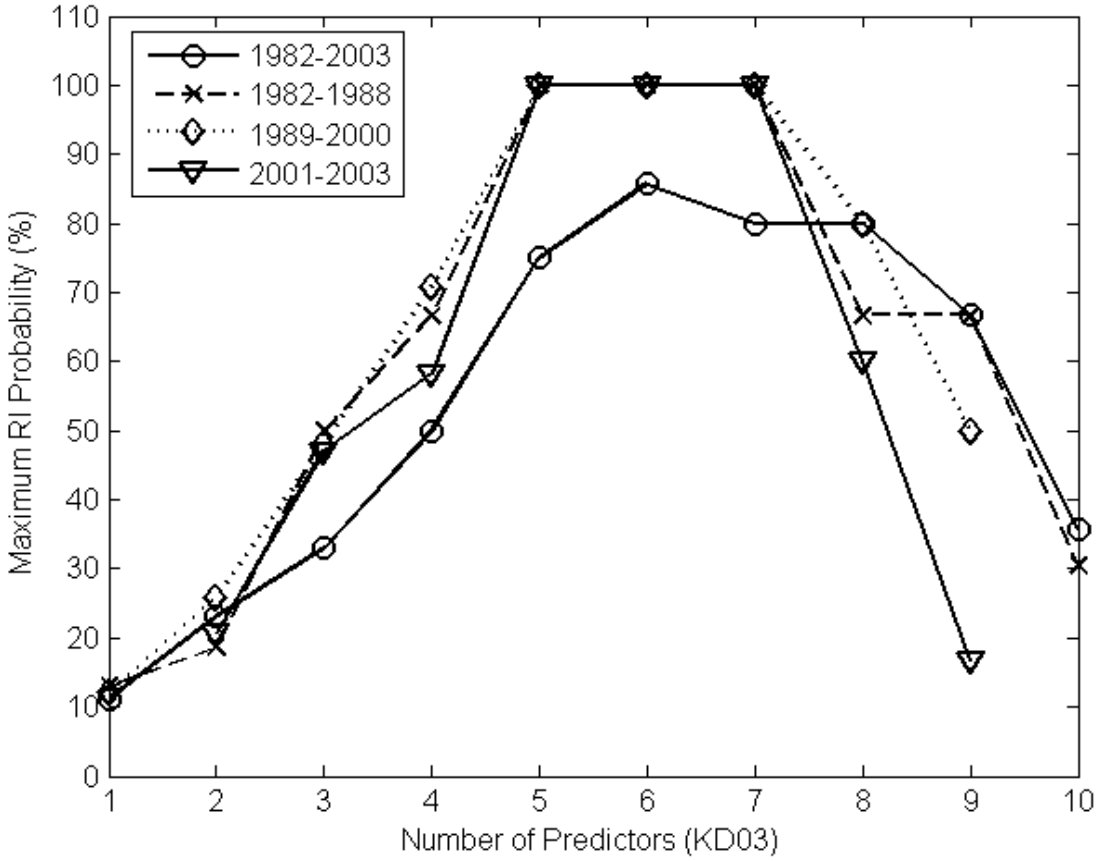


Figure 7-4 The highest RI probabilities for different number of thresholds and multiple time periods

The most striking result in Figure 7-4 is that the data mining algorithms identify certain cases with a 100% RI probability. These "perfect" results take place only for the three sub-periods but not for the whole period. This plausible result comes from the fact that the

detailed “optimal” conditions for each subset period are different from each other, and as a result, the conditions and the RI probability for the whole time period are also different from those in individual sub-periods.

Table 7-2 lists the detailed conditions for the N=6 cases for various time periods. By carefully checking, one can see that no two groups give the same conditions although the RIP (confidence) is 100% for all three short periods.

Table 7-2 Optimal conditions when N=6 for different time periods.

The columns represented by “s,” “c,” and “i” give the values of support, confidence, and lift in percentages, respectively.

Periods	Detailed Conditions for N=6	s	c	i
1982-1988	IV12=L,LAT=L,ZONX=L,POT=L,SHRD=L,RHLO=H	0.2	100	1950
1989-2000	POT=L,SHRD=L,IV12=H,ZONX=H,RHLO=H,PSLV=H	0.2	100	1956.2
2001-2003	IV12=L,LAT=L,POT=L,U200=L,SST=H,REFC=H	0.2	100	2858.3
1982-2003	LON=L,REFC=L,IV12=H,LAT=H,POT=H,PSLV=H	0.1	85.7	1780.6

Table 7-3 lists the support, confidence and lift of every optimal condition combination in other sub periods. Although these “optimal” results take place only in one of the three sub-periods but not for the whole period, they do appear as a good rule in most of other sub periods, although with a lower support and lower confidence. At some cases, the exact rules with all conditions satisfied were not found, rules will fewer conditions satisfied were found instead.

Table 7-3 Support, confidence, and lift of the optimal rules in other sub periods

	82-88 data	89-00 data	01-03 data	82-03 data
82-88 rule	0.2,100.0,1950.0	X (only 5 cond.)	0.1,25,714.6	0.1,42.9,890.3
89-00 rule	0.1,33.3,650.0	0.2,100.0,1956.2	X (only 5 cond.)	0.1,57.1,1187.1
01-03 rule	0.1,100.0,1950.0	X (only 4 cond.)	0.2,100.0,2858.3	0.1,25.0,519.3
82-03 rule	0.1,100.0,1950.0	0.2,83.3,1630.2	X (only 5 cond.)	0.1,85.7,1780.6

All of the 64 possible combinations from the 6 parameters selected in the optimal condition found in 1982-1988 period are listed in Table 7-4, along with the number of RI and non RI cases. The RI and non RI cases in other sub periods and the entire period are also listed. "T" column stands for the total cases in the sub period which have all the 6 conditions satisfied. "R" column stands for the total RI cases which have all of the 6 conditions satisfied. "N" column stands for the total non RI cases which have all of the 6 conditions satisfied. And the "E" column stands for the expected RI cases based on the sample mean (refer to the RIP values inTable 7-1), i.e., E = sample mean*T.

Several combinations are manually selected and highlighted in either green or yellow in Table 7-4. A combination highlighted in green is considered as an interesting combination to have some ability to indicate RI because the actual RI cases associated with that combination are much greater than the expected RI cases. Similarly, a combination highlighted in yellow is considered as an interesting combination to indicate non RI because the actual RI cases associated with that combination are far less than the expected RI cases.

Table 7-4 RI and non-RI cases of all combinations of the 6 parameters selected in the 1982-1988 optimal condition

		82-88				89-00				01-03				82-03			
	Condition Combination	T	N	R	E	T	N	R	E	T	N	R	E	T	N	R	E
1	IV12=L,LAT=L,ZONX=L,POT=L,RHLO=L,SHRD=L	9	9	0	0.5	49	43	6	2.5	0	0	0	0.0	58	52	6	2.8
2	IV12=L,LAT=L,ZONX=L,POT=L,RHLO=L,SHRD=H	33	31	2	1.7	178	174	4	9.1	11	8	3	0.4	222	213	9	10.7
3	***IV12=L,LAT=L,ZONX=L,POT=L,RHLO=H,SHRD=L***	***2***	***0***	***2***	***0.1***	***1***	***1***	***0***	***0.1***	***4***	***3***	***1***	***0.1***	***7***	***4***	***3***	***0.3***
4	IV12=L,LAT=L,ZONX=L,POT=L,RHLO=H,SHRD=H	3	2	1	0.2	3	3	0	0.2	12	12	0	0.4	18	17	1	0.9
5	IV12=L,LAT=L,ZONX=L,POT=H,RHLO=L,SHRD=L	16	16	0	0.8	83	79	4	4.2	5	5	0	0.2	104	100	4	5.0
6	IV12=L,LAT=L,ZONX=L,POT=H,RHLO=L,SHRD=H	112	110	2	5.7	203	194	9	10.4	21	21	0	0.7	336	325	11	16.2
7	IV12=L,LAT=L,ZONX=L,POT=H,RHLO=H,SHRD=L	14	13	1	0.7	15	13	2	0.8	19	19	0	0.7	48	45	3	2.3
8	IV12=L,LAT=L,ZONX=L,POT=H,RHLO=H,SHRD=H	7	7	0	0.4	24	23	1	1.2	65	64	1	2.3	96	94	2	4.6
9	IV12=L,LAT=L,ZONX=H,POT=L,RHLO=L,SHRD=L	0	0	0	0.0	2	2	0	0.1	0	0	0	0.0	2	2	0	0.1
10	IV12=L,LAT=L,ZONX=H,POT=L,RHLO=L,SHRD=H	6	6	0	0.3	57	55	2	2.9	0	0	0	0.0	63	61	2	3.0
11	IV12=L,LAT=L,ZONX=H,POT=L,RHLO=H,SHRD=L	4	1	3	0.2	0	0	0	0.0	0	0	0	0.0	4	1	3	0.2
12	IV12=L,LAT=L,ZONX=H,POT=L,RHLO=H,SHRD=H	1	1	0	0.1	19	19	0	1.0	10	10	0	0.3	30	30	0	1.4
13	IV12=L,LAT=L,ZONX=H,POT=H,RHLO=L,SHRD=L	0	0	0	0.0	18	18	0	0.9	0	0	0	0.0	18	18	0	0.9
14	IV12=L,LAT=L,ZONX=H,POT=H,RHLO=L,SHRD=H	21	21	0	1.1	72	71	1	3.7	6	6	0	0.2	99	98	1	4.8
15	IV12=L,LAT=L,ZONX=H,POT=H,RHLO=H,SHRD=L	4	4	0	0.2	25	23	2	1.3	10	10	0	0.3	39	37	2	1.9
16	IV12=L,LAT=L,ZONX=H,POT=H,RHLO=H,SHRD=H	3	3	0	0.2	55	49	6	2.8	47	45	2	1.6	105	97	8	5.1
17	IV12=L,LAT=H,ZONX=L,POT=L,RHLO=L,SHRD=L	7	7	0	0.4	69	69	0	3.5	23	23	0	0.8	99	99	0	4.8
18	IV12=L,LAT=H,ZONX=L,POT=L,RHLO=L,SHRD=H	196	195	1	10.1	511	509	2	26.1	116	116	0	4.1	823	820	3	39.6
19	IV12=L,LAT=H,ZONX=L,POT=L,RHLO=H,SHRD=L	7	7	0	0.4	2	2	0	0.1	4	3	1	0.1	13	12	1	0.6
20	IV12=L,LAT=H,ZONX=L,POT=L,RHLO=H,SHRD=H	18	17	1	0.9	18	18	0	0.9	29	29	0	1.0	65	64	1	3.1
21	IV12=L,LAT=H,ZONX=L,POT=H,RHLO=L,SHRD=L	14	11	3	0.7	30	28	2	1.5	6	6	0	0.2	50	45	5	2.4
22	IV12=L,LAT=H,ZONX=L,POT=H,RHLO=L,SHRD=H	79	73	6	4.1	116	107	9	5.9	39	38	1	1.4	234	218	16	11.3
23	IV12=L,LAT=H,ZONX=L,POT=H,RHLO=H,SHRD=L	8	6	2	0.4	4	3	1	0.2	2	2	0	0.1	14	11	3	0.7
24	IV12=L,LAT=H,ZONX=L,POT=H,RHLO=H,SHRD=H	5	5	0	0.3	18	15	3	0.9	22	22	0	0.8	45	42	3	2.2
25	IV12=L,LAT=H,ZONX=H,POT=L,RHLO=L,SHRD=L	36	35	1	1.8	107	105	2	5.5	21	21	0	0.7	164	161	3	7.9
26	IV12=L,LAT=H,ZONX=H,POT=L,RHLO=L,SHRD=H	145	144	1	7.4	536	527	9	27.4	118	117	1	4.1	799	788	11	38.5
27	IV12=L,LAT=H,ZONX=H,POT=L,RHLO=H,SHRD=L	0	0	0	0.0	2	2	0	0.1	8	8	0	0.3	10	10	0	0.5
28	IV12=L,LAT=H,ZONX=H,POT=L,RHLO=H,SHRD=H	15	15	0	0.8	13	13	0	0.7	17	15	2	0.6	45	43	2	2.2
29	IV12=L,LAT=H,ZONX=H,POT=H,RHLO=L,SHRD=L	16	16	0	0.8	64	59	5	3.3	16	15	1	0.6	96	90	6	4.6
30	IV12=L,LAT=H,ZONX=H,POT=H,RHLO=L,SHRD=H	101	96	5	5.2	161	156	5	8.2	102	99	3	3.6	364	351	13	17.5
31	IV12=L,LAT=H,ZONX=H,POT=H,RHLO=H,SHRD=L	2	2	0	0.1	12	12	0	0.6	10	10	0	0.3	24	24	0	1.2
32	IV12=L,LAT=H,ZONX=H,POT=H,RHLO=H,SHRD=H	9	9	0	0.5	49	48	1	2.5	77	77	0	2.7	135	134	1	6.5
33	IV12=H,LAT=L,ZONX=L,POT=L,RHLO=L,SHRD=L	10	8	2	0.5	76	57	19	3.9	0	0	0	0.0	86	65	21	4.1
34	IV12=H,LAT=L,ZONX=L,POT=L,RHLO=L,SHRD=H	9	9	0	0.5	106	102	4	5.4	6	6	0	0.2	121	117	4	5.8
35	IV12=H,LAT=L,ZONX=L,POT=L,RHLO=H,SHRD=L	11	9	2	0.6	7	4	3	0.4	12	5	7	0.4	30	18	12	1.4
36	IV12=H,LAT=L,ZONX=L,POT=L,RHLO=H,SHRD=H	8	6	2	0.4	11	11	0	0.6	10	10	0	0.3	29	27	2	1.4
37	IV12=H,LAT=L,ZONX=L,POT=H,RHLO=L,SHRD=L	5	3	2	0.3	23	20	3	1.2	3	3	0	0.1	31	26	5	1.5
38	IV12=H,LAT=L,ZONX=L,POT=H,RHLO=L,SHRD=H	16	12	4	0.8	37	34	3	1.9	10	10	0	0.3	63	56	7	3.0
39	IV12=H,LAT=L,ZONX=L,POT=H,RHLO=H,SHRD=L	1	1	0	0.1	3	2	1	0.2	4	4	0	0.1	8	7	1	0.4
40	IV12=H,LAT=L,ZONX=L,POT=H,RHLO=H,SHRD=H	0	0	0	0.0	12	10	2	0.6	7	7	0	0.2	19	17	2	0.9
41	IV12=H,LAT=L,ZONX=H,POT=L,RHLO=L,SHRD=L	1	1	0	0.1	1	0	1	0.1	0	0	0	0.0	2	1	1	0.1
42	IV12=H,LAT=L,ZONX=H,POT=L,RHLO=L,SHRD=H	2	2	0	0.1	22	22	0	1.1	1	1	0	0.0	25	25	0	1.2
43	IV12=H,LAT=L,ZONX=H,POT=L,RHLO=H,SHRD=L	5	4	1	0.3	10	3	7	0.5	1	1	0	0.0	16	8	8	0.8
44	IV12=H,LAT=L,ZONX=H,POT=L,RHLO=H,SHRD=H	2	2	0	0.1	2	2	0	0.1	5	3	2	0.2	9	7	2	0.4
45	IV12=H,LAT=L,ZONX=H,POT=H,RHLO=L,SHRD=L	0	0	0	0.0	3	3	0	0.2	0	0	0	0.0	3	3	0	0.1
46	IV12=H,LAT=L,ZONX=H,POT=H,RHLO=L,SHRD=H	1	1	0	0.1	12	12	0	0.6	0	0	0	0.0	13	13	0	0.6
47	IV12=H,LAT=L,ZONX=H,POT=H,RHLO=H,SHRD=L	0	0	0	0.0	14	7	7	0.7	4	4	0	0.1	18	11	7	0.9
48	IV12=H,LAT=L,ZONX=H,POT=H,RHLO=H,SHRD=H	4	4	0	0.2	6	5	1	0.3	9	8	1	0.3	19	17	2	0.9
49	IV12=H,LAT=H,ZONX=L,POT=L,RHLO=L,SHRD=L	1	1	0	0.1	35	29	6	1.8	9	9	0	0.3	45	39	6	2.2
50	IV12=H,LAT=H,ZONX=L,POT=L,RHLO=L,SHRD=H	47	44	3	2.4	120	116	4	6.1	27	26	1	0.9	194	186	8	9.3
51	IV12=H,LAT=H,ZONX=L,POT=L,RHLO=H,SHRD=L	3	3	0	0.2	2	2	0	0.1	5	4	1	0.2	10	9	1	0.5
52	IV12=H,LAT=H,ZONX=L,POT=L,RHLO=H,SHRD=H	7	7	0	0.4	12	12	0	0.6	9	9	0	0.3	28	28	0	1.3
53	IV12=H,LAT=H,ZONX=L,POT=H,RHLO=L,SHRD=L	12	11	1	0.6	12	9	3	0.6	2	2	0	0.1	26	22	4	1.3
54	IV12=H,LAT=H,ZONX=L,POT=H,RHLO=L,SHRD=H	22	19	3	1.1	23	20	3	1.2	3	3	0	0.1	48	42	6	2.3
55	IV12=H,LAT=H,ZONX=L,POT=H,RHLO=H,SHRD=L	1	1	0	0.1	2	1	1	0.1	0	0	0	0.0	3	2	1	0.1
56	IV12=H,LAT=H,ZONX=L,POT=H,RHLO=H,SHRD=H	4	4	0	0.2	10	10	0	0.5	5	5	0	0.2	19	19	0	0.9
57	IV12=H,LAT=H,ZONX=H,POT=L,RHLO=L,SHRD=L	10	8	2	0.5	40	36	4	2.0	2	2	0	0.1	52	46	6	2.5
58	IV12=H,LAT=H,ZONX=H,POT=L,RHLO=L,SHRD=H	46	44	2	2.4	96	92	4	4.9	22	22	0	0.8	164	158	6	7.9
59	IV12=H,LAT=H,ZONX=H,POT=L,RHLO=H,SHRD=L	0	0	0	0.0	0	0	0	0.0	4	4	0	0.1	4	4	0	0.2
60	IV12=H,LAT=H,ZONX=H,POT=L,RHLO=H,SHRD=H	9	9	0	0.5	6	6	0	0.3	13	10	3	0.5	28	25	3	1.3
61	IV12=H,LAT=H,ZONX=H,POT=H,RHLO=L,SHRD=L	2	2	0	0.1	19	10	9	1.0	0	0	0	0.0	21	12	9	1.0
62	IV12=H,LAT=H,ZONX=H,POT=H,RHLO=L,SHRD=H	35	30	5	1.8	59	53	6	3.0	19	15	4	0.7	113	98	15	5.4
63	IV12=H,LAT=H,ZONX=H,POT=H,RHLO=H,SHRD=L	0	0	0	0.0	4	3	1	0.2	1	1	0	0.0	5	4	1	0.2
64	IV12=H,LAT=H,ZONX=H,POT=H,RHLO=H,SHRD=H	3	3	0	0.2	5	4	1	0.3	16	15	1	0.6	24	22	2	1.2

When comparing the manually selected combinations with the output from association rule mining (Table 7-5), we can see that every highlighted combination in Table 7-4 does appear as a truly correlated association rule in Table 7-5. The number listed in the right end of each rule is the corresponding number of the combinations in Table 7-4. The rule in italicization in Table 7-5 is the optimal rule listed in Table 7-2.

Table 7-5 Association rules containing the 6 parameters in the 1982-1988 optimal condition

Intensifying rules containing any combinations of the 6 attributes (s=0.01%, c=10%, l=100%)	
FD24=RI <- IV12=L LAT_=L ZONX=L POT_=L RHLO=H SHRD=L (0.2,100.0,1950.0)	***#3***
FD24=RI <- IV12=L LAT_=L ZONX=L POT_=L RHLO=H SHRD=H (0.1,33.3,650.0)	#4
FD24=RI <- IV12=L LAT_=L ZONX=H POT_=L RHLO=H SHRD=L (0.3,75.0,1462.5)	#11
FD24=RI <- IV12=L LAT_=H ZONX=L POT_=H RHLO=L SHRD=L (0.3,21.4,417.9)	#21
FD24=RI <- IV12=L LAT_=H ZONX=L POT_=H RHLO=H SHRD=L (0.2,25.0,487.5)	#23
FD24=RI <- IV12=H LAT_=L ZONX=L POT_=L RHLO=L SHRD=L (0.2,20.0,390.0)	#33
FD24=RI <- IV12=H LAT_=L ZONX=L POT_=L RHLO=H SHRD=L (0.2,18.2,354.5)	#35
FD24=RI <- IV12=H LAT_=L ZONX=L POT_=L RHLO=H SHRD=H (0.2,25.0,487.5)	#36
FD24=RI <- IV12=H LAT_=L ZONX=L POT_=H RHLO=L SHRD=L (0.2,40.0,780.0)	#37
FD24=RI <- IV12=H LAT_=L ZONX=L POT_=H RHLO=L SHRD=H (0.3,25.0,487.5)	#38
FD24=RI <- IV12=H LAT_=L ZONX=H POT_=L RHLO=H SHRD=L (0.1,20.0,390.0)	#43
FD24=RI <- IV12=H LAT_=H ZONX=L POT_=H RHLO=L SHRD=H (0.3,13.6,265.9)	#54
FD24=RI <- IV12=H LAT_=H ZONX=H POT_=L RHLO=L SHRD=L (0.2,20.0,390.0)	#57
FD24=RI <- IV12=H LAT_=H ZONX=H POT_=H RHLO=L SHRD=H (0.4,14.3,278.6)	#62
Weakening rules containing the 6 attributes (s=10%, c=70%, l=100%)	
FD24=NR <- IV12=L LAT_=H ZONX=L POT_=L RHLO=L SHRD=H (16.7,99.5,104.9)	#18
FD24=NR <- IV12=L LAT_=H ZONX=H POT_=L RHLO=L SHRD=H (12.3,99.3,104.7)	#26

Once again, Table 7-2 shows that the support values for the optimal conditions are low, and the low support (sample size) for the highest RIP leads one to suspect the usefulness of mined results, or the representation of a general rule. One way to overcome this problem is to increase the sample size. For that reason, the discussions on the results are focused for the whole time period, 1982-2003.

The highest RIP for the period of 1982-2003 is 85.7% when six conditions are working together. Detailed investigation showed that there are seven cases satisfying the six conditions in the whole data set, and six of them underwent RI (6/7=85.7%). The seven cases are traced back to the original data and it is found that the seven cases are actually from four TCs (Yang et al. 2008). The fact that relatively large numbers of divers TCs underwent RI when the above given conditions were satisfied and the extreme high RIP lead us to believe that the results from data mining are significant. The results will not only shed light for understanding the RI processes and but also help to guide future RI forecasting.

7.6 Impacts of threshold values

In the stratified studies, the threshold values separating the binary "High" and "Low" value ranges are based on average of two group means. Here for the RI studies, the initial threshold values are based on the mean values of RI cases from KD03. The binary transform is very coarse although efficient, and it is expected that the threshold values determining the binary values will play a role in estimating the RI probability. Here, the parameter POT, the differences between the maximum possible intensity and the current intensity, is taken as an example to demonstrate the impact of the threshold values. The first threshold value for POT

is 47.7 m/s based on KD03. However, since RI is defined by a 30 knot (about 15 m/s) wind increase in 24 hours, a logical threshold for POT should be 15 m/s (M. DeMaria, private communications). Therefore, with the same threshold values for other parameters except POT, the above data mining procedure is repeated to reveal the impact by changing the POT threshold value from 47.7m/s (KD03 threshold) to 15.0m/s (physical threshold).

Figure 7-5 shows the variations of the highest RIP with number of conditions (N) for both the KD03 thresholds and the physical thresholds. Clearly, there is no difference in the results for N=1-3 and N=9. For N>3 cases, the largest RIP with the KD03 thresholds are higher than the corresponding values with the physical thresholds except for the case N=9. The overall trends of the RIP variation with N are the same, too. The uniform "no less" relationship between the RIP values with KD03 POT threshold values and those with the physical threshold may lead one to think the monotonic relationship is universal, and the exact threshold values are not important to the mining results.

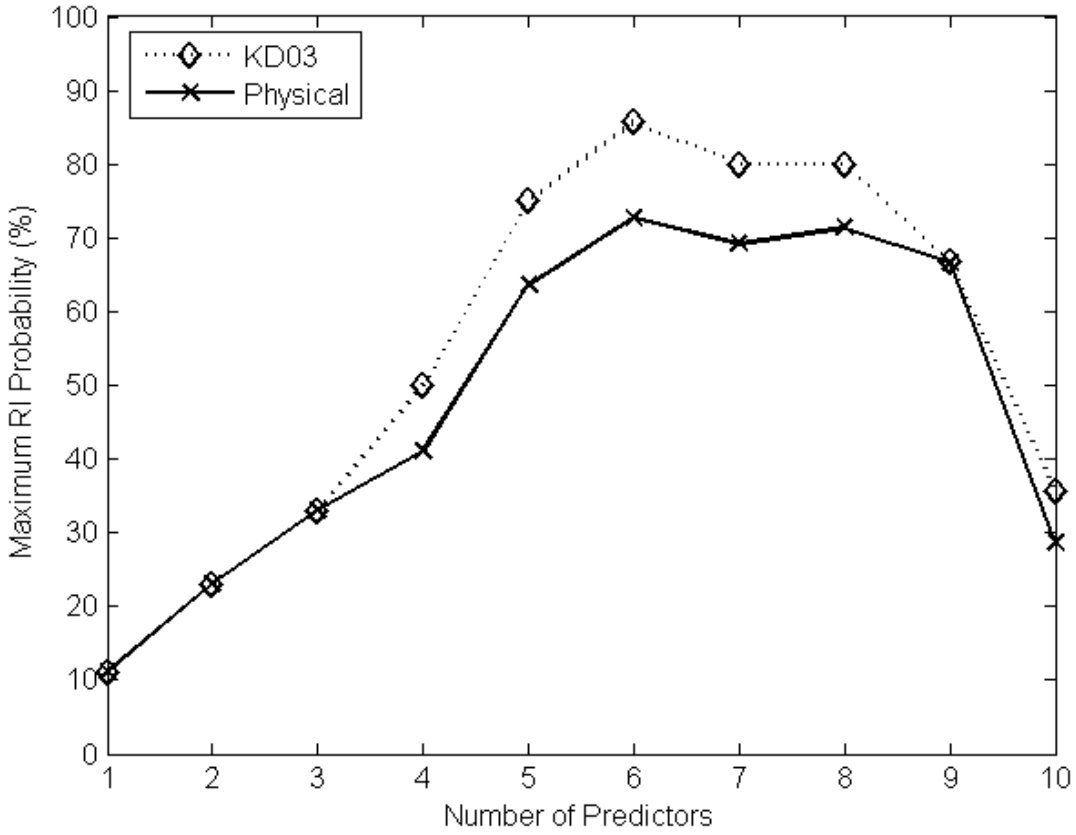


Figure 7-5 The highest RIP for two different POT threshold values. KD03: 47.7m/s; Physical: 15.0m/s.

However, the relationship between the RIP values with different threshold values is not always that simple. Figure 7-6 shows a non-monotonic relationship of RIP values for the two groups of threshold values. In this example, the original five conditions which resulted in the highest RIP in KD03 were chosen to repeat the KD03 results and to compare the result against the RIP values with the new physical POT threshold value. Once again, the original POT threshold value gives RIP values larger than those with modified physical POT threshold value for most of condition combinations. However, for the highest RI cases, i.e., when all

five conditions are satisfied, the RIP value with physical threshold is higher. This result suggests that more detailed studies are needed on both the threshold values and the corresponding RIP values.

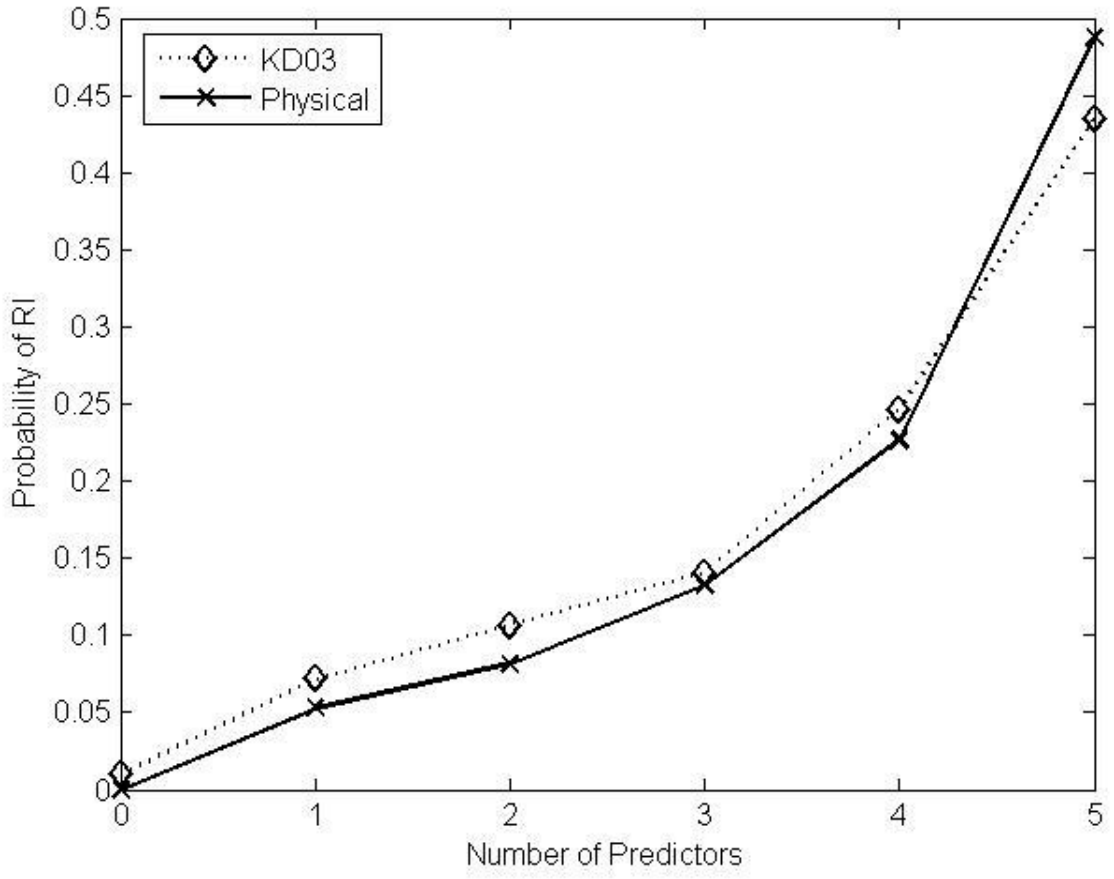


Figure 7-6 RIP variation with number of five predefined conditions (IV12=H, SST=H, POT=H, SHRD=L, RHLO=H) for the two different POT threshold values

CHAPTER 8 Summary, Discussion and Future Study

8.1 Summary

In this study, a new generation of multi-correlation data mining algorithm – Apriori (Agrawal et al. 1993) and its variation (Borgelt 2009) – has been applied to tropical cyclone data to discover "multiple-to-one" associations among a large number of geophysical characteristics that are associated with weakening, intensifying, or rapidly intensifying TCs.

The tropical cyclone data come from 3 major sources - the NHC best track data, the TRMM PR 2A25 rain rate profile and SHIPS 2003 data (DeMaria and Kaplan 1994b, 1999, and DeMaria et al. 2005). A case study of Hurricane Isabel (2003) using the structural parameters from TRMM PR 2A25 products has been conducted. Afterwards, several parameters of latent heat release of a tropical cyclone have been extracted from TRMM PR 2A25 data products. And Apriori algorithm has been applied to find the associations between the structural parameters and the intensity change in the next 6 hours. However due to the limited observations from TRMM PR, our study has shifted focus to use a more complete dataset – SHIPS 2003 dataset for association rule mining. The Apriori algorithm has been applied to study the stratified SHIPS 2003 dataset first and rules for each stratum are discussed. Subsequently the Apriori algorithm has been used in rapid intensification (RI) of tropical cyclone study.

The case study of Hurricane Isabel (2003) and the association rule mining of the structural parameters of latent heat release from TRMM PR 2A25 have produced very promising results if one wants to find additional indicators of tropical cyclone intensity change. The extracted structural information here contains the maximum height of the convective tower (HGH), the surface rain rate at the core area (CRR) and at the outer band (ORR), the Precipitation Intensity Parameter (PIP) at the core area (CPP) and at the outer band (OPP), from a total of 53 TRMM PR observations of 25 North Atlantic TCs over the period of 1998 to 2003. Our study has shown the capability of using the 3D structural information of latent heat release to identify the intensification and decaying of a TC. The association rules found by Apriori algorithm have shown the important role of these structural parameters of latent heat release when TC is growing, or when it is in stable state or when it becomes weaker. They have also shown that a stabilized tropical cyclone is the result when both the adverse and amicable factors can balance each other.

In the first study of SHIPS 2003 data, data is stratified into 7 subsets from tropical depression to category 5 hurricanes based on the Saffir-Simpson Hurricane Scale (NOAA/NWS 2007) and the Apriori algorithm is applied to find association rules for intensifying and weaking TCs. An intensifying case is identified if the TC intensity increases 5 knots or greater in the next 12 hours. A weakening case is defined as the intensity decreasing of 5 knots or greater in the next 12 hours. And a stable case is considered as the intensity change is less than 5 knots in the next 12 hours. Category 5 Hurricane data is omitted due to the too-small sample size. Listed below are the interesting results from the association rules mined from the 6 stratified tropical cyclone categories:

- The intensifying rules in all of the 6 stratified TC datasets shows that the persistence is the most prevalent and favorable factor for cyclone intensification, at least for TC of strength from tropical storms to category 2 hurricanes. Nevertheless, for tropical depressions the increasing past intensity condition should work together with another condition such as high 200 hPa divergence (D200=H), high 500-300 hPa relative humidity (RHHI=H), or large difference between the maximum possible intensity and the current intensity (POT=H, or POT2=H) to establish a two-factor condition to support TC intensification.
- The weakening rules in all of the 6 stratified TC datasets shows that the strong vertical shear condition (U200=H, SHRD=H, SRV0=H, or SRLA=H) and the not-so-strong past intensity increase condition (IV12=L, or VV=L) are the determining factors for cyclone weakening.
- Both the intensifying and weakening rules show that the parameters describing the differences between the current intensity and the maximum possible intensity, or intensity potential, POT and POT2 play a role mainly for weak TCs.
- Associated hyper-edge search shows a perfect bind between IV12 and VV in all subsets. It also finds strong intramural binds among U200, SHRD, SRV0 and SRLA. Another strong binding revealed by the hyper-edge relationships is POT and POT2 pairs.
- During pre-processing the two-side two-sample t test is performed to reexamine the 21 parameters for their ability to distinguish the intensifying and weakening stages. We find out that the t-test selected attributes are different in each stratum. This confirms

the suspicion that the attribute selection performed in SHIPS model is skewed by the dominant tropical storm records.

- During pre-processing while examining the values of the means for intensifying and weakening cases for each parameter (that is, the I-W relation in this study), we find out that most of the parameters have a consistent "I-W" relation over different TC categories, indicating a consistent function of that parameter in TC development. However, the "I-W" relations of the convergence of the upper atmosphere (REFC) and the meridional motion (MERY) change their directions from tropical storms to hurricanes, indicating a change of role of that parameter in TC development.

The rapid intensification study of SHIPS 2003 dataset closely follows the work of KD03 which examined the large-scale characteristics of rapidly intensifying Atlantic TCs from 1989 to 2003. The same eleven independent parameters are used in this study. And the same split value definition – the sample mean of the RI cases – is used here to divide each selected parameter into "High" and "Low" value ranges for further mining. The same RI definition as in KD03 where a RI case is identified if the intensity of a TC (defined by the maximum wind) has increased at least 30 knots (15.4 m/s) over a 24-hour period to classify RI and non RI cases in the dataset. In this study the NHC best track data and SHIPS 2003 dataset from 1982 to 2003 are merged based on the methodology identical to that described in KD03. The SHIPS 2003 data are interpolated into 6-h interval for the RI study. The association rule mining shows following important findings for rapid intensified tropical cyclones:

- The original rule proposed in KD03, which states as "RI <= SHRD=L, PD12=H, SST=H, POT=H, RHLO=H (supp=0.7%, conf=43.5%, lift=850.5%)," is found by the

association rule mining but then removed as a redundant rule during association rule pruning as part of post-processing in this study. A stronger and more concise rule with three constraints satisfied, "RI <- SHRD=L, PD12=H, RHLO=H (supp=1.3%, conf=47.6%, lift=931.5%)," remains. Similar conclusions are found after examining the datasets of 1982-1988, 1982-2003 and 2001-2003. This means that this rule with all 5 constraints satisfied in KD03 is not the best and a rule with fewer constraints fits more RI cases.

- As a successful scientific data mining example, a case is identified in this study that when three conditions, low vertical shear of horizontal wind (SHRD=L), high humidity in the 850-700hPa level (RHLO=H), and the TC being in intensification phase (IV12=H), are satisfied, the RI probability (RIP) is higher than what was found in KD03, which included the above three conditions and two additional conditions, high sea surface temperature (SST=H), and an intensity far away from the maximum possible intensity (POT=H).

The results in this studies show successful data mining examples in geosciences with association rules. In the stratified TC cases, the association rules did identify conditions combinations for either intensifying or weakening cases. The probabilities are generally much higher than the sample means as indicated by the lift values. In most cases, the identified condition combinations give higher probabilities than those based on single condition which are most likely the choices with traditional statistical studies. The higher RI probabilities with fewer conditions compared with the result from traditional statistical analysis are significant. This work demonstrates that data mining techniques can be used as exploration methods to

generate hypotheses, and the statistical analysis should be performed as confirmation of the hypotheses, as generally expected for data mining applications (e.g., Hand et al. 2001).

8.2 Discussion

In this study we have successfully applied the association rule mining technique to tropical cyclone intensity change. The study reveals several challenges along with the data pre-processing and post-processing since there is no clear guidance on what a scientist should be looking for before mining.

The first challenge happens at the stratification method at the data pre-processing stage. Here in this study we divide the SHIPS 2003 data into 7 stratums based on the Saffir-Simpson scales. However the category 5 Hurricane data are omitted due to the very small sample size. It is acceptable here since we are focusing on data exploration and analysis but it will not help much if we want to apply the findings to real world of tropical cyclone forecast in the future. One way to include the category 5 hurricane data, which can be part of our future study, may be to stratify the dataset into 4 categories: tropical depression (TD), tropical storm (TS), category 1 and 2 hurricanes, and the major hurricanes (category 3, 4 and 5).

The second challenge appears at how to choose the split value to separate a parameter into a “High” or “Low” value range during data pre-processing. Our study displayed in Figure 7-5 and Figure 7-6 show that the RIP values do depend on the threshold values of POT. Since the sample sizes of RI cases and non-RI cases are highly skewed, sophisticated methods such as the decision trees or Bayesian theory on two-class classification should be considered.

The third challenge is revealed when we try to identify the "interesting" mined rules. Although the fact that association rule mining exhausts all of the combinations of the parameters is the major advantage to traditional one-to-one statistical comparison method, it also presents a stack of hay for a needle finder which limits its own usage, especially when the number of parameters are large as in the tropical cyclone intensity change study here. Several rule pruning techniques must be considered in finding the most interesting rules. In our study the measurements of support, confidence and lift are used first as a standard practice to find truly correlated rules. In order to have a manageable number of truly correlated rules, the levels of support, confidence and lift are adjusted. Moreover, our study benefits from concise rule searching which further reduces the overlap among the truly correlated rules. Additionally, through association hyper-edge search the hidden relationships among multiple parameters are found even without any background knowledge of geosciences

8.3 Future work

Looking back at the 3D structural parameters we extracted from TRMM PR data in Chapter 4, the max height of the convective towers contains the only information at the vertical scale of a tropical cyclone. In order to fully explore the power of 3D structure revealed by TRMM PR, one should try different definitions of the height of a convective tower in addition to that of Kelley and Stout (2004). Some of the definitions can be (1) the maximum height of a columnar rain in the entire swath (with its distance to the eye center), (2) the average height (or the median) of a columnar rain in the entire swath, (3) the 75% or 25% height in the entire swath, (4) the averaged height of the columns with max rain rate, (5)

the averaged height of the columns with average rain rate, etc. And using the definition of the inner band and outer band around the eye center, the concepts of the different heights can also be applied to them to present a detail layout of heights from inner core to outer band for a tropical cyclone.

As far as the data mining methodology is concerned, the data pre-processing should consider the impact of various data stratification schema and data discretization schema. Importantly the climatologically long term means should be used instead of the sample means which are varied based on the sample sizes. The rule pruning should still utilize the support, confidence and lift measurements to find truly correlated rules. It still makes sense to use the concise rule pruning and hyper-edge searching. Furthermore, the log-linear model proposed by Wu, Barbara, and Ye (2003) and the chi-square test proposed by DuMouchel and Pregibon (2001) should be considered to explore infrequent but surprising association rules in the future.

Although the study of latent heat release structure of a tropical cyclone versus its intensity change is very promising, an integration of those parameters to real time TC intensity forecast model such as SHIPS still has a long way to go. When the future objective is changed to improve the current forecasting skills, the future work should be focused on SHIPS data.

Although SHIPS is the most skillful among the NHC operational intensity forecasting models and it has been improved in the past, the linear feature limits further significant improvements (DeMaria 2009). Furthermore, our stratification study show that the controlling parameters change with different initial intensity, the role of a parameter can change during

TC lifetime, and the large number of tropical storm cases skews the parameter selection in SHIPS model. Moreover the unanticipated RI process in TC development, which is one major challenge for intensity forecasting (KD03; Rappaport et al. 2009; Kaplan et al. 2010), is difficult to be correctly forecasted based on a global linear regression model. As such, an outline is given below for leveraging data mining techniques to improve statistical TC intensity forecasting in the future study.

First, the condition combinations found in association rules should be explored more and some kind of composite index may be established based on them. Recently Wu et al. (2007) defined a SAL effect index (SEI) by a linear combination of relative humidity and wind shear and claimed that SEI "can much better account for the trend in the mean peak intensity … than SST or AMO index." This result is highly consistent with our mining results of the improved RI probability by the combined contribution from the conditions "SHRD=L" and "RHLO=H" in Chapter 7. Considering these facts, we do believe that other associated conditions with high support and confidence can be used as a hypothesis to build new composite indices to account for TC intensity change. Furthermore log-linear model may be useful to identify whether a linear combination is enough or a non-linear combination is required among the combined conditions. Consequently the coefficient of the nonlinear terms can be used in the composite index formation. When a linear relationship is identified, CrystalVision (a visualization tool) and the grand tour method may be useful to establish the coefficients of a linear based index (Wegman 1991, 2003).

Secondly, several piecewise regression models (Breiman et al. 1993) against the stratified initial intensity should be established. In the future study, a set of association rules

which have the complete coverage of the intensifying cases should be found first. The qualified rules must have a lift greater than 1 to guarantee the correlation is true. The qualified rules must be concise rules to avoid redundancy. The rule selection can start at the concise rules with the most support then go to the concise rules with smaller support. The confidence for the selected rules may vary. In order to avoid the problem of over-fitting, a qualified rule is better to cover cases from different TCs. The goal is to cover most of the intensifying cases with as few rules as possible. The rules with large support may indicate the major processes controlling TC intensification. The rules with small support reflect a particular process controlling several TCs. From the set of selected association rules, we can gain more understanding of the processes behind the TC intensification. Later on a set of regression models can be established to estimate the actual amount of intensity change.

Finally in order to fight against the forecast error introduced by rapid intensification, an RI indicator should be established to identify the probability of rapid intensification for each forecast. When the probability is high enough, an artificial modification should be introduced in the regression model based on the rapid intensification probability (RIP) of TCs. As a result, the skill of TC intensity forecasting can be improved.

REFERENCES

REFERENCES

Aberson, S.D., 2001: The ensemble of tropical cyclone track forecasting models in the north atlantic basin (1976-2000). *Bulletin of the American Meteorological Society*, **82(9)**, pp1895-1904.

Agrawal, R., T. Imielinski, and A. Swami, 1993: Mining association rules between sets of items in large databases, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington D.C., May 1993, 207-216.

Bankert, R.L., and P.M. Tag, 2002: An automated method to estimate tropical cyclone intensity using SSM/I imagery, *Journal of Applied Meteorology*, **41(5)**, 2002, pp 461-472.

Borgelt, C., 2009: Apriori - Association Rule Induction/Frequent Item Set Mining, http://www.borgelt.net/apriori.html, last access on May 11, 2009.

Bosart, L.F., C.S. Velden, W.E. Bracken, J. Molinari, and P.G. Black, 2000: Environmental influences on the rapid intensification of Hurricane Opal (1995) over the Gulf of Mexico. *Mon. Wea. Rev*., **128**, 322–352.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1993: *Classification and regression trees*. Chapman & Hall, New York, USA.

Camargo, S.J., A.W. Robertson, S.J. Gaffney, P. Smyth, and M. Ghil, 2007a: Cluster Analysis of Typhoon Tracks. Part I: General Properties. *J. Climate*, **20**, 3635–3653.

Camargo, S.J., A.W. Robertson, S.J. Gaffney, P. Smyth, and M. Ghil, 2007b: Cluster Analysis of Typhoon Tracks. Part II: Large-Scale Circulation and ENSO. *J. Climate*, **20**, 3654–3676.

Camp, J.P., and M.T., Montgomery, 2001: Hurricane maximum intensity: past and present, *Monthly Weather Reviews*, **129**, 1704-1717.

Cecil, D. J. and E. J. Zipser, 1999: Relationships between tropical cyclone intensity and satellite-based indicators of inner core convection: 85GHz ice-scattering signature and lightning, *Monthly Weather Review*, **127(1)**, 103-123.

Chang, A.T.C., L.S. Chiu, G.R. Liu, and K.H. Wang, 1995: Analyses of 1994 typhoons in the Taiwan region using satellite data, In: *Space Remote Sensing of Subtropical Oceans (SRSSO*) 1995 September 12-16, Taipei, TAIWAN, C.T. Liu, Ed., Elsevier Science, New York, New York, 89-96.

Chakrabarti, D., and C. Faloutsos, 2002: F4: Large-scale Automated Forecasting Using Fractals, in *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2002)*, McLean, VA, USA, November 4-9, 2002, 2-9.

Cram, T.A., J. Persing, M.T. Montgomery, and S.A. Braun, 2007: A Lagrangian Trajectory View on Transport and Mixing Processes between the Eye, Eyewall, and Environment Using a High-Resolution Simulation of Hurricane Bonnie (1998). *J. Atmos. Sci.*, **64**, 1835–1856.

Davis, C., C. Snyder, and A.C. Didlake, 2008: A Vortex-Based Perspective of Eastern Pacific Tropical Cyclone Formation. *Mon. Wea. Rev.*, **136**, 2461–2477.

DeMaria, M., J.-J Baik, and J. Kaplan, 1993: Upper-level eddy angular momentum flux and tropical cyclone intensity change, *J. Atmos. Sci.*, **50**, 1133-1147.

DeMaria, M., and J. Kaplan, 1994a: Sea Surface Temperature and the Maximum Intensity of Atlantic Tropical Cyclones. *J. Climate*, **7**, 1324–1334.

DeMaria, M., and J. Kaplan, 1994b: A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin, *Wea. Forecasting*, **9**, 209-220.

DeMaria, M. and J. Kaplan, 1999: An Updated Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic and Eastern North Pacific Basins Mark, *Wea. Forecasting*, **14**, 326–337.

DeMaria, M., 1996: The effect of vertical shear on tropical cyclone intensity change, *J. Atmos. Sci.*, **53**, 2076-2087.

DeMaria, M., 1997: Summary of the NHC/TPC tropical cyclone track and intensity guidance models, November 26 1997, at http://www.nhc.noaa.gov/aboutmodels.shtml, last accessed on Sep 09, 2004.

DeMaria, M., M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further Improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS), *Wea. Forecasting*, **20**, 531–543.

DeMaria, M., 2009: A Simplified Dynamical System for Tropical Cyclone Intensity Prediction. *Mon. Wea. Rev.*, **137**, 68–82.

DeMaria, M., J.A. Knaff, and C.R. Sampson, 2007: Evaluation of Long-Term Trends in Operational Tropical Cyclone Intensity Forecasts. *Meteor. and Atmos. Phys.*, **59**, 19-28.

Dunion, J.P., and C.S. Velden, 2004: The impact of the Saharen Air Layer on Atlantic tropical cyclone activity, *Bull. Amer. Meteor. Soc.,* **90**, pp353-365

DuMouchel, W., and D. Pregibon, 2001: Empirical bayes screening for multi-item associations, in *Proceedings of the 7th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 26-29 2001, 67-76.

Dvorak, V.F., 1972: A technique for the analysis and forecasting of tropical cyclone intensities from satellite pictures, *NOAA Technical Memorandum NESS 36*, 15.

Dvorak, V.F., 1975: Tropical cyclone intensity analysis and forecasting from satellite imagery. *Monthly Weather Review,* **103(5),** 420-430.

Dvorak, V., 1984: Tropical cyclone intensity analysis using satellite data. *NOAA Technical Report NESDIS 11*, 47pp. NOAA/NESDIS, 5200 Auth Rd., Washington D.C. 20233.

Edson, R.T., 2004: Uses of microwave imagery as a supplement to the Dvorak technique, an integrated technique (extended abstract), in *the 26th conference on hurricanes and tropical meteorology*, Miami, FL, USA, May 2-7, 2004, 8A.2.

Emanuel, K.A., 1988: The maximum intensity of hurricanes. *J. Atmos. Sci.,* **45,** 1143–1155.

Faloutsos, C., 2004: Stream and Sensor data mining, *Tutorial for the 9th International Conference on Extending Database Technology (EDBT 2004)*, March 14-18, 2004, Heraklion-Crete, Greece.

FEMA, 1995: National mitigation strategy: partnerships for building safer communities. *Mitigation Directorate*, p.2, Federal Emergency Management Agency, Washington, D.C.

Frank, W. M. and E. A. Ritchie, 1999: Effects of environmental flow upon tropical cyclone structure. *Mon. Wea. Rev.*, **127**, 2044–2061.

Frank, W. M., and E. A. Ritchie, 2001: Effects of vertical wind shear on hurricane intensity and structure. *Mon. Wea. Rev.*, **129**, 2249-2269.

Franklin, J. L., 2008: 2007 National Hurricane Center forecast verification report. NOAA/NWS/NCEP/Tropical Prediction Center, 68 pp. online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2007.pdf, last accessed on October 21, 2008.

Gedzelman, S., J. Lawrence, J. Gamache, M. Black, E. Hindman, R. Black, J. Dunion, H. Willoughby, and X. Zhang, 2003: Probing Hurricanes with Stable Isotopes of Rain and Water Vapor, *Mon. Wea. Rev.*, **131**, 1112–1127.

Gray, W. M., 1968: Global view of the origin of tropical disturbances and storms, *Mon. Wea. Rev.*, **96**, 669–700.

Guard, C., 2004: The longevity of the Dvorak tropical cyclone intensity technique in the Pacific and Indian Ocean regions (extended abstract), in *the 26th conference on hurricanes and tropical meteorology*, Miami, FL, USA, May 2-7, 2004, 7A.3.

Han, E.H., G. Karypis, V. Kumar and B. Mobasher, 1997: Clustering based on association rule hypergraphs. In *Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 9-13, Tucson, Arizona, 1997.

Hand, D., H. Mannila, and P. Smyth, 2001: *Principles of Data Mining*. M.I.T Press.

Hanley, D., J. Molinari, and D. Keyser, 2001: A Composite Study of the Interactions between Tropical Cyclones and Upper-Tropospheric Troughs. *Mon. Wea. Rev.*, **129**, 2570–2584.

Herndon, D.C., C.S. Velden, K. Brueske, R. Wacker, and B. Kabat, 2004: Upgrades to the UW-CIMSS AMSU-based tropical cyclone intensity estimation algorithm (extended abstract), in *the 26th conference on hurricanes and tropical meteorology*, Miami, FL, USA, May 2-7, 2004, 4D.1.

Herbert, P.J., J.D. Jarrell, and M. Mayfield, 1996: The deadliest, costliest, and most intense hurricanes of this century (and other frequently requested hurricane facts). *NOAA Technical Memorandum*, NWS TPC-I, 1996, NOAA, Washington, D.C.

Holland, G.J., and R.T. Merrill, 1984: On the dynamics of tropical cyclone structural changes. *Quart. J. Roy. Meteor. Soc.,* **110**, 723-745.

Holland, G.J., 1997: The maximum potential intensity of tropical cyclones, *Journal of the Atmospheric Sciences*, **54**, pp 2519-2541.

Holton, J.R., 1992: An Introduction to Dynamic Meteorology. Academic Press, San Diego, California.

Hong, X., S. W. Chang, S. Raman, L. K. Shay, and R. Hodur, 2000: The interaction between Hurricane Opal (1995) and a warm core ring in the Gulf of Mexico. *Mon. Wea. Rev.*, **128**, 1347–1365.

Hoshino, S., and T. Nakazawa, 2004: Method to estimate tropical cyclone intensity using TRMM PR/TMI data (extended abstract), in *20th Conference on Weather Analysis and*

Forecasting/16th Conference on Numerical Weather Prediction, Seattle, WA, USA, Jan 10-15, 2004, 13.4.

Huang, Y., H. Xiong, S. Shekhar and J. Pei, 2003: Mining confident co-location rules without a support threshold, in Proc. 2003 ACM Symposium on Applied Computing (ACM SAC'03), Melbourne, Florida, March 9 - 12, 2003.

Huang, Y., S. Shekhar, and H. Xiong, 2004: Discovering co-location patterns from spatial databsets: a general approach, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, **16(2),** 2004, pp 1472-1485.

Jarvinen, B. R., and C. J. Neumann, 1979: Statistical forecasts of tropical cyclone intensity for the North Atlantic basin. *NOAA Tech. Memo*. NWS NHC-10, 22.

Jarvinen, B.R., C.J. Neumann, and M.A.S. Davis, 1984: A tropical cyclone data tape for the North Atlantic basin, 1886-1983: Contents, limitations, and uses, *NOAA Technical Memorandum* NWS NHC 22.

Jin, Y., W.T. Thompson, S. Wang, and C.S. Liou, 2007: A Numerical Study of the Effect of Dissipative Heating on Tropical Cyclone Intensity. *Wea. Forecasting*, **22**, 950–966.

Kaplan, J. and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin, *Wea. Forecasting*, **18**, 1093-1108.

Kaplan J., M. DeMaria, J.A. Knaff, 2009: A Revised Tropical Cyclone Rapid Intensification Index for the Atlantic and Eastern North Pacific Basins. *Wea. Forecasting*, In Press .

Kelley, O. and J. Stout, 2004: Convective towers in eyewalls of tropical cyclones observed by the TRMM Precipitation Radar in 1998–2001. In *20th Conference on Weather Analysis*

and Forecasting/16th Conference on Numerical Weather Prediction in the 84th AMS Annual Meeting, Jan 10-16 2004, Seattle, WA.

Kidder, S.Q., W.M. Gray, and T.H. Vonder Haar, 1978: Estimating tropical cyclone central pressure and outer winds from satellite microwave data. *Monthly Weather Review*, **106**, pp 1458-1464.

Kitamoto, A., 2001: Data mining for typhoon image collection, *Proceedings of the 2nd International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference*, San Francisco, USA, August 26, 2001, pp 68–77.

Kitamoto, A., 2002a: Spatio-temporal data mining for typhoon image collection, *Journal of Intelligent Information Systems*,**19 (1)**, pp 25–41.

Kitamoto, A., 2002b: Evolution Map: modeling state transition of typhoon image sequences by spatio-temporal clustering, *Discovery Science (DS2002), Lecture Notes in Computer Science 2534*, Lange, S., Satoh, K., and Smith, C.H. (Eds.), Springer, pp. 283-290.

Knaff, J.A., S.A. Seseske, M. DeMaria, and J.L. Demuth, 2004: On the Influences of Vertical Wind Shear on Symmetric Tropical Cyclone Structure Derived from AMSU. *Mon. Wea. Rev.*, **132**, 2503–2510.

Knaff, J.A., T.A. Cram, A.B. Schumacher, J.P. Kossin, and M. DeMaria, 2008: Objective Identification of Annular Hurricanes. *Wea. Forecasting*, **23**, 17–28.

Koperski, K., J. Han, and N. Stefanovic, 1998: An efficient two-step method for classification of spatial data, in Proc. 1998 International Symposium on Spatial Data Handling SDH'98, Vancouver, BC, Canada, July 1998.

Koperski, K., and J. Han, 1995: Discovery of spatial association rules in geographic information databases, in Proc. 4th Int'l Symp. on Large Spatial Databases (SSD95), Maine, Aug. 1995, pp. 47-66.

Landsea, C.W., 1993: A climatology of intense (or major) Atlantic hurricanes. *Monthly Weather Reviews*, **121**, pp. 1703-1713.

Landsea, C., 1997, FAQ: Hurricanes, typhoons and tropical cyclones, at website http://www.faqs.org/faqs/meteorology/storms-faq/part1/, last accessed on Feb. 24, 2010. Another updated FAQ website is http://www.aoml.noaa.gov/hrd/tcfaq/tcfaqHED.html, last access on Feb. 24, 2010.

Lee, W.-C., and M. M. Bell, 2007: Rapid intensification, eyewall contraction, and breakdown of Hurricane Charley (2004) near landfall, *Geophys. Res. Lett*., **34**, L02802, doi:10.1029/2006GL027889.

Lowag, A., M.L. Black, and M.D. Eastin, 2008: Structural and Intensity Changes of Hurricane Bret (1999). Part I: Environmental Influences. *Mon. Wea. Rev.*, **136**, 4320–4333.

May, D. A., J. Sandidge, R. Holyer, and J. D. Hawkins, 1997: SSM/I derived tropical cyclone intensities. *22d Conf. onHurricanes and Tropical Meteorology*, Fort Collins, CO, Amer.Meteor. Soc., 27–28.

McTaggart-Cowan, R., L.F. Bosart, C.A. Davis, E.H. Atallah, J.R. Gyakum, and K.A. Emanuel, 2006: Analysis of Hurricane Catarina (2004). *Mon. Wea. Rev.*, **134**, 3029–3053.

Merrill, R.T., 1988: Environmental Influences on Hurricane Intensification. *J. Atmos. Sci.*, **45**, 1678–1687.

Molinari, J., and D. Vollaro, 1989: External influences on hurricane intensity. Part I: Outflow-layer eddy angular momentum fluxes. *J. Atmos. Sci*., **46**, 1093–1105.

Molinari, J., and D. Vollaro, 1990: External influences on hurricane intensity. Part II: Vertical structure and response of the hurricane vortex. *J. Atmos. Sci*., **47**, 1902–1918.

Montgomery, M.T., M.M. Bell, S.D. Aberson, and M.L. Black, 2006: Hurricane Isabel (2003): New Insights into the Physics of Intense Storms. Part I: Mean Vortex Structure and Maximum Intensity Estimates. *Bull. Amer. Meteor. Soc.*, **87**, 1335–1347.

Nakazawa, T., and S. Hoshino, 2004: Tropical cyclone intensity estimation by TRMM/TMI microwave radiometer data (extended abstract), in *the 26th conference on hurricanes and tropical meteorology*, Miami, FL, USA, May 2-7, 2004, P1.48, 2 pages.

NOAA/NWS (National Oceanic and Atmospheric Administration/National Weather Service), 2007: Hurricanes...Unleashing Nature's Fury, online at http://www.nws.noaa.gov/om/hurricane/pdfs/HurricanesUNF07.pdf, last accessed on October 21, 2008.

Olander, T.L., C.S. Velden and J.P. Kossin, 2004: The Advanced Objective Dvorak Technique (AODT): Latest Upgrades and Future Directions (extended abstract), in *the 26th conference on hurricanes and tropical meteorology*, Miami, FL, USA, May 2-7, 2004, P1.19, 2 pages.

Pielke, R.A. Jr.; and C.W. Landsea, 1998: Normalized hurricane damages in the United States: 1925–1995, *Weather and Forecasting*, **13**, pp621–631.

Rao, G.V., and J.H. McCoy, 1997: SSM/I measured microwave brightness temperature (TB's), anomalies of TB's, and their relationship to typhoon intensification, *Natural Hazards*, **15 (1),** pp 1-19.

Rao, G.V., and P.D. MacArthur, 1994a: The SSM/I estimated rainfall amounts of tropical cyclones and their potential in predicting the cyclone intensity changes, *Monthly Weather Review*, **122**, pp 1568-1574.

Rao, G.V., and P.D. MacArthur, 1994b: Some characteristics of typhoons as revealed by the recent SSM/I microwave radiometry, *Natural Hazards*, **9**, pp 17-35.

Rappaport, E.N., J.L. Franklin, L.A. Avila, S.R. Baig, J.L. Beven, E.S. Blake, C.A. Burr, J.G. Jiing, C.A. Juckins, R.D. Knabb, C.W. Landsea, M. Mainelli, M. Mayfield, C.J. McAdie, R.J. Pasch, C. Sisko, S.R. Stewart, and A.N. Tribble, 2009: Advances and Challenges at the National Hurricane Center. *Wea. Forecasting*, **24**, 395–419.

Riehl, H., and J.S. Malkus, 1958: On the heat balance in the equatorial trough zone. *Geophysica* (Helsinki), **6**, 503–38, 1958.

Rodgers, E.B., and R.F. Adler, 1981: Tropical cyclone rainfall characteristics as determined from a satellite passive microwave radiometer, *Monthly Weather Review*, **109**, pp 506-521.

Rodgers, E.B., S.W. Chang and H.F. Pierce, 1994: A satellite observational and numerical study of the precipitation characteristics in western North Atlantic tropical cyclones. *Journal of Applied Meteorology*, **33**, pp129-139.

Rodgers, E.B., J-J. Baik, and H.F. Pierce, 1994: The environmental influence on tropical cyclone precipitation, *Journal of Applied Meteorology*, May 1994, **33**, pp573-593.

Rogers, R., S. Chen, J. Tenerelli, and H. Willoughby, 2003: A Numerical Study of the Impact of Vertical Shear on the Distribution of Rainfall in Hurricane Bonnie (1998). *Mon. Wea. Rev.*, **131**, 1577–1599.

Shay, L. K., G. J. Goni, and P. G. Black, 2000: Effects of a warm oceanic feature on Hurricane Opal. *Mon. Wea. Rev.*, **128**, 1366–1383.

Shekhar, S., and Y. Huang, 2001: Co-location rules mining: a summary of results, in *Proc. 7th Intl. Symposium on Spatio-temporal Databases*, 2001.

Silverstein, C., S. Brin, and R. Motwani, 1998: Beyond market baskets: generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery,* **2**, 39-68.

Srikant, R., and R. Agrawal, 1996: Mining Quantitative Association Rules in Large Relational Tables, *Proc. of the ACM SIGMOD Conference on Management of Data*, Montreal, Canada, June 1996.

Tan, P-N. M. Steinbach, V. Kumar, et al, 2001: Finding spatio-temporal patterns in earth science data, *KDD 2001 Workshop on Temporal Data Mining*, August 26, 2001, San Francisco, CA, USA.

Tang, J., R. Yang, and M. Kafatos, 2005: Data Mining for Tropical Cyclone Intensity Prediction, *AMS 2005 the Sixth Conference on Coastal Atmospheric and Oceanic Prediction and Processes*, January 8, 2005, San Diego, CA, USA, **7.5.**

Velden, C.S., and W.L. Smith, 1983: Monitoring tropical cyclone evolution with NOAA satellite microwave observations, *Journal of Climate and Applied Meteorology*, **22**, pp714-724.

Velden, C.S., T.L. Olander, R.M. Zehr, 1998: Development of an objective scheme to estimate tropical cyclone intensity from digital geostationary satellite infrared imagery,*Weather and Forecasting*, **13 (1)**, pp172-186.

Ventham, J.D., and B. Wang, 2007: Large-Scale Flow Patterns and Their Influence on the Intensification Rates of Western North Pacific Tropical Storms. *Mon. Wea. Rev.*, **135**, 1110–1127.

Wang, Y., 2002a: An Explicit Simulation of Tropical Cyclones with a Triply Nested Movable Mesh Primitive Equation Model: TCM3. Part II: Model Refinements and Sensitivity to Cloud Microphysics Parameterization *Mon. Wea. Rev.*, **130**, 3022–3036.

Wang, Y., 2002b: Vortex Rossby Waves in a Numerically Simulated Tropical Cyclone. Part II: The Role in Tropical Cyclone Structure and Intensity Changes. *J. Atmos. Sci.*, **59**, 1239–1262.

Wang, Y., J.D. Kepert, and G.J. Holland, 2001: The Effect of Sea Spray Evaporation on Tropical Cyclone Boundary Layer Structure and Intensity. *Mon. Wea. Rev.*, **129**, 2481–2500.

Wang, Y., and C.C. Wu, 2004: Current understanding of tropical cyclone structure and intensity changes – a review, *Meteorology and Atmospheric Physics*, **87**, 257–278.

Wegman, E. J., 1991: The grand tour in k-dimensions, *Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface*, 127-136.

Wegman, E. J., 2003: Visual data mining, *Statistics in Medicine*, **22(9),** 1383 – 1397.

Wilks, Daniel S., 2005: "*Statistical Methods in the Atmospheric Sciences: An Introduction*," Academic Press, December 2005.

Willoughby, H. E., J. A. Clos, and M. G. Shoreibah, 1982: Concentric eyewalls, secondary wind maxima, and the evolution of the hurricane vortex, *J. Atmos. Sci.*, **39**, 395–411.

Willoughby, H.E. and P.G. Black, 1996: Hurricane Andrew in Florida: Dynamics of a Disaster, *Bull. Am. Meteorol. Soc.*, **77**, 543–549.

Wimmers, A., and C.S. Velden, 2004: Satellite-based center-fixing of tropical cyclones: new automated approaches (extended abstract), in *the 26th conference on hurricanes and tropical meteorology*, Miami, FL, USA, May 2-7, 2004, 3D.3, 2 pages.

Wu, C.C., C.Y. Lee, and I.I. Lin, 2007: The Effect of the Ocean Eddy on Tropical Cyclone Intensity. *J. Atmos. Sci.*, **64**, 3562–3578.

Wu, X., D. Barbara, and Y. Ye, 2003: Screening and interpreting multi-item associations based on log-linear modeling, in *Proceedings of the 9th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, August 24-27 2003, pp 276-285.

Xiong, et al, 2004: A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects, in *Proc. 2004 SIAM Int'l Conf. on Data Mining (SDM 2004)*, pp. 78 - 89, Florida, USA, 2004.

Yang, R., D. Sun, and J. Tang, 2008: A "sufficient" condition combination for rapid intensifications of tropical cyclones, *Geophys. Res. Lett.*, **35**, L20802.

Yang, R., J. Tang, and M. Kafatos, 2007: Improved associated conditions in rapid intensifications of tropical cyclones, *Geophys. Res. Lett*., **34**, L20807.

Yi, B-K, N.D. Sidiropoulos, T. Johnson, H.V. Jagadish, C. Faloutsos, 2000: Online data mining for co-evolving time sequences, *in IEEE Conf. on Data Engineering (ICDE 2000),* San Diego, CA, 2000.

Zehr, R., 1989: Improving objective satellite estimates of tropical cyclone intensity (Extended abstracts), in *the18th Conference on Hurricanes and Tropical Meteorology*, San Diego, CA, American Meteorologist Society, pp J25-J28.

Zehr, R.M., 2003: Environmental Vertical Wind Shear with Hurricane Bertha (1996). *Wea. Forecasting*, **18**, 345–356.

Zehr, R.M., and J.A. Knaff, 2007: Atlantic Major Hurricanes, 1995–2005—Characteristics Based on Best-Track, Aircraft, and IR Images. *J. Climate*, **20**, 5865–5888.

Zeng, Z., L. Chen, and Y. Wang, 2008: An Observational Study of Environmental Dynamical Control of Tropical Cyclone Intensity in the Atlantic. *Mon. Wea. Rev.*, **136**, 3307–3322.

Zeng, Z., Y. Wang, and C.C. Wu, 2007: Environmental Dynamical Control of Tropical Cyclone Intensity—An Observational Study. *Mon. Wea. Rev.*, **135**, 38–59.

Zhu, Y., and D. Shasha, 2002: StatStream: statistical monitoring of thousands of data streams in real time. In Proceedings *of the 28th International Conference on Very Large Data Bases (VLDB, 2002).*

Zhu, T., and D.L. Zhang, 2006: Numerical Simulation of Hurricane Bonnie (1998). Part II: Sensitivity to Varying Cloud Microphysical Processes. *J. Atmos. Sci.*, **63**, 109–126.

Zhu, T., D.L. Zhang, and F. Weng, 2004: Numerical Simulation of Hurricane Bonnie (1998). Part I: Eyewall Evolution and Intensity Changes, *Mon. Wea. Rev.,* **132**, 225–241.

CURRICULUM VITAE

Ms. Jiang Tang received her Bachelor of Science in Environmental Sciences from Peking University, Beijing, China, in 1997. She received her Master of Science in Cartography and Geographical Information System from Peking University, Beijing, China, in 2000. In May 2005, Ms. Jiang Tang received her Master of Science in Information System from George Mason University, Fairfax, Virginia. She was employed as a software engineer in the Society of Worldwide Interbank Financial Telecommunication, Inc., Manassas, Virginia, for four years and received her Doctor of Philosophy from George Mason University in 2010.